Tools for massive bacterial comparative genomics: Development and Applications.

Amandine Perrin

Sorbonne Université

Ecole doctorale Complexité du vivant

Génomique Evolutive des Microbes (GEM)
Hub de Bioinformatique et Biostatistique
Institut Pasteur

Tools for massive bacterial comparative genomics : Development and Applications.

Par Amandine PERRIN

Thèse de doctorat de Bioinformatique

Dirigée par Eduardo ROCHA

Présentée et soutenue publiquement le 21 février 2022

Devant le jury composé de :

Hélène CHIAPELLO	IR	*Rapportrice*
Pierre PETERLONGO	CR	*Rapporteur*
Ingrid LAFONTAINE	PR	*Examinatrice*
Krister SWENSON	CR	*Invité*
Eduardo ROCHA	DR	*Directeur de thèse*

Remerciements

Cette histoire a commencé en l'an deux mille quatorze dans le Grand Nord Lillois, dans une équipe dénommée Bonsai. C'est là que j'ai rencontré 3 personnes très importantes dans ma carrière scientifique. JS, Aida et Samuel, je voudrais vous remercier pour la confiance que vous avez accordée à la petite débutant que j'étais, et de m'avoir donné goût à la recherche. Du campus de Villeneuve d'Ascq à Montréal en passant par Belo Horizonte (et Baisieux !), vous m'avez fait découvrir différentes facettes du monde de la recherche. Si je soutiens cette thèse dans quelques jours à Pasteur, c'est en grande partie grâce à vous, qui m'avez mis en tête l'idée de faire une thèse, et m'avez incitée à postuler à Pasteur.

Embauchée en tant qu'ingénieure de recherche au Hub, j'ai tout de suite été adoptée par une petite équipe qui porte bien son nom : GEM. J'ai ainsi eu la chance de rencontrer Eduardo, qui m'a proposé de faire une thèse en parallèle de mon travail d'ingénieure dans l'équipe. Je tiens à te remercier pour cette opportunité que tu m'as offerte. Tu as aussi su prendre en compte mes intérêts, et m'aider à construire un super sujet de thèse. Malheureusement, comme dans beaucoup de thèses, je n'ai pas eu le temps d'aller au bout. Mais ce n'est que partie remise, je reste au labo pour la suite !

Cette thèse n'aurait bien sûr pas été la même sans tous les collègues. Je tiens à remercier tout particulièrement Marie et Jorge, mes colocs du bureau 07g. Marie, merci de m'avoir fait faire mes tout premiers pas GEMiques, et pour ton entrain quotidien ! Jorge, encore félicitations pour ton poste confirmant le statut permanent du 07g, et merci pour ton apport de bonne humeur dans le bureau (et les pasteis de nata) ! Merci également à tous les autres membres de l'équipe, qui ont contribué d'une manière ou d'une autre à mon travail. Ils sont nombreux à être passés pendant ces années, et je ne peux donc pas tous les citer. Ces derniers mois, je tiens à remercier Eugen, Charles, Manuel, Olaya, Amandine, Jorge, Matthieu, Eloi et Fanny pour les repas conviviaux et animés, qui permettent de reprendre une bouffée d'oxygène avant de se replonger dans la rédaction. Un grand merci aussi à Brigitte, pour toutes les démarches et l'organisation, concernant ma thèse, mais aussi toute l'équipe en général. Tu vas nous manquer. Merci aux membres du Hub avec qui j'ai pu interagir, que ce soit pour des collaborations, des cours, ou tout simplement des déjeuners. Je tiens à remercier Bertrand pour son aide sur la diffusion de code : de longues discussions très intéressantes, autant techniquement qu'apnéiquement ! Merci aussi de m'avoir fait confiance pour donner les cours Python avec toi ! Merci aussi à maître Jedi (aka ./ed) de m'avoir initiée à ce langage qui m'a été indispensable pour ma thèse !

Je tiens aussi à remercier les membres de mon comité de suivi de thèse, Marc, Philippe et Krister, qui ont veillé au bon déroulement de celle-ci en faisant un point chaque année. Merci aussi à mon tuteur Pasteur, Stéphane. Merci pour tes échanges, tes retours d'expériences et tes encouragements.

Je peux maintenant répondre à la question que tu m'as posée à l'entretien d'embauche : c'était sur ça (voir pages *) ma thèse, avant mon "post-doc" ;)

Je tiens aussi à remercier (oui, ce début de phrase revient très souvent dans cette partie...) le « BIM Power » pour les nombreux échanges bio-info mais pas que, les sessions zoomiques depuis les 4 coins du monde, les animations pendant le confinement...et tout le reste ! C'est toujours un réel plaisir de recevoir de vos nouvelles. Un merci plus spécifique à Matthieu et Hélène pour leur soutien chaleureux.

Je tiens tout particulièrement à remercier Marion (même si tu vas me dire que tu n'y es pour rien alors que si) et Nathalie pour leur grand soutien et leurs encouragements tout au long de ma thèse. (Un petit coucou à Aurélien, qui a vu le jour pendant mes dernières semaines de rédaction !)

Comme toute thèse, les derniers mois ont été consacrés à l'écriture du manuscrit. Bien sûr, celui-ci n'aurait pas pu voir le jour sans l'aide précieuse d'Eduardo. Merci encore pour tes conseils, tes suggestions et corrections ! Merci également à Yoann, qui m'a beaucoup aidée pour certaines parties. Un grand merci à Maggie, Hélène et Yoann d'avoir accepté de relire des parties de mon manuscrit. Merci à Ginon, Sushi et Maki pour leur regard bienveillant pendant la rédaction, même si leurs retours ne m'ont pas beaucoup aidée. Merci à Affinity Designer de m'avoir permis de réaliser des figures... au grand désarroi du ventilateur de mon ordinateur (mais au moins j'ai eu un peu de chauffage...) ! Merci également à Manon d'avoir transformé mes vagues idées en 2 magnifiques dessins de phages (pour les voir, il faut lire la thèse ;)). Merci encore à mes parents, puis à Yoann (oui, encore toi) qui a pris le relais pour assurer toute l'intendance pendant mes derniers jours de rédaction.

Merci à Pierre et Hélène d'avoir pris le temps de relire attentivement mon manuscrit, et de l'avoir validé en l'accompagnant de rapports très intéressants. Merci à Ingrid et Krister d'avoir accepté d'être, avec Hélène et Pierre, dans mon jury. Krister, merci pour le prêt du pointeur "porte bonheur" lorsqu'on s'était croisés il y a 10ans ! Qui aurait deviné, à l'époque, que j'aurais fait une thèse, et que tu serais dans mon jury (oui, malgré les contraintes administratives, tu es entièrement membre de mon jury) ?! Enfin, si vous lisez ce manuscrit en version papier, c'est grâce à Christophe : merci !

Enfin, je veux (re)remercier tout particulièrement 4 personnes sans qui cette thèse n'aurait pas pu être menée à son terme.

Yoann, je t'ai déjà cité plusieurs fois, mais encore merci infiniment pour ton aide (et encore plus pendant les phases critiques), et pour le soutien que tu m'as apporté pendant ces années, et que tu continues à m'apporter encore aujourd'hui.

Mes parents, merci de nous avoir hébergés pendant le confinement, sans quoi je n'aurais pas pu avancer sur ma thèse. Mais surtout, merci infiniment (je me répète, mais je ne trouve pas d'autre mot) pour votre soutien permanent. Vous faites vraiment partie des personnes clé, sans qui je ne serais jamais allée au bout de cette thèse.

Enfin, Eduardo, sans toi, cette thèse n'existe tout simplement pas. Tu m'as apporté la possibilité de faire une thèse, tu m'as aidée à trouver un sujet adapté à mes envies, mais tu as surtout été là tout au long de celle-ci. Cette thèse a été freinée par une période très difficile de ma vie, mais tu as continué de croire en moi, et tu m'as poussée à continuer malgré tout. On m'a dit un jour que j'avais le meilleur des directeurs de thèse, et je veux bien le croire. Muito obrigada !

Contents

Part I

Large scale comparative genomics of bacterial genomes

Introduction

"Large scale comparative genomics of bacterial genomes": six words encompassing a wide range of subjects. This opening chapter relates, in a way, the story that leads to the birth of my PhD project. In order to make it accessible to non-biologists, I start from the very beginning... which corresponds to the two last words of the title: "bacterial genomes".

In this first part, Bacterial genomes, I focus on bacteria, from their external appearance to their innermost components, introducing the fundamental notion of *genome*. This is of course a short introduction to some notions that I use in the following parts, and in no instance a full course on bacteria.

The second part, Genome evolution, tackles the main mechanisms by which the previously described bacterial genomes evolve to adapt over time and environmental conditions.

The third and last part, Comparative genomics, combines biology with computer science and mathematics to introduce different Bioinformatics methods by which we can compare these evolving genomes. The increasing number of genomes available, now reaching hundreds of thousands, explains the last two words of this introducing chapter title: large-scale.

Providing unrivalled information, this huge amount of data nonetheless requires the development of methods handling extremely large datasets. The first methods, developed in the last two decades, have reached their limits both in terms of computation time and space. Generic and automatic methods able to reliably handle such amount of data in a reasonable amount of time have to be developed, which has spurred the birth of this PhD.

1 Bacterial genomes

In this first chapter, I will introduce the main protagonists of our story: *Bacteria*. I will start from a general overview of these living organisms as seen from the human perspective (part 1.1), and progressively zoom in at ever smaller scales until the most basic elements composing them (part 1.2). Then, I will make the link between the two scales, to understand how the basic elements can influence observed behaviors, called phenotypes (part 1.3). In doing so, I will define the basic notions and vocabulary of microbiology necessary to understand the rest of this manuscript.

Before going further, I want to clarify one point. The style of this first chapter is purposely not in a classical PhD thesis style. I know that reading a PhD manuscript is not particularly pleasant (if not boring), as it is a rather long document, on a very specific subject. However, it is a quite important document for its author and potentially for the scientific community, which deserves to be read. It is relating several years of work of a growing researcher. In order to lighten the reading, I added figures when possible, and wrote it in a more lively (and thus less formal) style. I hope this will make it a bit less painful to read.

While the main text contains the essential elements, some figures provide a few more details, for those who want a more advanced understanding of the mechanisms. However, if you are already familiar with terms such as (in random order, and non-exhaustive) capsule, genetic code, bp, binary fission, plasmid, replicon, translation, operon, cocci and nucleotide, you might want to skip this chapter. In that case, let's directly meet in chapter 2. Otherwise, let me introduce you to the fascinating bacterial world.

1.1 Bacterial ID card

1.1.1 Age, Address and Population

Inhabiting Earth for billions of years, these microscopic living organisms are ubiquitous. They are found almost everywhere: soil, ocean, rocks, on or inside other living organisms (including bacteria), and can even adapt to extreme environments. For example, some bacteria found in Arctic ocean waters can grow at very low temperatures, but die when temperature exceeds 20°C. On the contrary, other bacteria, living in hot springs, have adapted to live

with temperatures higher than 60°C. The number of bacteria on Earth nowadays has been estimated to be approximately 10^{30} [199]. Well, I know that this kind of numbers have no real meaning, as they are just unimaginable. Let's try to make an analogy.

I once saw this nice sandcastle Eiffel Tower (see on the right). Apart from the fact that there are bacteria living on sand, I guess you are wondering what is the connection with the total number of bacteria. Now, imagine that we make a full-size sandcastle of this same shape (to simplify, we consider it as a square pyramid of 324m high, with base sides of 125m). You can imagine this would require a huge number of sand grains. Considering that a grain of sand is a sphere of around 0.05mm of diameter (this is the diameter of small particles of sand in Sahara), this would require 10 billions times less sand grains than there are bacteria on Earth. In other words, each grain of sand would represent more bacteria than the total number of human on Earth! This is of course an approximation to give an idea of the huge number. The shape of the Eiffel Tower has been simplified, real sand grains are not all spherical and their size can vary of an order of magnitude.

Figure 1.1: Seen on Pinterest. Unidentified artist(s), but congratulations to him/her!

I agree that it is still unimaginable, but maybe a bit more manageable than a crude "10^{30}" number? Anyway, despite this huge population, the existence of bacteria was totally ignored before 1684, when Leeuwenhoek discerned 'animacules' through his homemade microscope on a sample of mouth biofilm [115].

1.1.2 Size

If a microscope was needed to see them, it is because bacteria are... microscopic. A typical bacteria like *Bacillus subtilis*, *Staphylococcus aureus* or *Escherichia coli* has a volume between 0.4 and 3 μm^3 [118]. If we go back to our grains of sand, the volume of a single grain is equivalent to the volume of more than 500 millions of bacteria taken together. In other words, more bacteria (of a "standard" size) than there are inhabitants in the USA fit in a single sand grain.

However, the size of bacteria varies by many orders of magnitude. The smallest bacteria observed so far, which are marine ultramicrobacteria called *Candidatus Actinomarina*, have an average volume of 0.013 μm^3, which is 100 times smaller than *E. coli*. In contrast, giant bacteria *Thiomargarita namibiensis* can be up to 0.22 mm^3, being eight orders of magnitude bigger than *E. coli*, or in other words, in the same order of magnitude as our grain of sand. Obviously, like sand, these giant bacteria can be seen by the naked human eye. Levin et al. illustrate the difference of sizes between the smallest and the biggest bacteria as being equivalent to the difference of sizes between a mouse and the Empire State Building [118].

1.1.3 Relationships

As already mentioned, bacteria are found on or inside other organisms, including for example plants, animals, but also other bacteria. This close relationship between bacteria and another organism, called *symbiosis*, can have advantages and disadvantages for both the bacterium and its host.

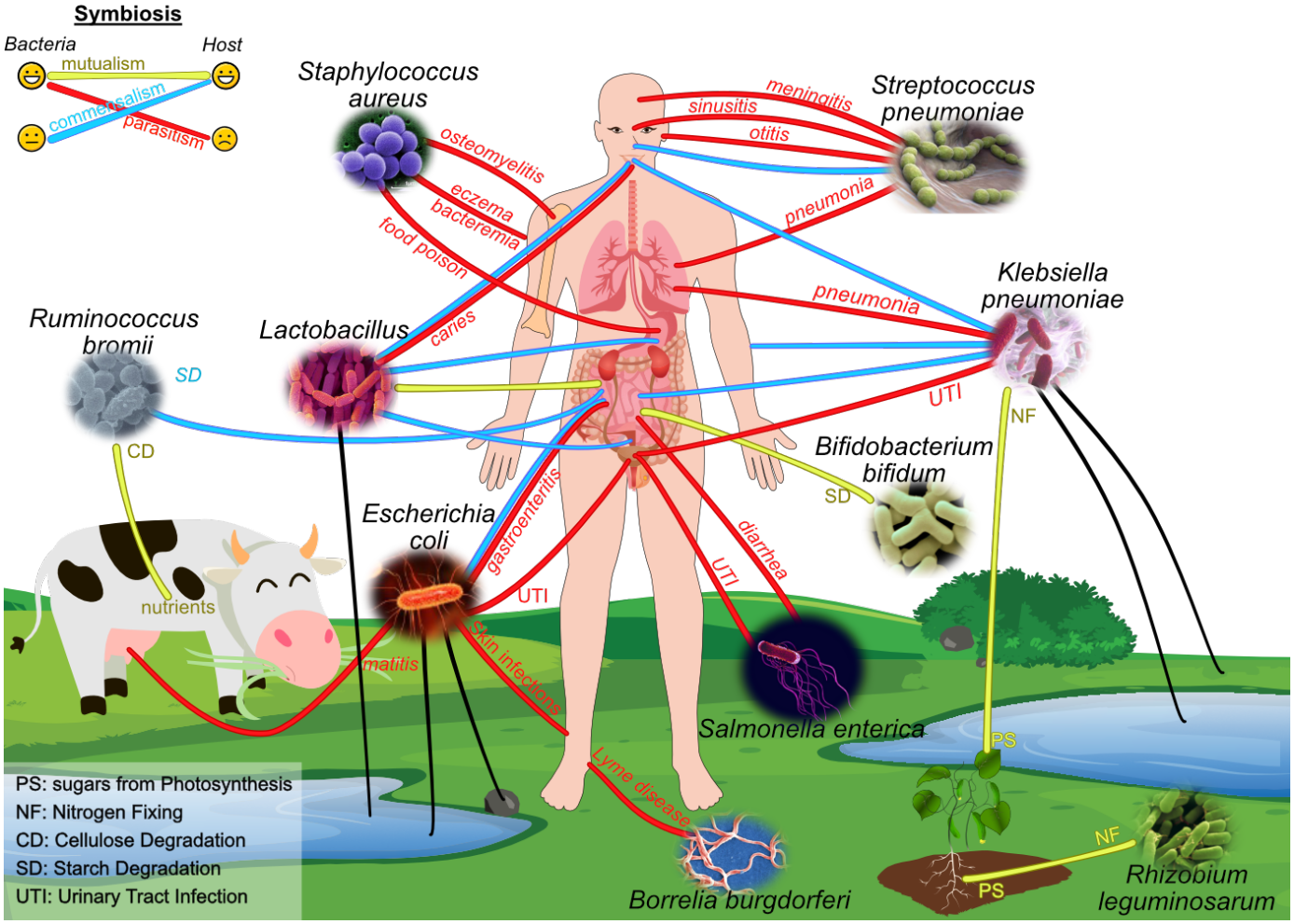


Figure 1.2: A few examples of the complex symbiosis relationships between some bacteria and other living organisms (human, animals and plants). Black arrows show examples of environmental habitats of several bacteria.

For instance, although they are invisible to the naked human eye, bacteria are generally seen as the "bad guys". And indeed, we cannot deny it: nobody is spared by diseases caused by bacteria. Even if some of them, like pneumonia, meningitis or diphteria, are less common, I would not believe that you never caught any bacterial infection. What about sinusitis, otitis, or urinary tract infections? Or maybe gastroenteritis, food poisoning and/or diarrhea? Well, they definitely deserve some place in our worst memories and nightmares. This is the

parasitism side of *symbiosis*: bacteria take benefit from their host(s) to the detriment of the latter (see fig.1.2).

But, hopefully, not all bacteria are (always) like that. Even if we ignore it (or tend to), bacteria are also more than our best friends: we could not even live without them. For instance, we could not breathe if cyanobacteria had not played a crucial role in the production of oxygen in the atmosphere [161]. We could not assimilate most food if some bacteria, our gut microbiome, were not continuously helping us with digestion. Like so, we host approximately as many bacteria in our gut than we have human cells in our whole body. These *commensal* (we take benefit without penalizing them) and *mutualist* (mutual benefit) residents can weight up to 200g [164].

Also, less vital but still enjoyable, bacteria are used in food industry to make many key ingredients. They are for example helpful to convert milk into yogurts, cheese or cream by fermentation. Yes, you can even thank bacteria for your ice-cream!

I must clarify one last thing: a microbe is not necessarily a bacteria. *Microbe*, meaning micro-organism, is a generic term to define any living organism too small to be seen with an unaided eye. Together with bacteria, protists (like *Plasmodium*), fungi (like *Candida*), some plants (like some green algae) and even micro-animals (like dust mites) are part of the microbial world... and are also responsible for several diseases (see figure 1.3)! Bacteria are far from being the only disease-causing micro-organisms. Malaria, mycosis, toxoplasmose, but also some pneumonia and otitis are caused by other types of microbes. Besides, although their "living organism" status (and consequently their affiliation to microbes) is controversial, we cannot omit viruses (like the too famous SARS-CoV-2 causing covid-19, but also common flu viruses) from our list of agents of infectious diseases. Finally, infectious diseases can even be caused by non-living molecules: Creutzfeldt–Jakob disease is caused by a prion, which is no more than a misfolded protein.

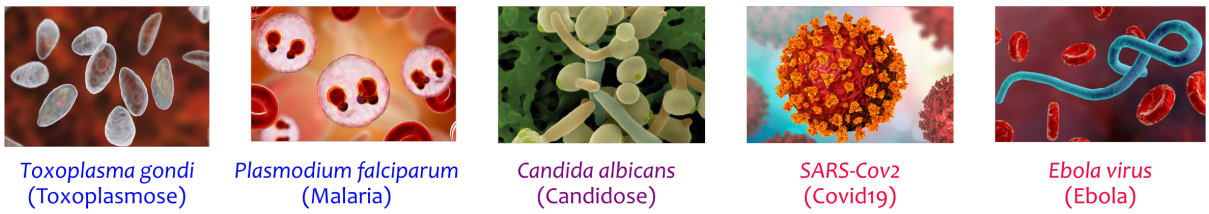


Figure 1.3: Different types of microbes others than bacteria: protists, fungus (including yeast), viruses. Pictures from science photo library

Whatever their nature, these other pathogens do not seem really nicer than bacteria, do they?

So, I hope I reconciled you, at least a little, with bacteria, because they will be with us in the whole manuscript.

1.1.4 ID photo

I previously mentioned that Leeuwenhoek saw bacteria with his microscope. But what did he see exactly, what does a bacterium look like? Concretely, he saw, or rather barely distinguished nothing more than little rods. And many scientists observed the same after him, to the extent that in 1838, Ehrenberg gave them their name *bacteria*, from the Greek $\beta\alpha\kappa\tau\rho o\nu$ meaning stick.

Later, improvements not only in microscopy but also in laboratory techniques unveiled a wide diversity of bacterial shapes. Indeed, on top of the low magnification of microscopes at that time, bacteria were also difficult to observe because they blend in with the tissue cells. In 1884, Hans Christian Gram, who was working on lung tissue from patients suffering from pneumonia, developed a staining method aimed at bacterial cells to make them more visible. During his experimentation, he discovered that some bacteria retain the stain even after decolorization, due to the characteristics of their cell wall [75]. His method, originally designed for observing bacterial shape, turned into a routine laboratory technique which has been used for decades to quickly classify bacteria into two main groups: Gram-positive (hereafter Gram+, retaining the stain) and Gram-negative (Gram-) bacteria (see figure 1.5).

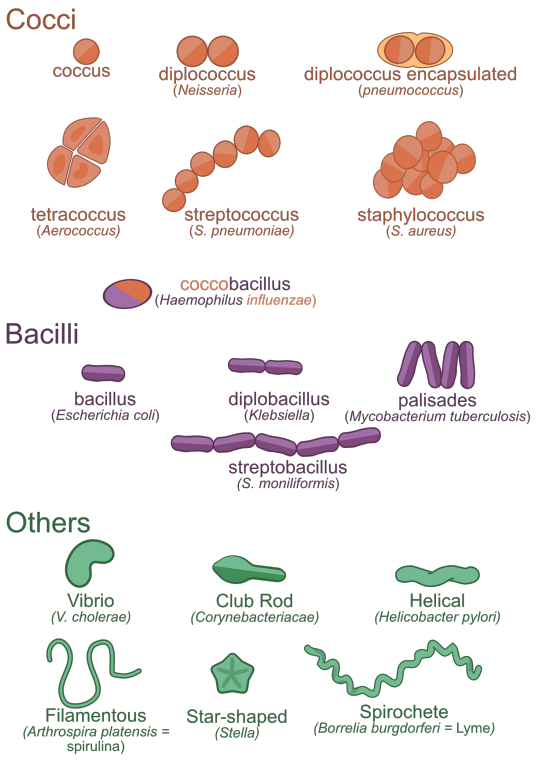


Figure 1.4: Different morphologies of bacteria, with example of a species for each one.

Better conditions for the observation of bacteria thus allowed researchers to distinguish many different shapes. The most common are spherical (*Cocci*) and rod-shaped (*Bacilli*) bacteria, but they can also be spiral-shaped (*Spirochaetes*) or comma-shaped (*Vibrio*). Some have fancier shapes like corkscrews, helices or stars, and others can even exhibit different forms over time (pleomorphic).

Beyond their individual shape, some bacteria also have specific group arrangements. For example, some bacteria live in pairs (*Diplo*), some grow in chains (*Strepto*), and others group together in grape-like clusters (*Staphylo*). Observed morphologies are thus a combination of an individual shape and an arrangement. For example, *Streptobacilli* are chains of rod-shaped bacteria, while *Staphylococci* are groups of spherical bacteria.

Figure 1.4 gives an idea of this diversity.

Zooming in a little more on the cell membrane reveals the presence of appendages on some bacteria (see figure 1.6). Long hollow tube-shaped appendages called *flagella* are mostly dedicated to bacterial motility. Some bacteria also have *pili* or *fimbriae*, smaller hair-like appendages used for cell adhesion. They are, for example, responsible for the formation of biofilms when attaching on infected host surfaces [148]. They are also involved

in the exchanges of DNA by attaching to other bacteria, as we will see in 2.2.1. In that case, they are called *pili* or *sexpili*.

However, observing bacteria is not enough to explain their behavior. Neither their morphology (shape, appendages, arrangement) nor their chemical characteristics (Gram stain) can give us a clue on how the bacterium will behave. For example, pneumonia can be caused by *Streptococcus pneumoniae* (chain of spheres, Gram+), *Haemophilus influenzae* (coccobacilli, Gram-) or *Klebsiella pneumoniae* (rods, Gram-). Conversely, the same bacterium can cause multiple diseases. For example, *Streptococcus pneumoniae* is responsible for pneumonia, but also for otitis or sinusitis [31]. Other examples are showed in figure 1.2. Even more stunning, the behavior of a bacteria evolves over time. For example, *E. coli*, a widespread commensal inhabitant of our intestine, can sometimes turn into a dangerous pathogen in this same environment, causing various forms of diarrhea, or even sepsis. It is also responsible for several extra-intestinal diseases like skin infections or urinary tract infections (see figure 1.2) [49]. Another example is the acquisition of antibiotic resistance by bacteria which were previously sensitive to these same drugs.

If the external properties like morphology are not enough to understand the different phenotypes of bacteria, we can suggest that the latter are the result of events happening inside the bacterium. So, let's cross the membrane and explore the inside of bacteria!

1.2 Inside the bacterium

Well, wait a little... Going inside a bacterium is not always that easy! Some bacteria are protected by a capsule, a solid layer of polysaccharides, which are complex sugars (like starch in potatoes or cellulose in vegetables) (see figure 1.6). They are tightly packed according to a well-organized structure which makes it hard to wash away the capsule, and provides a real shield to the bacteria.

Behind the capsule, the bacterial cell wall also has different architectures defining two types of bacteria (see figure 1.5). Monoderm bacteria have a cytoplasmic cell membrane surrounded by a thick cell wall composed of peptidoglycans. As the latter often retains the Gram stain, monoderm bacteria are often assimilated to Gram+. On the contrary, diderm bacteria have two cell membranes (cytoplasmic and outer membrane), but with a thin cell wall in between. The thinness of the latter makes it permeable to the Gram stain, such that diderm bacteria appear in pink on Gram-stained culture, and are often assimilated to Gram- (see figure 1.5). However, some monoderm bacteria can also appear in pink on Gram-stained cultures. For example, *Mycoplasma* bacteria lack a cell wall around their single membrane, and thus do not retain the Gram stain. On the other way around, *Deinococcus* bacteria have a thick cell wall between their two membranes, making them retain the Gram stain (Gram+). Nowadays, bacteria are rather distinguished by the nature of their cell envelop than by their Gram-staining response. This has been shown to be more relevant to explain the evolution of bacteria over time [83].

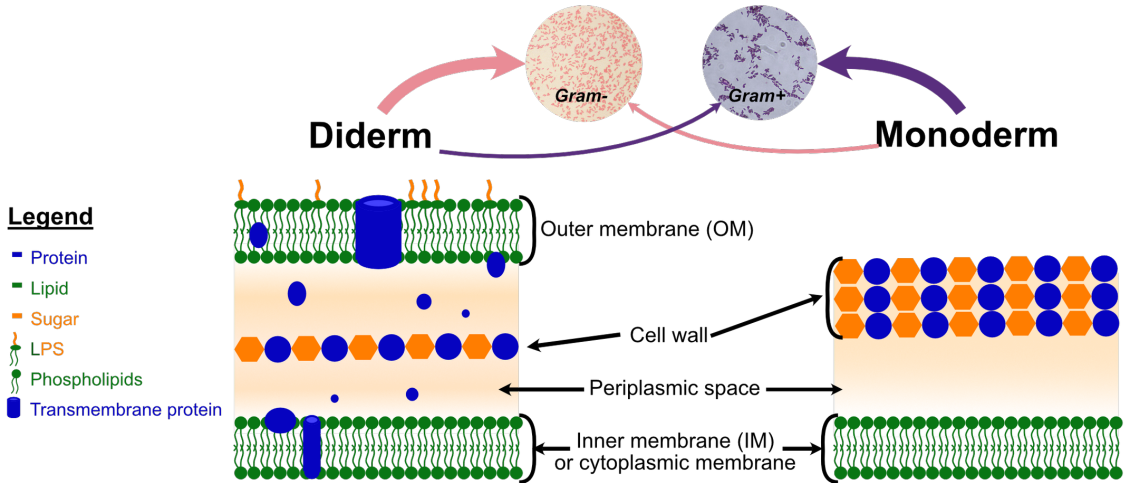


Figure 1.5: Typical cell envelop composition of Monoderm and Diderm bacteria. Because of variations in monoderm and diderm structures, the association monoderm/Gram+ versus diderm/Gram- is not always true. Gram-stained culture photos are from CDC and Riraq25/Wikimedia. LPS means lipopolysaccharides.

The cell wall has a role in protecting the bacterium from its environment, but also in giving the bacterium its shape (rods, spheres...). For information, peptidoglycans (cell wall shaping bacteria) are composed of peptides (parts of proteins) and sugars. The outer membrane of diderms often contains lipopolysaccharides (LPS) in its outer leaflet. These are composed of lipids (fatty acids) and... sugar again! Who would have thought? With their cell wall and capsule, bacteria are sometimes sweet in both senses of the word!

1.2.1 Swimming in the cytoplasm

Now that we have passed through the cell walls, we can finally dive inside the bacterium. The most striking thing you would notice while entering this gel-like matrix called *cytoplasm* is the simplicity of the structure of the organism. Like so, bacteria belong to *prokaryotes*, unicellular organisms devoid of any internal membrane-bound compartment (as opposed to eukaryotes which contain several specialized sub-units called organelles, like the nucleus). All steps for bacterial replication and growth occur in the same medium (see figure 1.6).

Despite I said there are no compartments, there is still a condensed irregularly shaped region in the cytoplasm. It corresponds to the main genetic material of the bacterium, which is localized (but not separated by any membrane) in this small region called *nucleoid*.

Apart from this localized genetic material, the cytoplasm contains enzymes, gases, nutrients, and more complex structures like ribosomes (used for protein synthesis, see 1.3.2) and plasmids (small extra-chromosomal DNA molecules). Recently, small organelles (like carboxysomes used for carbon fixation) have also been observed, but only in specific bacteria [33].

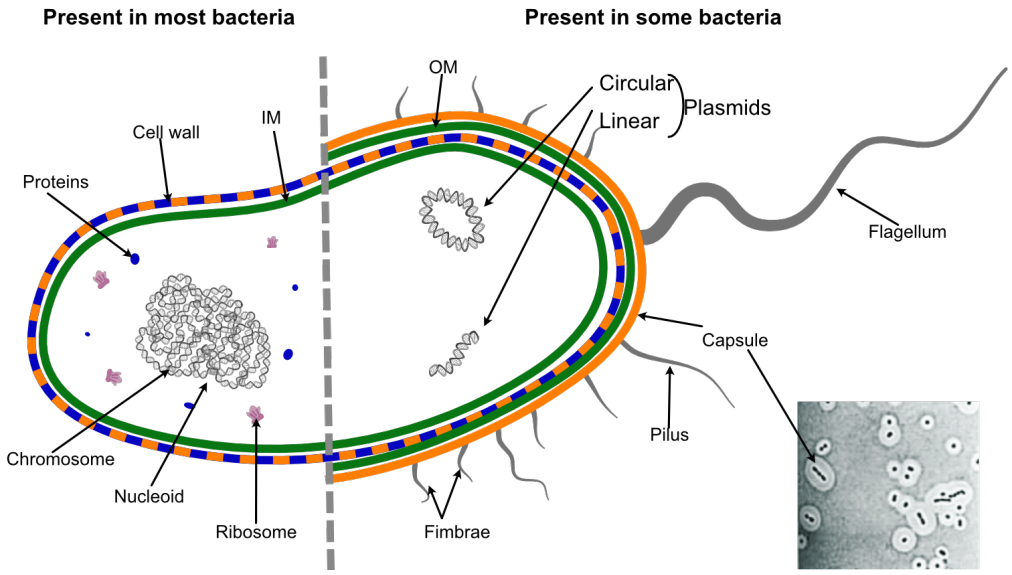


Figure 1.6: Inside the bacterial cell. Colors used are the same as figure 1.5: sugars, proteins, and lipids. Figure adapted from biologyease.com, with a photo from [179].

1.2.2 Genetic material

Let's zoom in one last time, to see how the genetic material we saw in the nucleoid is made. Bacteria have most of (or sometimes all) their genetic material in a single circular DNA molecule called *chromosome*. DNA (Deoxyribonucleic acid) is composed of four different nucleotides (see figure 1.7): Adenine, Thymine, Cytosine, Guanine. A small detail: each nucleotide is actually composed of a base, a phosphate group and... a sugar (deoxyribose)! In practice, the term *nucleotide* is used interchangeably with *base*. The most stable form of DNA, called double-stranded DNA (or dsDNA), is a double helix made of two chains (strands) of nucleotides coiled around each other according to pairing rules: A with T and C with G. The latter was put forward by Watson and Crick in 1953, based on Rosalind Franklin experimental observations [197]. This 3D structure with paired bases is essential for most cellular functions (like protein synthesis or bacterial reproduction, as we will see in part 1.3).

The size of a dsDNA molecule is given in base pairs (bp), corresponding to the number of bases in one strand. Due to the asymmetry of deoxyribose sugar, each DNA strand has two distinct extremities (3' and 5', see figure 1.7), imposing a direction for many mechanisms, as we will see in part 1.3.1.

Many bacteria host, in addition to their chromosome, one or several extra-chromosomal DNA molecule(s) (mostly circular, but sometimes linear) like *plasmids* or secondary chromosomes [128] [87] (see figure 1.6). All these molecules contain genetic information, instructions necessary for all cellular functions, from basal functions like growth to specific ones like molecule secretion.

A typical bacterial genome, like *E. coli* K12, contains around 4.6Mbp. However, this size is highly variable, from a hundred kbp (like *Candidatus Nasuia*) to more than 14Mbp (like

Sorangium cellulosum) [178] [109]. This includes the potential plasmid(s), which range from 2kb to up to 10% of their host chromosome [16]. However, even those particularly "large" genomes remain relatively small compared to most eukaryotic genomes. For example, the human genome is around 3300Mbp.

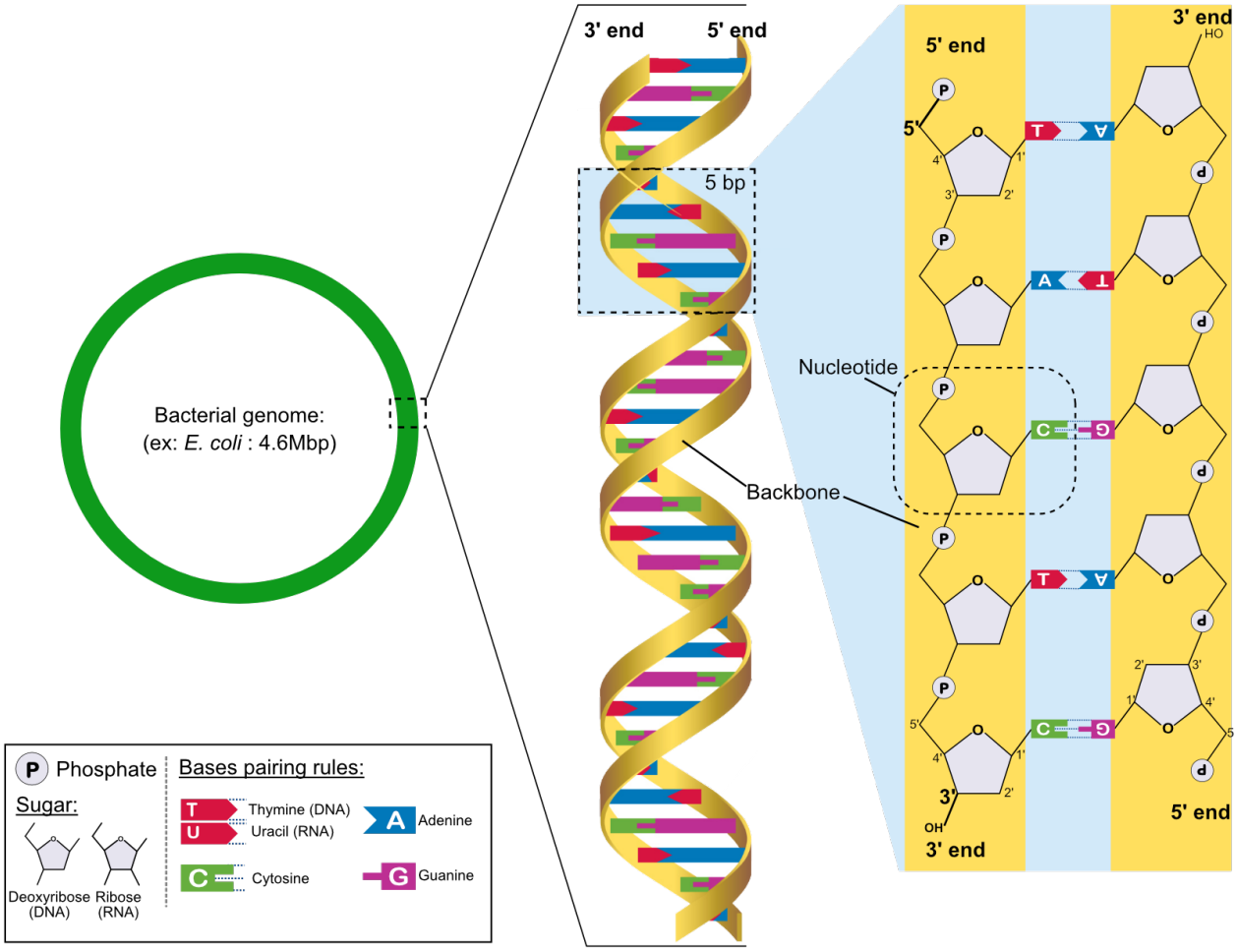


Figure 1.7: **Detailed composition of nucleotide-based molecules found in bacteria: DNA and RNA** (Deoxy)ribose sugars have five carbons, numbered from 1' to 5'. Phosphate group is bound to carbon 5', and carbon 3' binds to the phosphate of the next nucleotide: molecules are oriented from 5' to 3'. A double-stranded DNA consists of antiparallel strands of nucleotides held to one another according to the pairing rules. RNA molecule differs from DNA in 3 ways: it is single stranded, the thymine base is replaced by uracil, and the nature of the sugars is different. All together, these DNA molecules form the bacterial *genome*. Figure adapted from Encyclopaedia Britannica.

Now that we have seen the smallest components (bases) of bacteria necessary to understand the rest of the manuscript, we can take a very little step back. A DNA molecule is composed of several *genes*, DNA sequences of typically 1000 nucleotides carrying information to make specific products. The latter can be proteins or RNA (see figure 1.9B), two essential

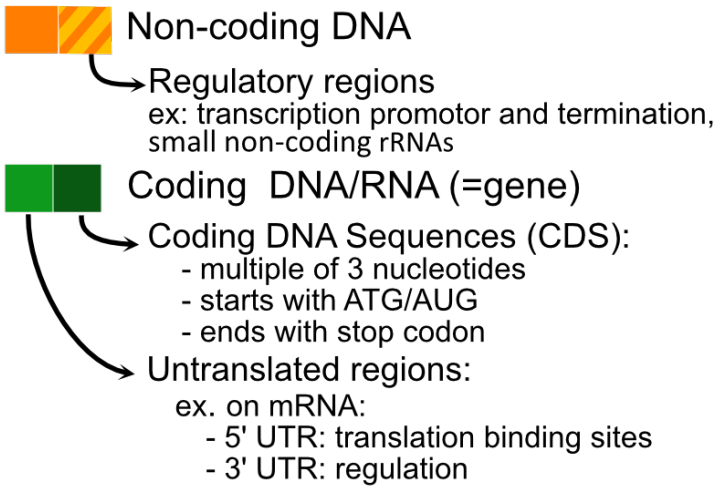


Figure 1.8: Color code common to figures 1.9, 1.11 and 1.12.

molecules that will be described in the next parts.

While the chromosome carries the essential house-keeping genes, plasmids encode dispensable (but sometimes beneficial for the host) genes. For example, antibiotic resistance genes are often encoded on plasmids rather than on the chromosome [16].

Unlike most eukaryotes, bacteria have a very high density of genes in their genome (see figure 1.9A), with in average more than 87% of coding content (compared to 2% for some eukaryotes like humans) [109].

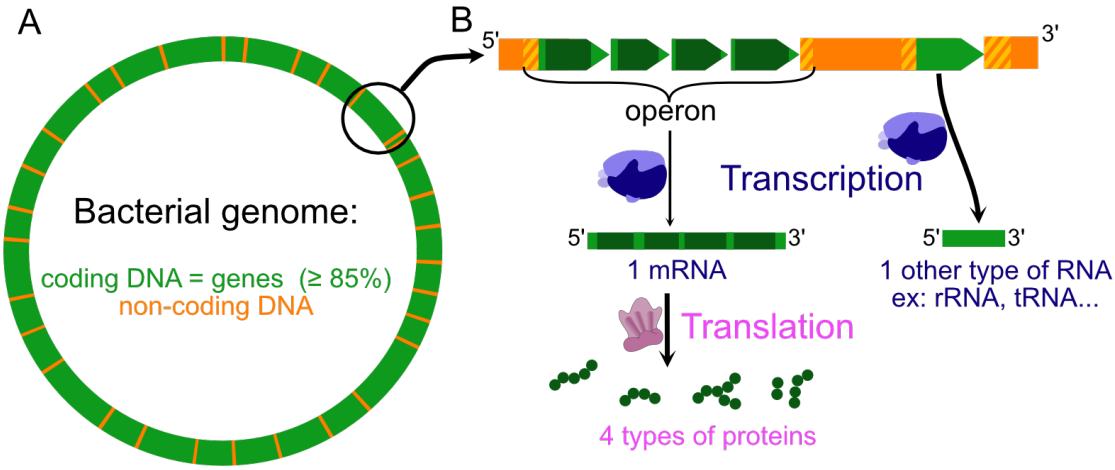


Figure 1.9: A. Bacterial DNA at the gene scale (for details at the nucleotide scale, see figure 1.7). B. Operon organization of bacterial genes (see part 1.3.4). The mechanisms for transcription and translation are detailed in figures 1.11 and 1.12 respectively. For color code, see legend in figure 1.8.

1.2.3 Proteins

Apart from the genetic material, we observed many proteins scattered throughout the cytoplasm. Together with nucleotides (DNA and RNA), fatty acids and sugars, proteins are major types of molecules essential for life.

Products of gene expression, they are made of chains of amino-acids (aa). There are 20 standard amino-acids used by all living organisms (several more have already been observed in some bacteria [5]). Each amino-acid is coded by a codon, a group of three nucleotides.

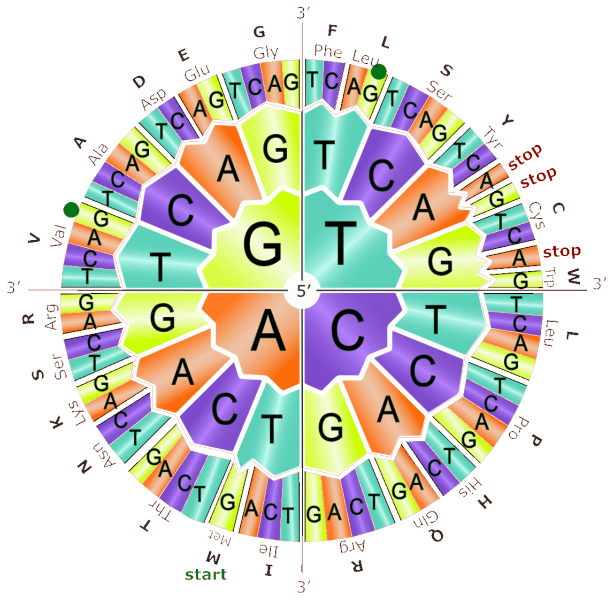


Figure 1.10: The "universal" genetic code employed by most organisms. Green dots correspond to alternative start codons used by some bacteria. Picture adapted from Wikimedia Commons.

The universal genetic code (see figure 1.10) determines which amino-acid corresponds to each codon of nucleotides. As there are 64 different triplets of nucleotides but only 20 amino acids to encode, the genetic code is degenerated: a given aa can be encoded by several codons. On the other hand, each codon encodes a unique aa (except three which are stop codons). This property is very important for genome analyses, as some species preferentially use one specific codon for a given amino-acid.

As we said, genes are DNA sequences carrying information to make a product. When this product is a protein, the portion of the gene coding for the protein itself is called *CDS*, for Coding DNA Sequence. Except for particular cases out of our scope, this coding region must fulfill specific conditions: have a multiple of 3 nucleotides, start with a start codon (most of the time ATG, but sometimes GTG or TTG for some bacteria [14]), and end with a stop codon (TAG, TAA or TGA) (see gene and CDS in figure 1.8 and 1.9).

Proteins affect the bacterial phenotype at different levels. Some of them directly contribute to cell morphology (like flagellin forming the flagellum), while others take part in complexes of proteins.

Some of them can be used by the bacterium to harm its host. For example, some *Vibrio cholerae* strains are able to synthesize the different proteins forming the cholera toxin, causing the typical watery diarrhea of Cholera infection [196]. In addition to damaging the host, some bacteria also produce proteins to protect themselves. To take a trendy example, some bacterial genomes contain *bla* genes, which code for β-lactamases proteins. Interacting with other elements, these enzymes break the structure of β-lactam antibiotics (such as penicillins), preventing them from harming the bacterium, and thereby making it resistant to these antibiotics [153].

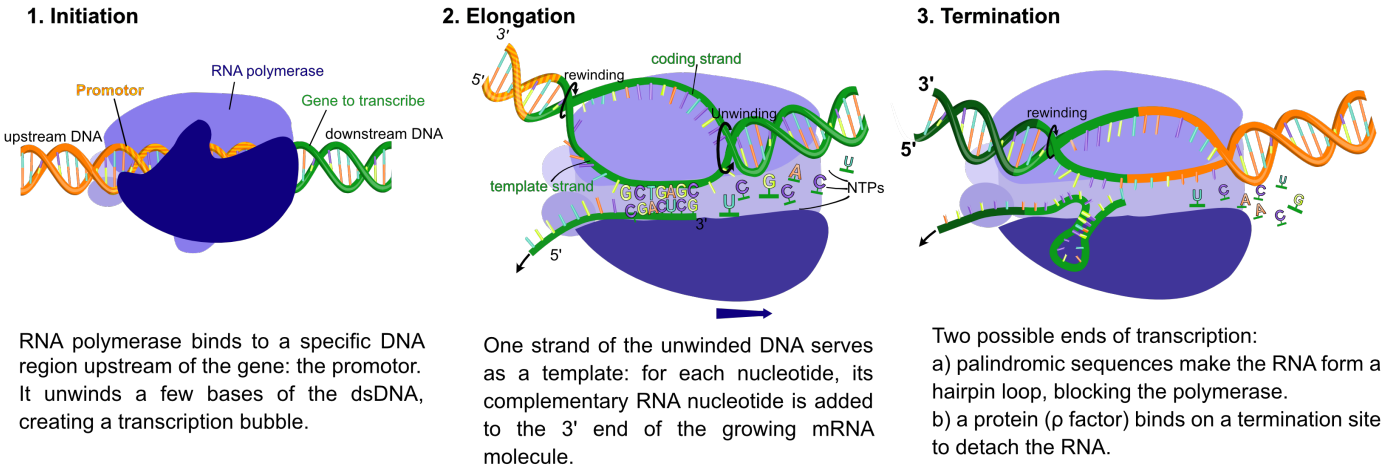
Another example, shared by all living organisms, is that the mechanism used to synthesize proteins is itself dependent on several proteins. Speaking about this mechanism, you might understandably wonder how a DNA sequence can be transformed into a protein. This is precisely the purpose of the next chapter.

1.3 From DNA to phenotype

Synthesis of proteins starting from a DNA gene is not straightforward. It requires two main steps, which are common to all living organisms: transcription and translation (see figure 1.9B). Some organisms (mainly eukaryotes) use additional mechanisms, resulting in the same gene potentially coding for many different proteins. However, as bacteria usually do little more than those two main steps, I will not tackle the others in this manuscript. I will just very briefly describe the two main processes, in order to introduce some vocabulary needed thereafter.

1.3.1 Transcription

The gene is first copied (or transcribed) into a short-lived RNA molecule, called messenger RNA (*mRNA*). The mRNA molecule differs from DNA in 3 ways: it is single stranded, the thymine base is replaced by uracil, and the nature of the sugars is different (ribose vs deoxyribose) (see figure 1.7). During transcription, one strand of the bacterial dsDNA is used as a template, and is complemented by RNA nucleotides (present in the cytoplasm as Nucleoside Triphosphate (NTPs) according to the previously cited pairing rules (A with U or T and C with G). If you want, you will find more details on this process in figure 1.11.



RNA polymerase binds to a specific DNA region upstream of the gene: the promotor. It unwinds a few bases of the dsDNA, creating a transcription bubble.

One strand of the unwinded DNA serves as a template: for each nucleotide, its complementary RNA nucleotide is added to the 3' end of the growing mRNA molecule.

Two possible ends of transcription:
a) palindromic sequences make the RNA form a hairpin loop, blocking the polymerase.
b) a protein (ρ factor) binds on a termination site to detach the RNA.

Figure 1.11: Transcription of a gene into a mRNA. For color code of DNA/RNA molecules, see figure 1.8.

1.3.2 Translation

Barely transcribed, the mRNA already serves as a template for the second step of protein synthesis: translation. This step takes place inside of structures called *ribosomes* (see figure 1.6), which are themselves composed of (r)RNA and proteins. The mRNA is read by the ribosome, and each codon is translated (giving its name to the process) according to the genetic code: the corresponding amino-acid, brought by a tRNA (transfer RNA), is added to the growing protein. More details on this process are provided in figure 1.12. Due to its single-stranded state, the mRNA is quickly degraded by specific enzymes (RNase) after a few rounds of translation.

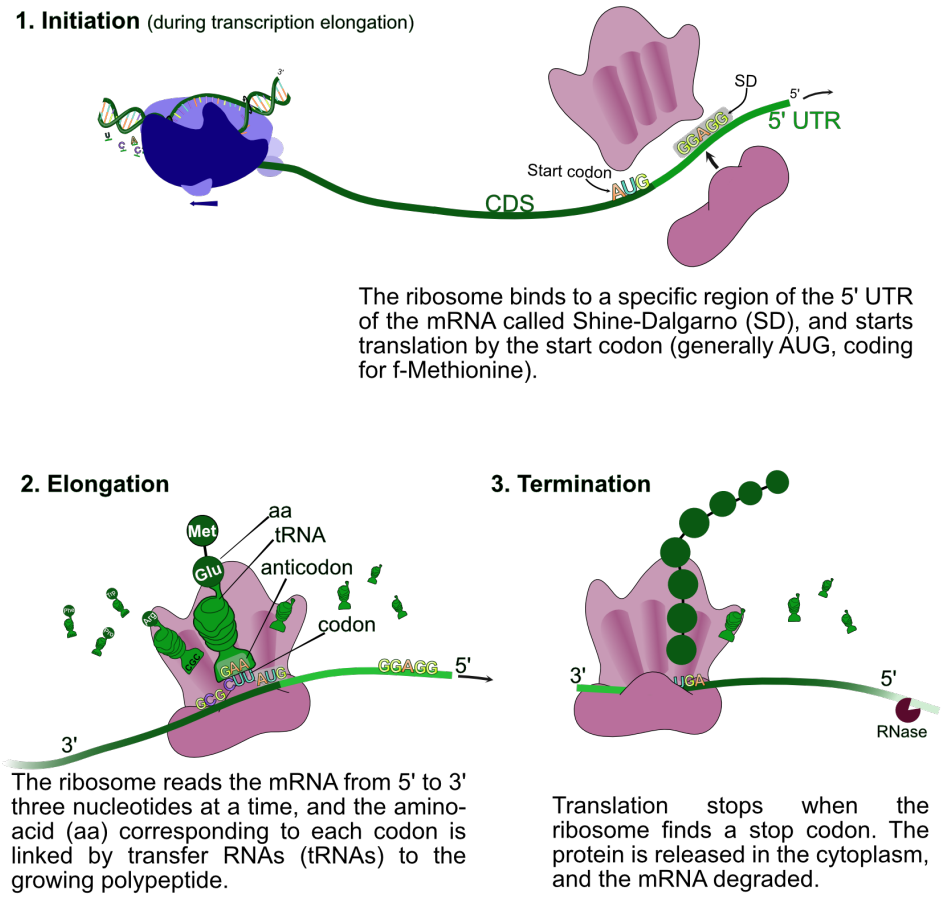


Figure 1.12: Translation of a mRNA into a protein. For color code of DNA/RNA molecules, see figure 1.8. Ribosome and tRNA shapes from Wikipemedia Commons.

Let's go back to our bacterium secreting the Cholera toxin. We now understand how this protein is synthesized from the bacterial gene. However, given the size of the human intestine, if only one bacterium was producing this toxin, it would not cause significant damage. So, it is quite obvious that in a sick intestine, there is not only one, but a huge quantity of bacteria having these same genes. But how do they manage to have such a huge population with the

exact same genes? In other words, how do bacteria reproduce?

1.3.3 Bacterial reproduction

Just like their internal structure, many bacteria have quite simple life cycles. They use an asexual mechanism called *binary fission*: they first grow, copy their genetic material (replication), and divide into two daughter cells, each one having one copy of the genome (duplication). In most cases, the two daughter cells are identical, and are ready to re-initiate a replication step. However, some species like *Caulobacter crescentus* asymmetrically divide, leading to two cells physiologically and morphologically distinct: one can directly replicate whereas the other has to go through a differentiation step before [170]. Other species have even more complex life cycles, involving several development phases for each cell. One of the best known is *Myxococcus*, with its well studied fruiting-body state [166].

Population growth is generally characterized by the generation time or doubling time (DT), which corresponds to the time required for one division. It varies widely among bacteria and according to the conditions (e.g. temperature, oxygen, nutrients or nature of environment). For example, in optimal growth conditions, a typical *E. coli* strain has a DT of 20 minutes, while *Syntrophobacter fumaroxidans* divides every 140h [72].

For the first step, replication, there are several different mechanisms. The most common, sometimes called *Theta replication*, starts at a single site on the chromosome (locus called *Ori*), and then proceeds in both directions in parallel, until the two replication forks meet at the terminal site (see figure 1.13 for more details). For example, the main chromosome replicates by this mechanism. Incidentally, a DNA or RNA molecule able to autonomously replicate is commonly called *replicon*. A bacterium can have several replicons: its chromosome(s) and its plasmid(s).

DNA replication must be initiated as often as the cell divides. In rapidly growing bacteria, a new round of chromosomal replication begins before an earlier round is even completed, resulting in nested replication bubbles. Other species, like *Caulobacter*, have a strict once-and-only-once replication behaviour [170].

Another common mechanism is *Rolling circle* replication. This system, much simpler, is particularly used for plasmid replication. It only replicates one strand at a time [157].

During both types of replication, new DNA is synthesized thanks to polymerases. Many of the latter self-check the base they just added, and immediately replace it if they detect a mis-pairing. However, a few errors can still slip through this *proofreading* system. To rectify this, the *mismatch repair system* (MMR) performs a second check right after replication [69]. In addition to replacing wrongly paired bases like proofreading process, it can also correct some small insertions or deletions that may happen if the polymerase slid on the template strain. MMR is also used to repair some DNA damages, and a defecting MMR system can have high impacts for the bacterium (see 2.3.1).

After the end of replication of all replicons, and before cell division, the latter must be partitioned. Like so, each daughter cell receives one copy of each replicon, leading to identical daughter cells.

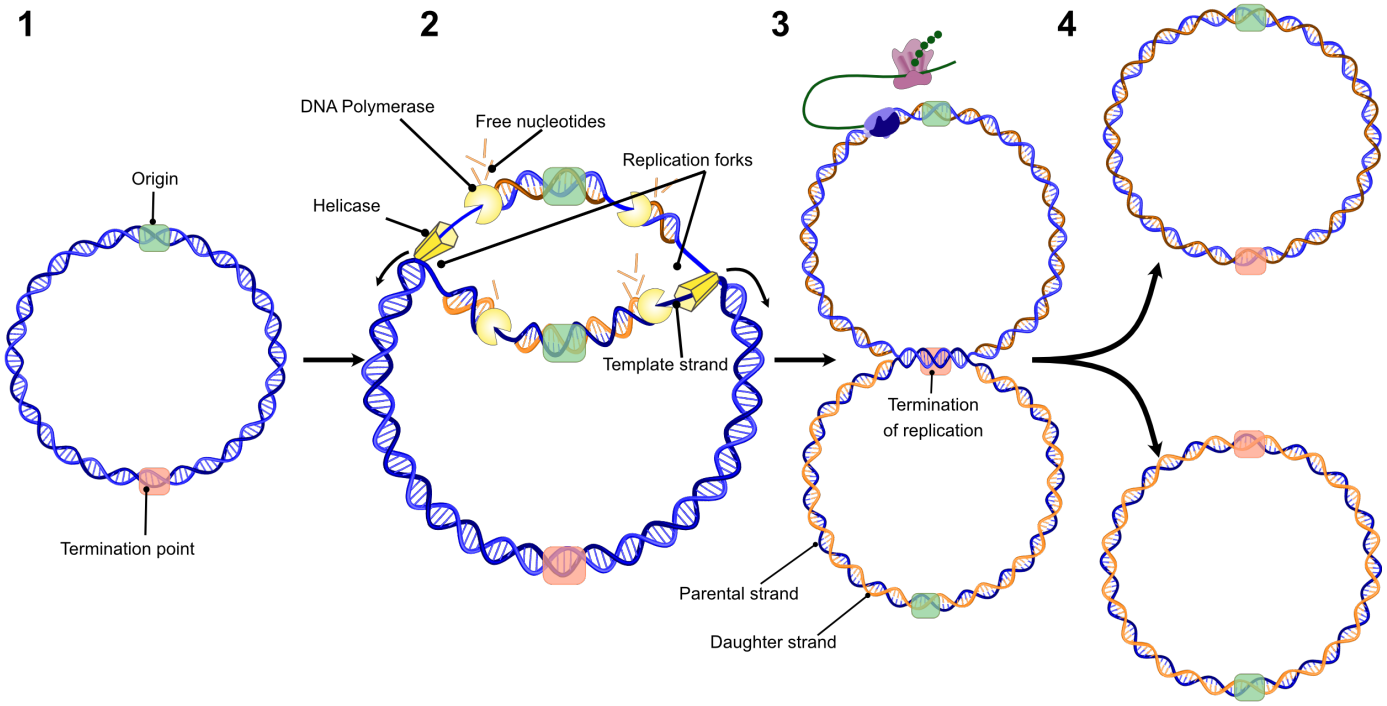


Figure 1.13: Replication of bacterial DNA, also called *Theta replication*. 1 Initiator protein binds on a particular locus of the chromosome called *Ori*. 2 The dsDNA is unzipped by a DNA helicase, creating two forks of replication in opposite directions. Each strand serves as a template to synthesise a new complementary DNA strand. A "proofreading" process checks each newly added base. 3 Replication stops at the *ter* locus, leaving two interlinked circles. 4 Both replicons are separated by a type II topoisomerase. Figure adapted from figures 7.6 and 7.7 of [12].

1.3.4 Organisation

To allow fast growth, these different processes (replication, transcription, translation) occur at the same time, as showed in figure 1.13 step 3. Moreover, bacterial genomes are highly organized. For instance, most bacterial genes are grouped in *operons*, condensed arrays of genes (usually sharing complementary functions) transcribed as a single unit (see figure 1.9B). One of the most famous operons is the *lac* operon of *E. coli* [96]. The expression of the genes of this operon allows the bacterium to assimilate lactose. However, if lactose is not present in the environment, bacteria do not need these genes to be expressed... and they actually do not express them. Indeed, bacterial genomes have a powerful system of regulation: the expression of operons and individual genes is regulated by other genes called regulatory elements, themselves optimized [149]. Like so, a given regulatory element can regulate several genes, being on the same DNA molecule (*cis regulation*) or on another DNA molecule via an intermolecular interaction (*trans regulation*). Conversely, a gene can depend on several regulatory elements.

The genome is also highly organized in relation to the replication mechanism. For ex-

ample, as mentioned before, fast-growing bacteria usually have nested replication bubbles, each one being translated at the same time. Like so, genes near the origin of replication are over-expressed compared to the others. The distribution of genes along the chromosome follows this characteristic: highly expressed genes tend to be near the origin of replication in fast growing bacteria (those with multiple replication bubbles) [155].

1.4 Classification

There is no consensus on how to class bacteria. They can be classified according to their pathogenicity, their membrane structure (mono/diderm), their growth temperature range or any other trait like those tackled in parts 1.1.4 and 1.2. However, the names given to bacteria are regulated by the International code of nomenclature of prokaryotes (ICNP) and indexed in the official LPSN database (List of Prokaryotic names with Standing in Nomenclature) [138] [58]. Bacterial names follow a rank-based classification, or taxonomy, a system proposed by Carl Linnaeus to classify all living organisms in the 18^{th} century. This classification is based on a wide range of criteria. Those include morphology, Gram-staining, physiology (temperature of growth, need of oxygen etc.) but also genetics (DNA properties, proteins synthesized etc.) and, above all, thanks to the sequencing technologies, genome comparison.

Along with Eukaryota and Archea, Bacteria constitute one of the three domains (the highest taxonomic rank) of the famous "tree of life" devised by Woese et al. in 1990 [202]. Figure 1.14 shows one of its most recent version, with details for the bacterial branch. Each domain is then divided into hierarchical ranks: phylum, class, order, family, genus, species. Bacteria are commonly named using their genus and species name.

However, for bacteria, this taxonomy is still controversial in the microbiology community. And I did not know how right I was by the time I wrote the previous sentence: a month later, an emendation of some rules of the ICNP regarding the phylum rank was published [136]. This means that maybe, if you are reading this manuscript many years after 2022, you may not recognize the names I use. Do Firmicutes and Bacteroidetes still exist?

Beyond the name, the definition of the different ranks, and in particular that of a *species* remains an ongoing debate. This will be quickly tackled in chapter 3.3.

For now, let's make an assessment on what we have seen so far. We know how the bacterial genetic material (its genome) affects its phenotype, how genomes are organised to optimize this expression, and how bacteria reproduce to spread in the environment. We can also easily observe that those genomes evolve over living conditions and time.

For example, even if most bacteria studied, like *E. coli* or *Listeria monocytogenes* are mesophiles (growing with moderate temperatures), other bacteria have adapted to extreme environments.

Regarding evolution over time, the arrival and spread of antibiotic resistance is a good witness of it. At the beginning of the antibiotic era, bacteria were all sensitive to antibiotics, and they progressively adapted to escape their attacks.

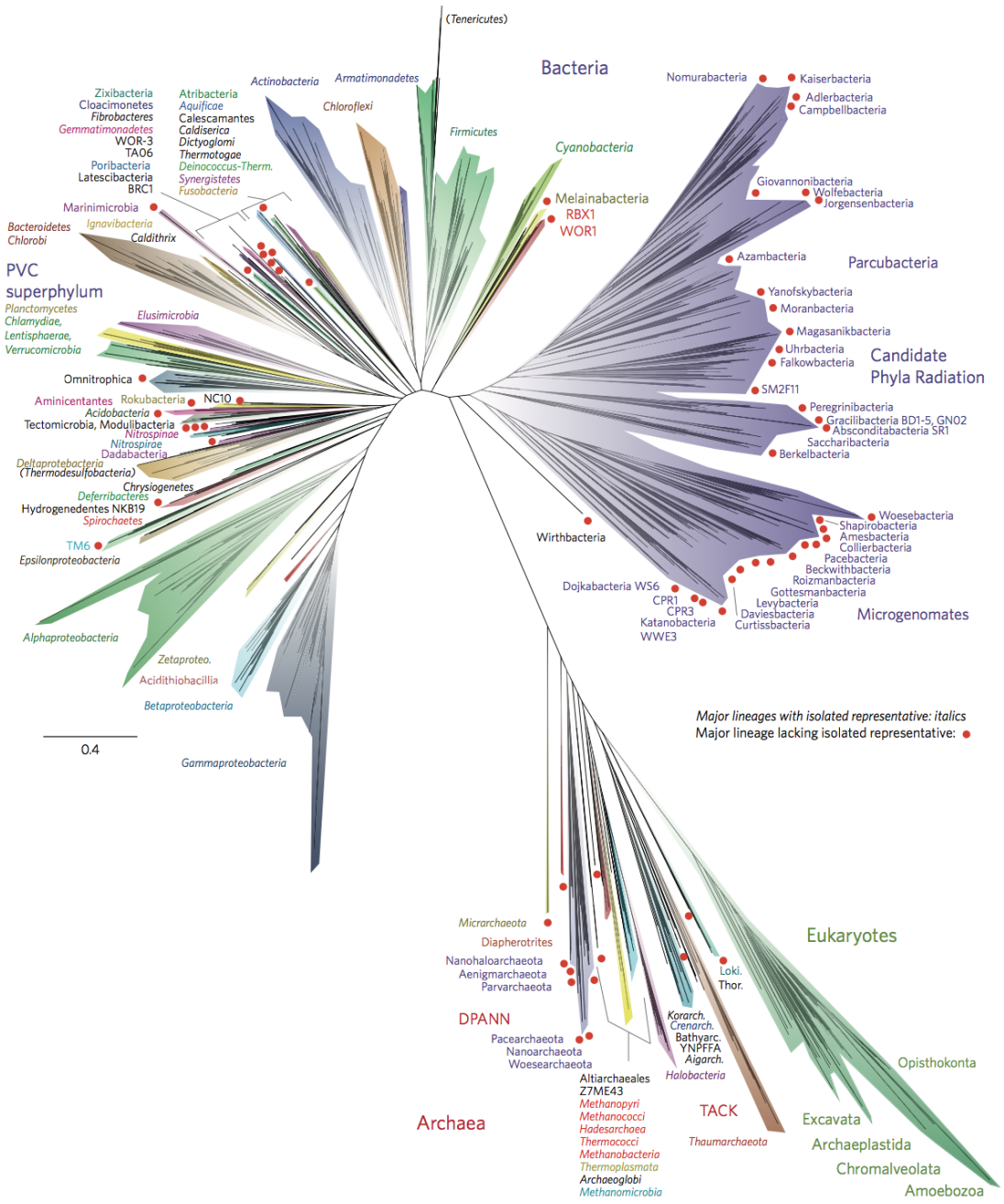


Figure 1.14: The tree includes 92 named bacterial phyla, 26 archaeal phyla and all five of the Eukaryotic supergroups. Major lineages are assigned arbitrary colours and named, with well-characterized lineage names, in italics. Figure from [92]

However, these observations lead to a paradox. Can you see why? Well, a high diversity of organisms still evolving over time is quite easy to understand with organisms growing by sexual reproduction (most eukaryotes): two cells exchange genetic material, giving a third

organism different from its two parents. But bacteria do not exchange genetic material during replication. Even worse, their binary fission mode of reproduction implies that they are all clones. So, how all these differences are possible? Where do those new genes come from? What are the mechanisms allowing such evolution? Let's meet in the next chapter to see that!

2 Genome evolution

A characteristic of bacteria is the high plasticity of their genome, despite their binary fission mode of reproduction supposed to create clones. They have a high ability to adapt when facing environmental changes (like new soil, air composition or temperature changes) or stressful events (presence of antibiotics, heavy metals).

The first mechanism coming to mind while speaking about genome changes is mutation. And yes, bacteria are affected by mutations. These can range from local changes, affecting a specific nucleotide, to larger scale mutations involving longer DNA sequences called mobile genetic elements (hereafter called MGEs), including deletions, inversions and duplications. However, this is not the main mechanism driving bacterial genome evolution.

A particularity of bacteria is that some of their mobile genetic elements can not only move within the bacterial genome, but also between organisms, by horizontal gene transfer (hereafter called HGT). This unidirectional process allows bacteria to acquire new genetic information at high rates. In addition to MGEs coming from other cells, the newly acquired DNA sequences can also be taken up from the environment, increasing the diversity of new genes acquirable.

It is now known that HGT (bringing new DNA), and not mutations, is the main process responsible for bacterial genome versatility [174]. A trendy example of the consequences of HGT is the rapid spread of antibiotic resistance genes, threatening human health.

In this chapter, I first present mobile genetic elements, which can be responsible for large changes in gene repertoires. Then, I describe by which mechanisms they can move within and between organisms, with an accent on the main mechanisms allowing HGT.

2.1 Mobile Genetic Elements

The generic definition of a Mobile Genetic Element (MGE) is a DNA sequence able to move within and/or between genomes. Some MGEs are autonomous, meaning that they encode all genes necessary for their own transfer from a genome to another one. Among them, we can cite conjugative elements and bacteriophages, who are driving bacterial evolution by generating HGT events (see 2.2). Their presence in bacterial genomes is highly variable.

For example, small plasmids (a few kbp) are sometimes over 100 copies per cell, whereas

larger ones are only a few copies per cell [183]. Regarding phages, a study on 2110 bacterial genomes found that 46% of them host at least one phage, with up to 15 prophages per genome [188].

Other mobile elements need systems encoded in trans to move within/between replicons, and/or to take advantage of self-transmissible MGEs (or other HGT mechanisms) to be horizontally transferred.

2.1.1 Conjugative elements

There are two main types of conjugative elements: Conjugative Plasmids (CPs) and Integrative and Conjugative Elements (ICEs). Although ICEs are the most abundant systems found in bacteria, CPs have been much more studied [81]. This is due to historical reasons, but also to the fact that plasmids are widely used in microbiology experiments, mainly for DNA cloning.

Conjugative Plasmids (CP)

As a reminder, a plasmid is an extra-chromosomal DNA molecule capable of self-replication (see chapter 1.3.3). They do not encode genes involved in essential processes, but often encode genes that can be beneficial for their host under certain circumstances.

Beyond their capacity to autonomously replicate, some plasmids also encode genes necessary for their own transfer from a donor cell (their host) to a recipient cell. Those plasmids are called self-transmissible or Conjugative Plasmids (CPs), as they are involved in the eponymous mechanism of HGT: conjugation (see 2.2.1). While transferring to a recipient cell, conjugative plasmids bring with them all the beneficial genes, providing new functions to their new host. For example, conjugative plasmid pOLA52 found in *E. coli* confers the capacity to form type III fimbriae upon bacteria carrying it, enhancing biofilm formation [133]. Like any plasmid, CPs are also replicated and partitioned between the daughter cells during cell division. Thereby, conjugative plasmids are transmitted both vertically (parent to offspring) and horizontally (via conjugation), making them important actors of genome evolution.

Microbiologists use CPs as vectors to introduce new genes (like antibiotic resistance genes, capsule genes...) in another bacterium. With the expression of those new genes in the bacterium, they can study their impact on the bacterium phenotype [38] [37].

Integrative and Conjugative Elements (ICE)

Many elements were described in the literature as being able to be horizontally transferred via a secretion system while being, unlike CPs, integrated into the bacterial chromosome. Those conjugative transposons, integrative 'plasmids', genomic islands and other unnamed elements were finally unified by the term Integrative and Conjugative Elements (ICE), coined by Burrus et al in 2002 [23].

The *Integrative* part of "ICE" describes their latent phase: integrated in the bacterial genome. In this state, they are passively replicated with the chromosome and vertically

transferred to the daughter cells. However, ICEs can also be induced, explaining the *Conjugative* part of their name. Upon induction, ICEs "excise by site-specific recombination, transfer the resulting circular form by conjugation and integrate by recombination between a specific site of this circular form and a site in the genome of their new host" [23] (see 2.2.1 for conjugation mechanism and 2.3.2 for site-specific recombination). In this way, they are also horizontally transmitted, increasing their spread in the bacterial population. They encode their own system of transfer, similar to the one found in CPs, but with additional genes for their specific excision and integration. In addition to those genes linked to their life style, ICEs code for other cargo genes, usually conferring new phenotypes to host cells.

A conjugation system is composed of at least four essential elements, responsible for the main steps of a conjugation event:

- an origin of transfer (*oriT*)
- a relaxase (also called *mob* gene)
- the Type 4 Coupling Protein (T4CP)
- the Type 4 secretion system (T4SS). This multi-molecular system is responsible for transferring macro-molecules through the bacterial membrane. Above its use in conjugation systems, it also has two other functions described [29]. It can be used for the transport of DNA and/or proteins to eukaryotic infected cells and it is also involved in exchanges of DNA (imports or exports) with the extracellular space (see transformation 2.2.2).

Also widely found in bacterial genomes, Integrative and Mobilizable Elements (IMEs) encode, as ICEs, their own excision and integration system. However, unlike ICEs, they do not have their own conjugation system, but use the conjugation machinery of their host for their own transfer [80].

However, the distinction between CPs and ICEs could be less obvious than commonly thought. Indeed, it has been suggested that ICEs could switch into CPs and stabilize in this state, even if the evolutionary pressures driving this phenomenon remain unknown [28]. At the other end, plasmids (like, for example, the F-plasmid or sex factor) can sometimes be integrated in the chromosome, in strains then called high-frequency recombination (Hfr) bacteria [77]. Those strains can transfer the entire chromosome by conjugation.

2.1.2 Phages

As already mentioned at the very beginning, human infectious diseases are not always due to bacteria: they can also be caused by viruses. I am mentioning this here because we are not the only ones suffering from viruses: bacteria also have theirs! In 1915, Twort discovered that small organisms were killing the *Micrococcus* bacteria he was studying [193]. Due to World War I, he was not able to investigate the nature of those organisms. Independently, D'Herelle discovered them in 1917 at Institut Pasteur Paris, while studying *Shigella dysenteriae*. Considering them as "bacteria eaters", he called them *bacteriophages* (*phágos* meaning *eater* in Greek), or *phages* for short [45] [46]. As for human (and any other living organism), they are submicroscopic parasites, only capable of reproducing if they enter host cells and

harness the transcription, translation and replication machinery of this host. Even if they are not as known as bacteria and viruses, probably because they do not directly infect human, bacteriophages are the most abundant life forms on Earth. There are $\simeq 10^{31}$ individuals, and they are responsible for $\simeq 10^{25}$ infections per second [89].

Phages encapsulate their genome inside a capsid. This genome can be single or double-stranded DNA or RNA, linear or circular. It ranges between a few kb to hundreds of kb in length. The majority of laboratory-studied phages have genomes of a few tens of kb. One of the shortest known phage genome, phage MS2, contains a single-stranded RNA genome of less than 5 kb, and encodes for only three proteins [60]. On the other hand, *jumbo phages* refer to phages with genomes larger than 200 kb, and *megaphages* those larger than 500kb. The largest genome sequenced so far is more than 700 kb [165].

Several ways have been suggested to classify phages. The International Committee on Taxonomy of Viruses (ICTV) proposes a classification into orders and hundreds of families according to their morphology and nature of genetic material [116]. More than 95% of the phages currently described belong to *Caudovirales* order (or tailed phages, *cauda* meaning "tail" in Latin). Those dsDNA phages have a typical morphology. Each virion has an icosahedral head capsule attached to a flexible tail through which the phage introduces its DNA into the infected bacteria. They are sub-divided into families, according to their tail particularities (contractility, size...). Other phages, like the above mentioned MS2 phage (*Norzivirales* order), also encapsulate their ssRNA inside an icosahedral capsule, but do not have any tail. On their side, filamentous phages (*Tubulavirales* order, with *Inoviridae* the most known family), called after their long and thin shape, directly encapsulate their circular ssDNA genome into their filamentous coat. Figure 2.1 shows the main morphologies of bacteriophages.

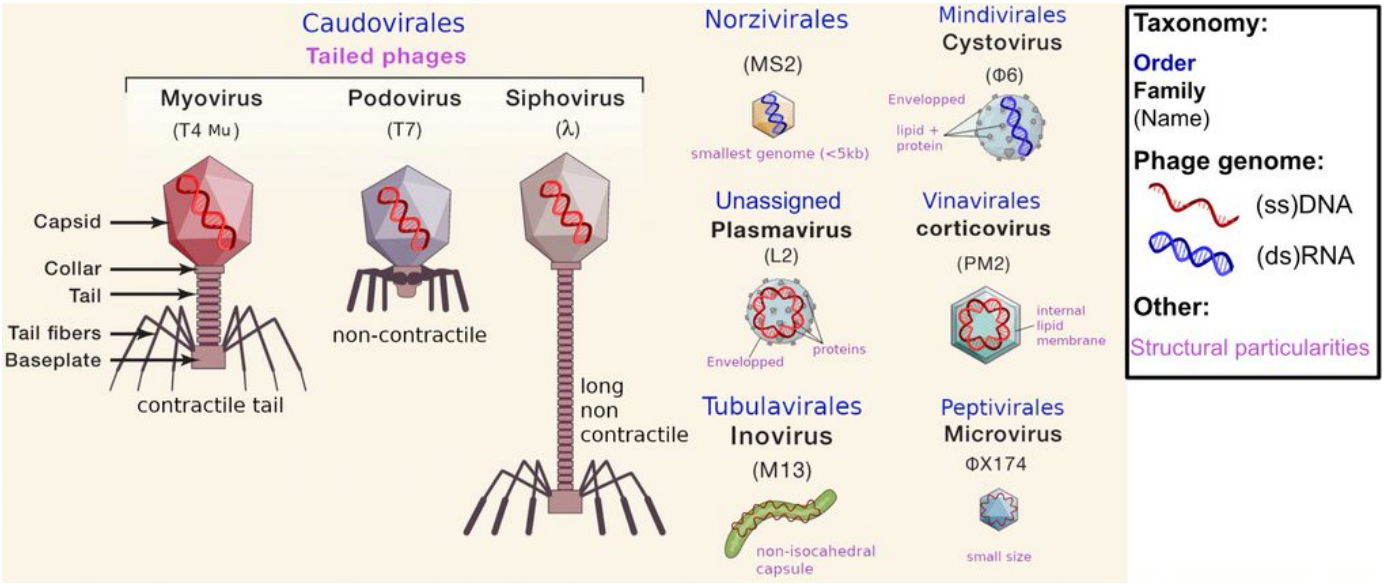
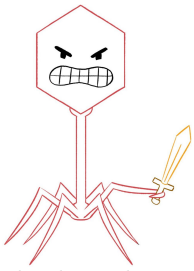


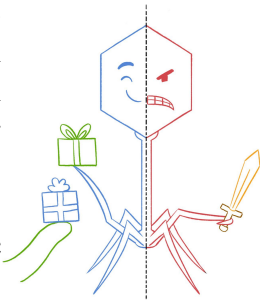
Figure 2.1: Different phage morphologies according to ICTV classification [116]. Figure adapted from [203].

A few essential genes are required for a phage to be functional. Among them, we can find gene(s) to bind and enter the host cell, genes to hijack the host replication machinery (or sometimes their own replication machinery), and gene(s) for DNA packaging [66].

Phages (especially dsDNA ones) can also be classified in function of their behavior during infection: temperate and virulent phages.



Virulent phages have the typical behaviour we have in mind when thinking of a virus: they infect a bacteria, and end by killing it. To do so, they enter the cell and harness the bacterial machinery to reproduce extensively. Once replicated, they package their DNA into particules called virions, and ultimately cause the cell lysis, spilling out its content (bacterial DNA and phage particules) in the environment. The new phage particules released will be able to infect other bacteria, and reproduce this *lytic* cycle. Coliphage T4 is one of the best known virulent phage [129].



Temperate phages, in contrast, lead a dual life. When entering a bacterium, they have two possible behaviors. On the one hand, they can act in the same way as virulent phages, resulting in the cell lysis, the bacterium death, and many new viral particles infecting other cells. On the other hand, they can reversibly integrate into the host chromosome as prophages or remain in a plasmid-like form [112] [17]. In both cases, they enter their "latent" phase, also called *lysogenic* cycle. Some phages need a specific insertion site, and integrate by site-specific recombination. Others can integrate almost any place in the chromosome by transposition (see 2.3.2). During lysogeny, the prophages provide new traits such as protection from other phages to their host. For example, lysogens (bacteria containing at least one prophage) are often found in unstable environments (low temperature, few nutrients, environmental stress...). In those poor growth conditions, it can be beneficial for the phage to remain integrated and keep the few hosts alive, in anticipation of better conditions to induce the lytic cycle. Thus, lysogeny can be beneficial for both the phage and its host. Replicated with the bacterial chromosome, prophages are vertically transmitted to the offspring. Later, the prophage can be induced by an environmental stimulus (ex: DNA damaging agents like UV or high temperatures, antibiotic treatment...). Upon this signal, it will excise from the chromosome, and switch to the lytic cycle. Therefore, temperate phages fluctuate between antagonistic and mutualistic (or at least commensal) relationship with their host. They code for genes to protect themselves (and by extension the host bacteria) from other viruses, but also for genes to bypass the host defense system to enter the lytic cycle. According to the state of the phage (lysogeny or lysis), different genes are expressed [25]. Temperate phages are quite common in bacterial genomes [188] [160]. Two model temperate prophages are λ [112], a prophage which integrates inside *E. coli* genome, and P1 [17], which lysogenizes *E. coli* as a plasmid.

Being temperate or lytic, the very first step of phage infection is the same. Indeed, before being able to inject its genome, the phage has to recognize and bind to the bacterium membrane. The bacterial population on which a given phage is able to bind determines its host range [94]. Some phages, like the previously mentioned λ phage, are specific to a few hosts [32]. They only infect a bacterial species (or even only some strains) having a specific

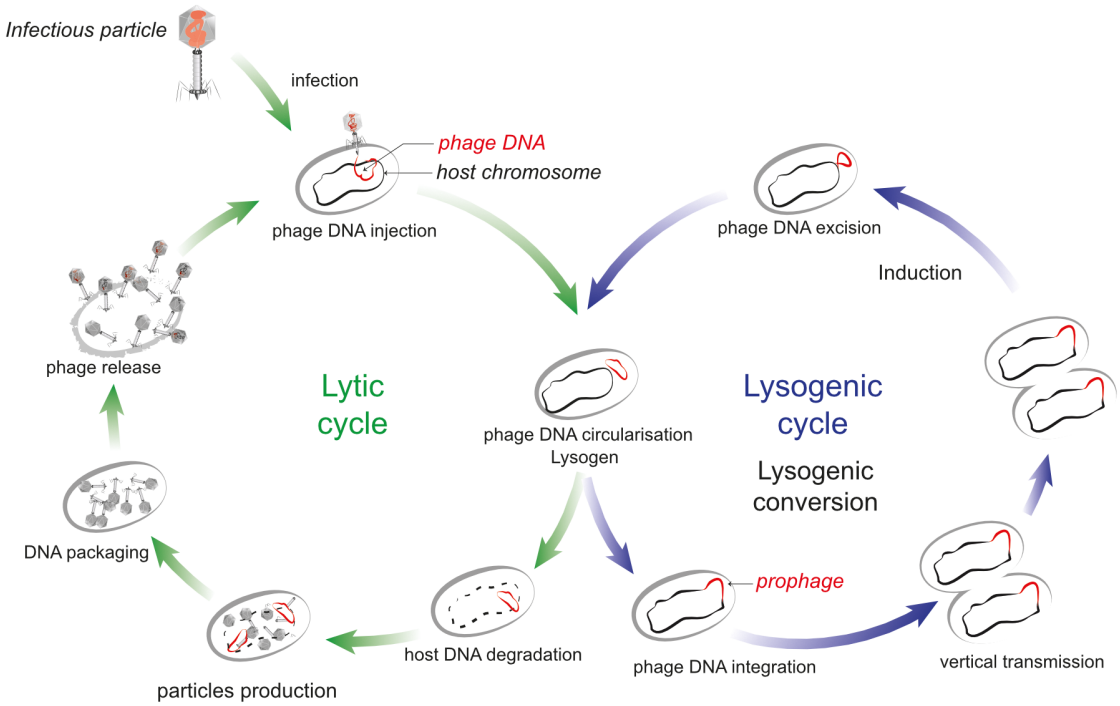


Figure 2.2: Phage behavior during infection. Figure from [189]

receptor on which they can bind. For temperate ones, they can even have a unique insertion site in the bacterial chromosome. Other phages have a wider host range. When they are temperate, they can insert almost anywhere in the host genome. This is for example the case with phage Mu, which can infect, among others, bacteria of genera *Escherichia* or *Salmonella* [86].

Finally, phages (mostly virulent ones) are also studied in the aim of fighting against pathogenic bacteria. This application, called *phage therapy*, was already proposed by D'Herelle himself, who cured a 12-year-old boy from severe dysentery in 1919 [177]. However, with the development of antibiotics, phage therapy trials ceased, at least in Western Europe. Nowadays, with the alarming emergence and spread of (multi) antibiotic resistant bacteria, this field is progressively gaining a new momentum [177]. Phages have the advantage, among others, to be more specific than antibiotics, which could avoid the destruction of the whole microflora of the patient during treatment. However, even if many successful experiments have been reported, many challenges remain to be solved in this "new" field [25]. For instance, as bacterial genomes constantly evolve, emergence of phage-resistant strains is unavoidable.

And to end, guess what? The Russian doll of viruses does not end with phages: they also have their own parasites! Those so called phage satellites do not even have genes to self-mobilize. They hijack the phage (then called helper-phage) to make virions and allow their spread. The most famous phage satellite is P4, using P2 as a helper phage, which is itself using *E. coli* to replicate [169].

2.1.3 Jumping DNA

So far, I have presented MGEs able to move between cells, as they are the ones driving genome evolution by bringing new genes to bacteria. However, bacterial genomes also host other types of MGEs, which can move from place to place in the genome, but cannot transfer themselves to other cells [168]. Even if they do not bring new genes to the bacteria, they still have an important impact on the genome evolution. For example, they can increase (or decrease) the amount of DNA, and/or generate mutations potentially leading to gene inactivation (see 2.3.2).

Those elements, called with the generic term of Transposable Elements (TEs), are DNA sequences which can repeatedly move within a chromosome (or between two replicons when there are several in the cell), via a mechanism called transposition (see 2.3.2) [9]. The discovery of those "jumping genes" in 1983 by Barbara McClintock earned her a Nobel Prize [124].

There are several categories of TEs. Insertion Sequences (IS) are the simplest MGE. They only code for a transposase, an enzyme required for their transposition, and are flanked by short inverted repeats (IR), used by the transposase to initiate transposition [168] [121]. A bacterial genome can contain hundreds of copies of a same IS.

Two IS flanking several accessory (not essential for transposition) genes can constitute a DNA compound transposon or composite transposon (Tn). Genes in the intervening DNA segment are called passenger genes, or cargo. The composite transposon is transposed as a unit, thanks to the transposases of the two flanking IS acting in concert (see 2.3.2). Tn*10* (flanked by IS*10*) and Tn*5* (flanked by IS*50*) are among the most famous composite transposons, as they carry antibiotic resistance genes [9]. Another common type of Tn are non-composite or unit transposons, like Tn*3* family. Contrary to composite ones, they are not flanked by IS, but encode their own transposase among their passenger genes.

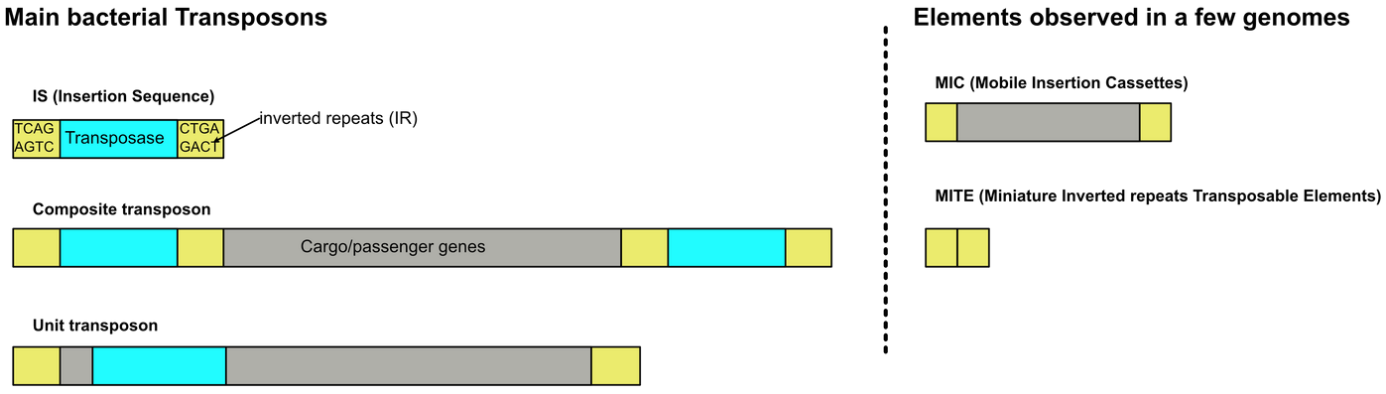


Figure 2.3: Different types of transposons

Whatever the type of transposon, passenger genes are often beneficial for the bacteria. For example, studies on high-level vancomycin-resistant *Staphylococcus aureus* strains revealed that this resistance was provided by genes coded in transposon *Tn1546* [200]. As for

conjugative elements, the distinction between the different TE types is becoming less clear with the discovery of new elements. In this way, some unit transposons lack passenger genes, some are instead devoided of transposase (MICs for Mobile Insertion Cassettes), and others lack both (MITEs for Miniature Inverted repeats Transposable Elements) [167].

Even if TEs do not carry genes necessary to move between genomes, they are still actively participating to genome evolution via intercellular exchanges. To do so, they use self-transmissible elements such as plasmids and phages, by "jumping" and integrating inside their DNA sequence (see 2.3.2).

Integrons, known for their important role in antibiotic resistance dissemination, also participate to intracellular diversity. Those adaptive elements are not mobile by themselves, but they are managing an array of gene cassettes, which are the smallest mobile genetic elements known so far. Those cassettes, able to move within and between integrons, usually carry a single gene, most of the time conferring antibiotic resistance to the host [53]. Thanks to a specific integrase *intI*, integrons are able to capture, remove and reorganise the cluster of cassettes (which are flanked by *attC* recombination sites) [53] (see figure 2.4). Besides, integrons (in particular class 1 integrons) can even be captured and integrated into transposons and/or plasmids, providing them intra and inter-cellular mobility [119].

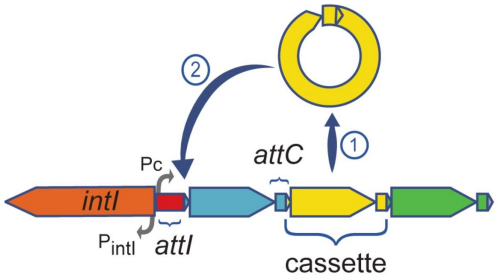


Figure 2.4: Schema of a typical integron. (1) The integrase can excise a cassette (2) and/or integrate it at the *attI* recombination site. Figure from [44]

Our little meeting with the main actors of genomic diversity comes to an end. We saw that genes carried by MGEs provide new traits to their hosts to adapt to their environment: resistance to antibiotics, use of new nutrients as food, resistance to heavy metals, etc. Those MGEs have a very important consequence on bacterial phenotype, sometimes at the cost of humans. For example, the only difference between a harmless commensal bacterium and a deadly pathogen can be the presence of a bacteriophage (diphteria [65]) or a plasmid (antrax [134]). But how do these little pieces of DNA end up in those strains? How do they manage to transfer between bacteria? I guess that you are suspecting the already mentioned "Horizontal Gene Transfer". And yes, HGT is (part of) the answer. Thus, I think that it is now the right time to explore a little bit what is hidden behind those three words.

2.2 Horizontal Gene Transfer

Horizontal Gene Transfer (HGT) is the mechanism by which bacteria can acquire new DNA (being integrated into the chromosome or as a free replicon) from their environment and/or neighbouring bacteria [174].

HGT can happen within but also between two different bacterial species. Although it was first considered as a side mechanism of genome evolution, it is now known as the most important mechanism driving bacterial diversity. There are three major HGT process: conjugation, transduction and transformation.

And last but not least: several mechanisms need those much-vaunted MGEs!

2.2.1 Conjugation

Conjugation, often referred as "bacterial sex", is a mechanism by which two cells directly exchange DNA. It was discovered in the late 40s, while Lederberg et al. studed recombination (see 2.3.2) on a mixed culture of *E. coli* strains [113]. Observing the recombined mutants, they saw unexpected ones. The genomes of the latter were the result of recombination of genes, but those genes were not from the same initial strain. They concluded that this implied a "sexual process" in the bacterium: two different cells were exchanging their DNA. At this point, they did not know how this 'cell fusion' could happen. Further studies clarified this process, which was called conjugation [47].

Figure 2.5: E. coli strains undergoing conjugation via a pilus ©Dennis Kunkel Microscopy

This mechanism is not species specific. Even if it is less frequent, two bacteria from two different species can exchange DNA via conjugation, as long as they are in close contact [79] [48]. It has even been shown to participate to trans-kingdom horizontal gene transfer. First experiments in the late 80s reported that *Escherichia coli* can transfer DNA to *Saccharomyces cerevisiae* yeast [88]. Many years after, in the early 2020s, a trans-kingdom conjugation was observed in nature, where it was shown that virulent strains of *Agrobacterium tumefaciens* cause tumours in plants [147]. Briefly, thanks to the expression of genes coded on the bacterium Ti plasmid, a DNA region is transferred via the T4SS (see below) to the plant, and integrated into its chromosome. This DNA region contains genes which, once translated by the host, generate enzymes who induce tumor growth. Hence, more than a simple transfer of DNA, trans-kingdom conjugation can even change the phenotype of the eukaryotic host.

Compared to the two other HGT mechanisms, conjugation is the one which can transfer the largest amount of DNA per event (up to almost an entire chromosome, e.g. for *E. coli* Hfr strains), and over the broadest range of organisms [192].

We already met the two MGE responsible for conjugation in 2.1.1: conjugative plasmids (extra-chromosomal system) and ICEs (integrated inside the bacterial chromosome).

1. The first step for conjugation of an ICE is to excise from the chromosome and circularise, to become a "plasmid-like" independent DNA molecule. Apart from this pre-requiste (and its re-integration to the chromosome after), the conjugation process is the same for both ICEs and plasmids.

2. At first, in the host cell, the relaxase introduces a specific nick (cutting one strand of the DNA) next to the *oriT*, unwinds the nicked strand from the unbroken one, and binds to it [24].

3. The T4CP protein brings the complex *oriT*/relaxase to the T4SS. If another bacteria is nearby, the latter uses its appendage (called pilus or sex pilus), to create an 'inter-cellular cytoplasmic bridge' between the donor cell and its recipient, forming a mating pair. This process is consequently called the mating pair formation (Mpf) [162]. The single-stranded DNA complex is secreted through the membrane pore created by the T4SS to enter the recipient cell, in an unidirectional way [48].

4. Once in this new host cell, the relaxase detaches from *oriT* and helps to circularise the ssDNA.

5. Both circular single-stranded DNAs (remaining strand in donor and transferred strand in recipient cell) are replicated to get back to their dsDNA state.

Finally, if the MGE involved is an ICE, it is (re-)integrated in (the donor and) the recipient chromosomes.

The process is depicted in figure 2.6.

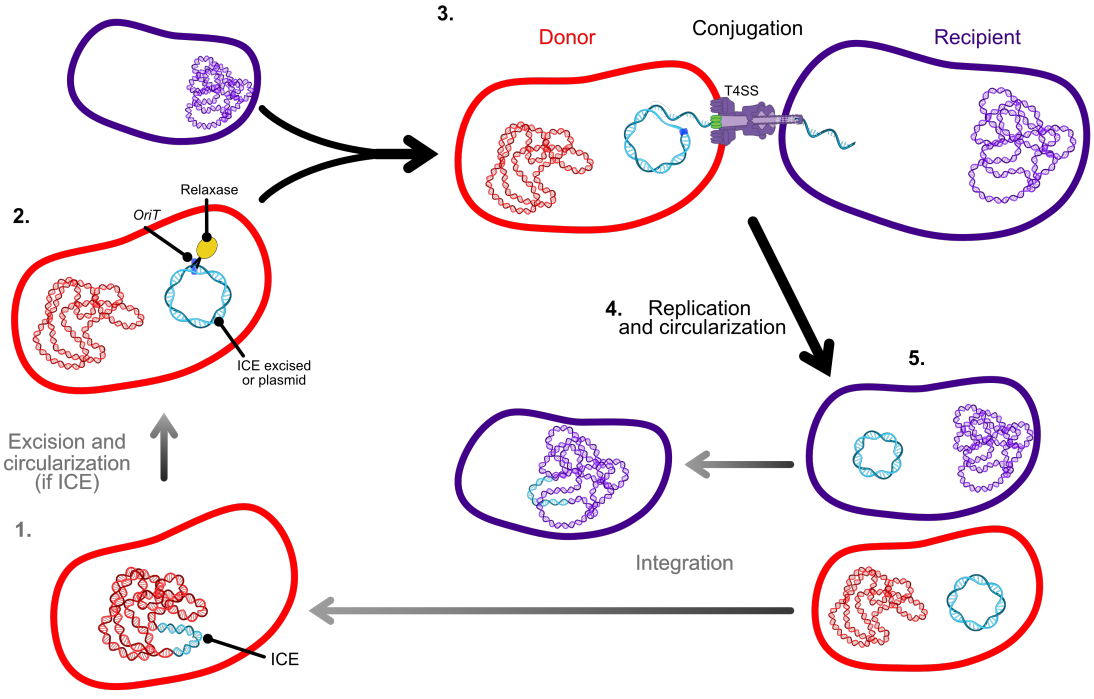


Figure 2.6: Conjugation process. See detailed steps above. Grey arrows represent optional steps (only for ICEs).

At the end of conjugation, both the donor and recipient bacteria have the plasmid or ICE, and are able to spread it to their progeny by vertical transmission (or of course again by HGT). In that way, conjugative systems are major players in the spread of antibiotic resistance, metabolic pathways, symbiotic traits, and even other mobile genetic elements like TEs (carrying themselves potentially pathogenic genes) [15] [117].

2.2.2 Transformation

Unlike conjugation, transformation does not require cell-to-cell interactions, as it consists of importing DNA fragments released in the environment.

This was the first HGT mechanism discovered, in 1928, when Frederick Griffith was studying virulent (S strains) and harmless (R strains) *Streptococcus pneumoniae* bacteria by observing their effect on mice [76]. While inoculating some mice with both alive R strains, and heat-killed S strains, he observed that the mice died of pneumonia, while they survived to the injection of each of these two types of strains separately. Even more stunning, he discovered that the live bacteria he took up and plated from the dead mice blood were of the virulent type. He concluded that such apparently haphazard results should be due to some "transforming principle" being taken up from the heat-dead cells to convert R strains into virulent ones.

It is only after Griffith's death, in 1944, that Avery, McCarty, and MacLeod identified the "transforming principle" as being the genetic material (i.e. the DNA) of the dead S strains. The process of integrating free DNA into a bacterial genome is now known as transformation, in reference to Griffith's "transforming principle" [8].

Natural transformation does not require any MGE: it is a property inherent to the bacterium. Even if a bacterium is transformable, it does not necessarily mean that it can uptake DNA from its environment at any time and anywhere [173]. Indeed, this capacity to capture DNA, named competence, depends on the expression of a specific set of genes, called *com* regulons [98]. Most naturally transformable bacteria are not permanently expressing those genes, but rather need some specific conditions to become competent (i.e. cell-cell signalling, stressful conditions, nutritional depletion, high cell density...) [98] [173].

The process of natural transformation consists in three main steps.

1. First, for transformation to be able to occur, some naked DNA is required. The latter is most of the time in the environment following the lysis of a dead cell. However, other mechanisms have been described. Some bacteria, like *Neisseria gonorrhoeae*, "donate" their DNA on purpose, via autolysis, or via type IV secretion (through the T4SS already described in 2.2.1) [85]. Other bacteria commit fratricide: they kill their siblings (cells genetically identical) by allolysis. This has been described on the Gram-positive *Streptococcus pneumoniae* species, where some genes of the *com* regulon, expressed in the competent state, code for products killing the non-competent sister cells [36].

2. The second step is on the competent bacterium side, and consists in up-taking the environmental DNA. Elements forming the transformation system vary a lot according to bacteria, and little is known about most of them. However, it has been described that the process of transformation generally involves a transformation pilus, *Tfp*. The latter binds to

the exogenous dsDNA (with a specific sequence or not), and, thanks to other little known proteins, imports it in the cytosol and processes it as a single-stranded form [98].

3. The last step is the integration of the ssDNA in the chromosome, often by homologous recombination (see 2.3.2).

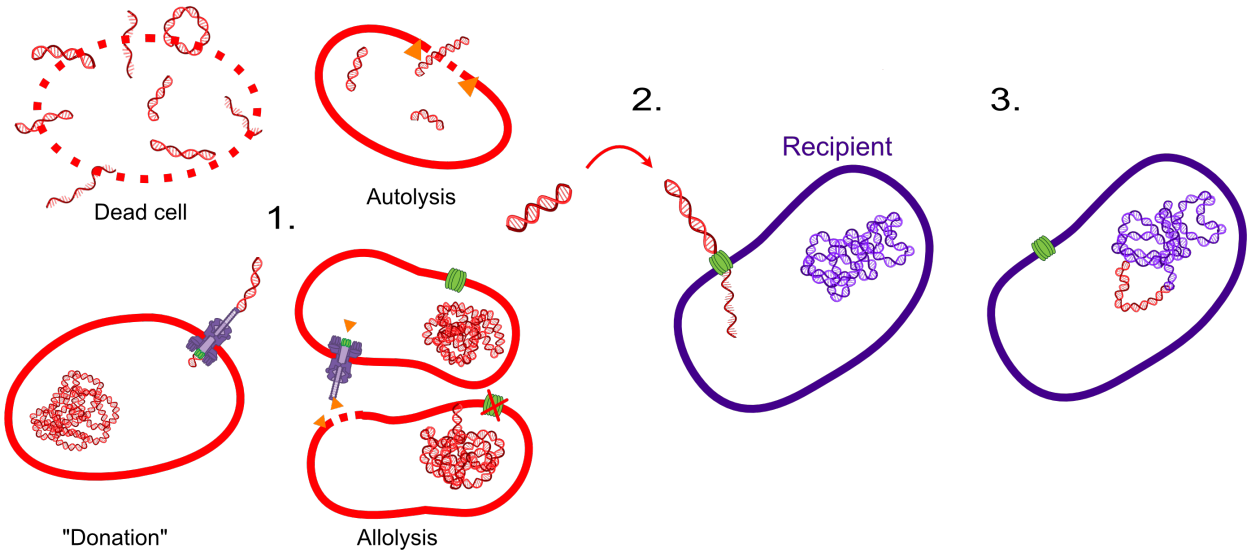


Figure 2.7: Transformation process. See detailed steps above.

Beyond facilitating genetic material exchanges, the role of transformation is still, nowadays, not completely known. It was thought to be used for nutrition [61] [152], for genome maintenance [130] or else for genome diversification [98]. As any other HGT mechanism, transformation enables bacteria to acquire new genes. The latter can help them to adapt to their environment (virulence genes, resistance to antibiotics or else evasion of vaccines) [76] [175] [195].

2.2.3 Transduction

The third HGT mechanism, transduction, is the transfer of DNA from a donor to a recipient bacterial cell via a viral vector, the latter being none other than the much-vaunted bacteriophage [189] (see 2.1.2 p.35).

Transduction was discovered and named in 1952 by Zinder and Lederberg. The latter, also being the discoverer of bacterial conjugation on *E. coli* six years before (see 2.2.1), decided to do the same experiments but with two mutants of *Salmonella typhimurium* strains [210]. Similar to with *E. coli* strains, recombination events between the two mutants were observed while plating both strains together. They wanted to go further by doing an experiment following Davis nonfiltrability U-tube design [47]. A U-tube consists in two pieces of curved glass tubes (each one receiving one of the two mutants) fused at their base, forming a 'U' shape. This time, the result was more disconcerting. Despite the filter (with pores smaller

than a bacterium) they inserted between the two tubes, they observed recombination events. This meant that what they called a 'filtrable agent' acted as a vector between the two cells. By varying the size of the pores, they identified this agent to be about the size of the P22 temperate phage of *Salmonella*, and further studies confirmed the viral nature of the agent [210].

As we already saw in 2.1.2, phages package their genetic material into their head or capsid. However, while doing so, they sometimes erroneously take some of the bacterial DNA with them. This accidentally packaged DNA is at the origin of transduction. Due to the different lifestyles of phages, we can distinguish two types of transduction.

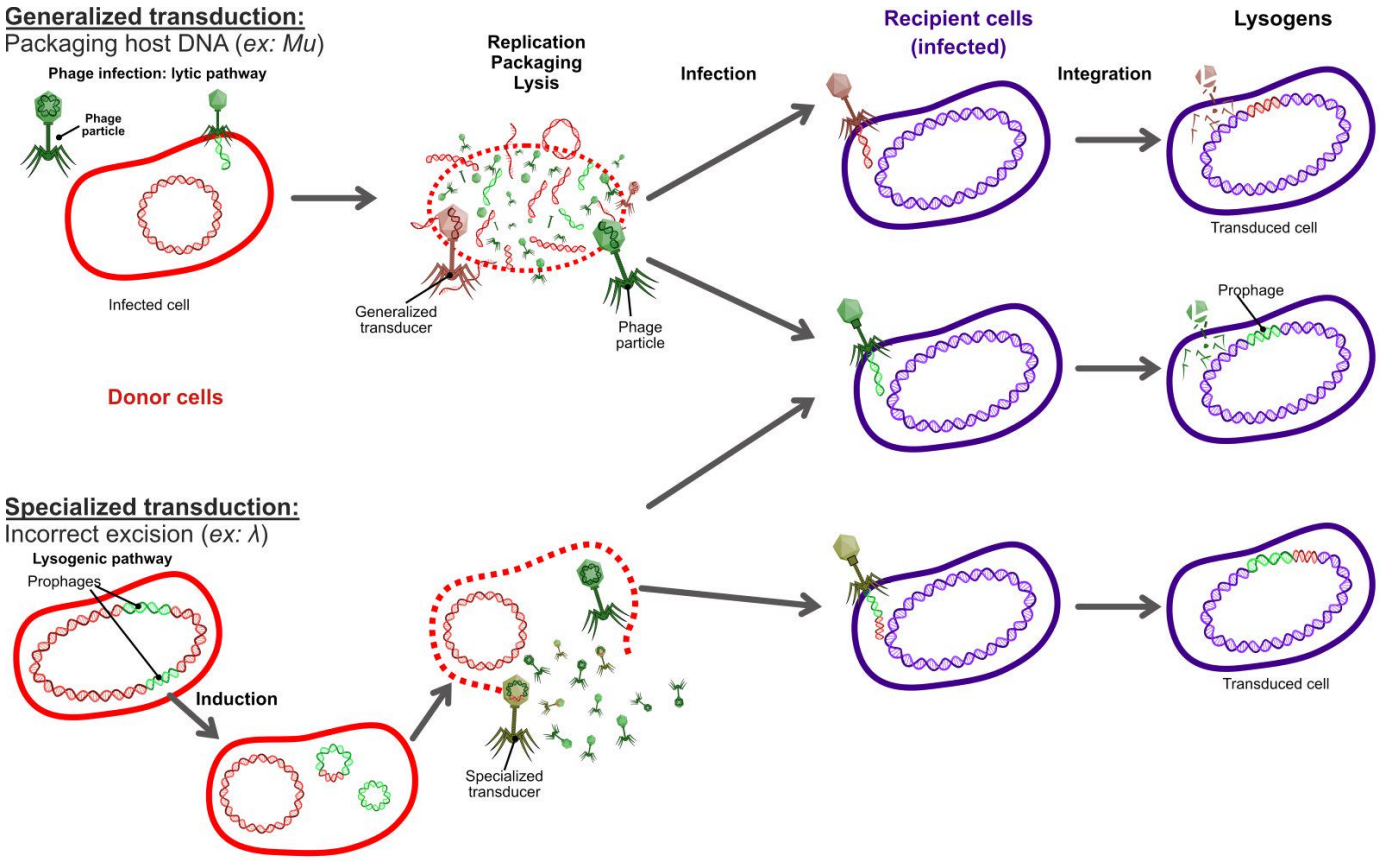


Figure 2.8: Transduction process

First, when a phage, being virulent or temperate, enters its lytic cycle, it harnesses the cell machinery to replicate and create new virions. This leads to the bacterial lysis, with the bacterial DNA being fragmented into many small pieces. While encapsulating their genetic material, some phages may take, instead, some of those random DNA fragments with them. This mechanism is called *generalized transduction*, as it can transduce any region of the bacterial genome [27] (see figure 2.8). The amount of DNA generally transduced is variable. It can be up to tens of genes co-packaged, depending on the capsid size. Generalized transduction frequency is also very variable according to species, due to the different types of

prophages they carry [30]. Some species have a majority of their phages capable of generalized transduction. For example, a study showed that 99% of the phages of *Salmonella* they used are able to perform generalized transduction, partly due to their permissive *pac* packaging system [160]. Other phages (like λ of *E. coli*) use a packaging system much more specific, almost always preventing generalized transduction [95] [143]. Besides, during an infection, only a small fraction of the phages able to generally transduce are packaging some bacterial DNA [102], which also reduces final generalized transduction frequency.

Another type of transduction exists, this one only being possible with temperate phages. When it is induced, the temperate prophage must excise from the chromosome to enter its lytic cycle. Sometimes, this excision is inaccurate, and the phage also takes with it a few genes flanking its attachment site *attB* [27] (see figure 2.8). Those bacterial genes will be co-packaged with the phage DNA in the virion, the volume of the capsid being large enough for a DNA molecule longer than the phage DNA itself [59]. This mechanism, called *specialized transduction* as it can only transduce specific parts of the bacterial DNA, has been described in a few phages, the most studied being λ phage of *E. coli* [59].

There is a third phage-mediated HGT mechanism. When (generally or specialized) transduced fragments are packaged in a temperate phage, this one can later lysogenize as a prophage in a new bacteria (see 2.1.2), and provide those new genes to the host (see figure 2.8). In that way, transduction contributes significantly to the genetic diversity of many bacteria. When the expression of those new genes engenders phenotypic changes to the bacteria, we talk about *lysogenic conversion*. This mechanism can potentially transform a harmless bacteria into a dangerous pathogen. For example, presence of corynephage β allows previously inoffensive *Corynebacterium diphteria* strains to produce a new toxin, causing diphteria [65]. Although the role of transduction in the spread of antibiotic resistance is less clear than the role of conjugation, a few studies found ARGs in phages or prophages [39].

As transformation, transduction does not need any cell-to-cell contact. This allows transfers of genetic material over longer distances, and longer time periods than the two other HGT mechanisms. Indeed, conjugation is limited both in space (cell-to-cell contact needed) and in time (the pore opened between the two cells does not stay for a long time). Regarding transformation, it is less constrained in space, as free DNA can move in the environment, and potentially come from a cell which died a bit further. However, DNA alone in the environment is quickly degraded by deoxyribonucleases (DNase), so transformation must happen in a quite restricted area and time period. Regarding transduction, DNA can stay longer in the environment and travel along longer distances, because it is protected by the phage capsid. On the other hand, transduction has the narrowest host range of the HGT mechanisms, due to the limited host range of most temperate phages [94].

2.3 Intragenomic evolution

Even if Horizontal Gene Transfer is the main mechanism driving bacterial evolution, other mechanisms can impact the evolution of genomes. Contrary to HGT where bacteria acquire

new DNA sequences, these other mechanisms reshuffle the existing genome of a bacterium: they introduce mutations. With the advent of the Covid19 pandemic, *mutations* is quite frequently used in everyday conversations. This makes sense, as mutations are the main motor of evolution in viruses. And it was, for a long time, also considered as the main mechanism of bacterial genome evolution before it became clear that gene repertoires are mostly modified by HGT.

By definition, a mutation is any heritable change in any part of the genome sequence (chromosome or plasmid), inducing or not a phenotypic change. (As an aside, this means that the broadcasted SARS-Cov2 mutations are only a few of all mutations really occurring. The ones not causing any phenotypic change do not deserve to be named.) Mutations can affect regions of different sizes and have diverse consequences. Those mutations are then spread in the population via vertical descent, which can also itself introduce a few changes.

2.3.1 Point mutations

Point mutations are the major source of Single Nucleotide Polymorphism (SNPs). Their different types are detailed below, and illustrated in figure 2.9.

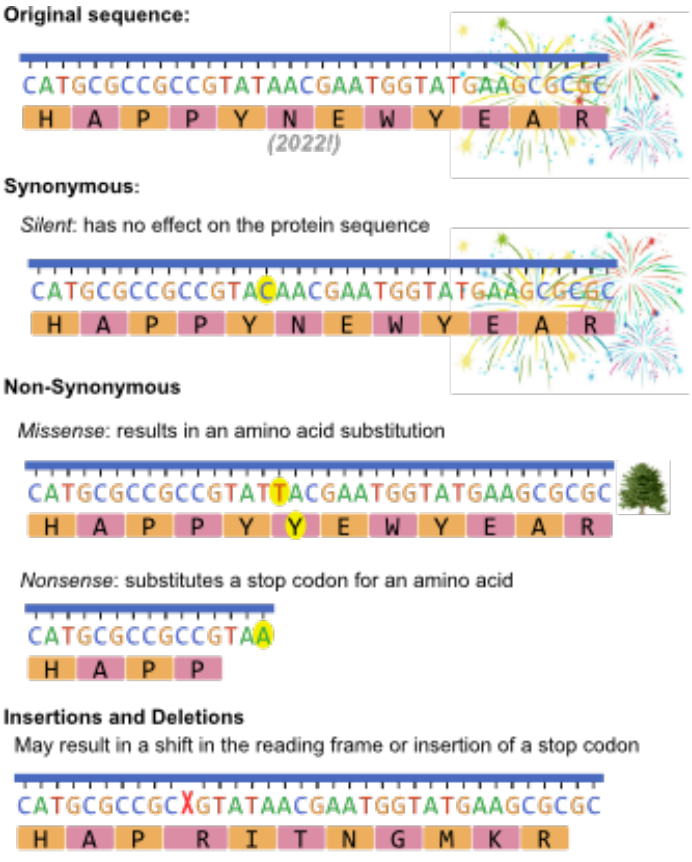


Figure 2.9: Different types of point mutations. See text on the right for more details.

Silent (or *synonymous*) mutations have no effect on the amino acid composition of the protein, as, thanks to the redundancy of the genetic code, they replace a codon with another one coding for the same amino-acid. On the other side, *non-synonymous* mutations change the protein sequence and can have a high impact on its function. They can replace a codon by another one coding for a different amino-acid (*missense* mutation) or by a premature stop codon (*non-sense* mutation). This type of mutation are very rarely adaptive. They sometimes have a neutral effect (if the new amino-acid is close enough from the original one), but are most of the time deleterious (result in the loss of the gene function).

Other point mutation types, like insertions or deletions of a nucleotide inside a gene, are most of the time deleterious: they lead to a frameshift, generally inactivating the gene (see figure 2.9).

At first, mutations were assumed to appear randomly. However, although spontaneous mutations exist, they are

most of the time the result of an event (during DNA replication, following an erroneous DNA repair, or induced by a DNA damaging agent like X-rays or UV) [114].

Most mutations, being deleterious, tend to be counter-selected and to disappear from the bacterial populations. Sometimes, they can be affected by reverse mutations (reversions) especially by DNA repair systems (see 1.3.3). In both scenarios, they are not kept and spread into the population. Yet, for some strains, called *mutators*, mutations fixed in the population are more frequent. Those strains have a particularly high rate of mutations, most of the time as a consequence of a defective DNA repair system [50] (see 1.3.3). Mutating frequently has the advantage to faster explore many combinations. Like so, there is a higher opportunity to find positive mutations that can be latter transmitted to progeny and thus fixed in the population. However, this does not change the fact that the vast majority of those mutations are deleterious, meaning that being a mutator still has a real cost.

2.3.2 Large scale mutations

Large scale mutations involve not only one nucleotide but whole DNA segments that are rearranged. We will from now on use the term *rearrangement* instead of large scale mutation, to avoid confusion with point mutations. Those rearrangements, including translocations, inversions or duplications of DNA, can lead to gene function loss. This can occur when the sequence of a gene is directly affected (insertion of DNA in the middle of the gene, moving only a part of the gene etc). When it does not affect the sequence of the gene by itself, the rearrangement can still have an impact on gene expression. For example, an insertion can split an operon, or change a promotor sequence. Therefore, even if, as for point mutations, some rearrangements can be silent, the majority have a high phenotypic impact. For example, large-scale rearrangements can help pathogenic bacteria bypass their host defences [139].

Rearrangements are mainly the result of recombination, a process consisting in joining two DNA segments which were previously separated. Horizontal gene transfer, however powerful to exchange genetic material between cells, would not be that effective without the help of recombination. Indeed, except for autonomous plasmid-like MGEs, horizontally transferred DNA must integrate the host chromosome to be able to persist in the next generations. This can be done by several recombination processes.

Homologous recombination

The most famous is probably homologous recombination (HR), which consists in the exchange of two DNA sequences (not necessarily on the same replicon) flanked by nearly identical regions. It is sometimes called general recombination. Originally described in the late 40s, homologous recombination was at the origin of the discovery of an horizontal gene transfer mechanism: conjugation (see 2.2.1) [113].

Homologous recombination naturally occurs not only in bacteria, but also in eukaryotes and even some viruses. Whether responsible for new DNA combinations during eukaryotic meiosis (cell-division used for sexual reproduction) or for integration of DNA after HGT in prokaryotes (see figure 2.10), homologous recombination has an universal critical role in

producing genetic diversity.

It was thus considered for many years as an equivalent of the eukaryotic sexual process, which role is to mix DNA in order to diversify genomes. In the mid 60s, while observing that recombination-deficient mutants were more sensible to DNA damage, scientists realised that the molecular mechanism behind DNA repair could have common steps with homologous recombination [34]. Subsequent studies finally revealed that, in fact, homologous recombination is primarily a major DNA repair system [107]. Therefore, even if they are often associated, it is important to understand that HGT and HR are distinct mechanisms. Unlike HGT which brings new DNA material, HR only exchanges two existing homologous sequences, not necessarily provided by an HGT event.

The molecular mechanism has been originally studied in *E. coli*, which remains the reference for HR [105]. This process is catalyzed by several sets of enzymes, the central one being the recombinase *RecA*. Its sequence is highly conserved among all organisms, even beyond bacteria domain [154]. Homologous recombination mechanism involves three main stages, each one involving multiple enzymes assembled in complexes [35]. First, *presynaptic* enzymes prepare the parent sequence by unwinding an extremity, converting it to a partially single-stranded sequence. In our case, this parent sequence can be, for example, some DNA introduced by a phage after generalized transduction (see Transduction). Recombinase *RecA*, who has access to the single-stranded parts, coats them, forming a nucleoprotein filament. In the second phase (*synapsis*), the coated presynaptic complex will scan the host replicons for homologous regions, and bind to it. This results in a joint molecule made of four dsDNA arms, called Holliday structure. Finally, *postsynaptic* resolvases cleave and join the appropriate strands to resolve the Holliday junction, and get a viable hybrid recombinant.

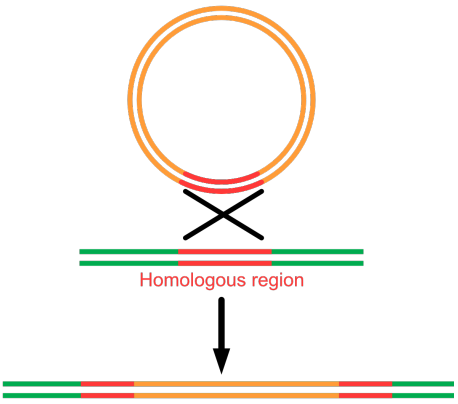


Figure 2.10: Insertion of a MGE by Homologous Recombination

Specialised recombination mechanisms

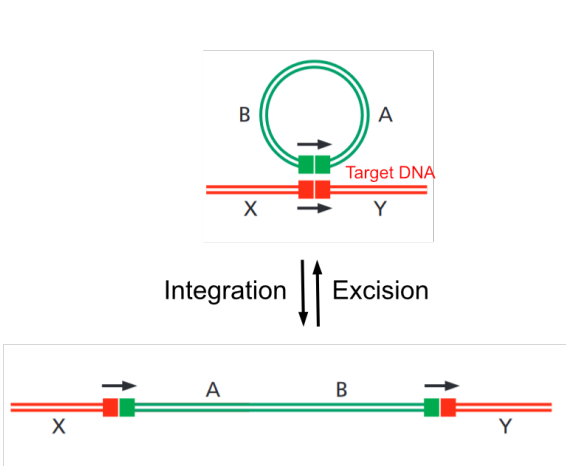
In addition to homologous recombination, other mechanisms are commonly reshuffling bacterial DNA. Contrary to HR, those specialized recombination mechanisms do not rely on extensive homologous sequences, and use relatively simple machineries.

Site-specific recombination needs precise DNA sequences of tens of bp. The latter, called recombination sites, define the specific positions at which the site-specific recombination occurs.

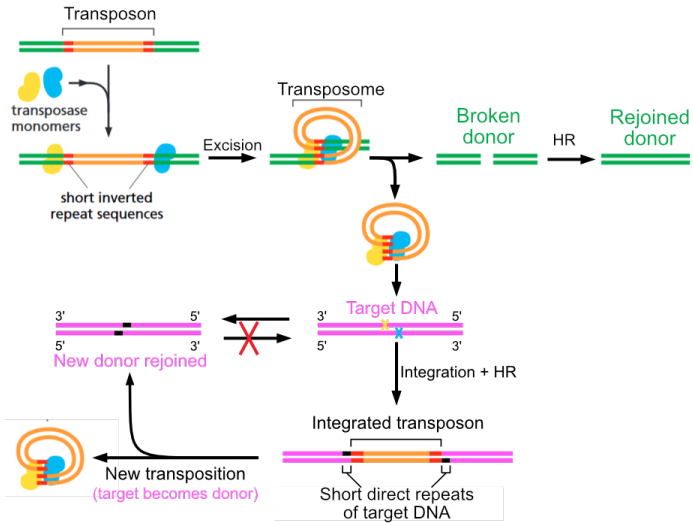
The process is much simpler than HR mechanism (see figure 2.11a). Most of the time, it only requires one enzyme, called site-specific recombinase (SSR), and two specific recombination sites. Two different types of SSRs have been described in literature. Although they have different mechanisms to execute each step, the processes are analogous. The SSR first

binds to the two specific recombination sites on the sequence, forming a synaptic complex. It then catalyzes the cleavage of the DNA sequence at both bound sites, exchanges the DNA segment in between, and rejoins the DNA strands [78].

Site-specific recombination is a conservative mechanism: it does not involve any DNA gain or loss. The original configuration can be re-established by a reciprocal recombination. This mechanism is used by most ICEs and most prophages to reversibly integrate or excise the bacterial chromosome at specific sites [26].



(a) Site-specific recombination



(b) Transposition

Figure 2.11: Specialised recombination mechanisms. Figures adapted from [2]

Transposition is another recombination mechanism, by which transposable elements (see 2.1.3) move within or between replicons (see figure 2.11b). Transposases, enzymes catalysing transposition, are generally self-encoded by the TE.

Conservative (or non-replicative) transposition is a "cut-and-paste" mechanism by which the TE excises from the chromosome, and inserts into a new non-homologous genome locus called *target site*. This process is catalysed by the transposase, which binds on the inverted repeats flanking the TE, and cleaves the dsDNA strand to excise it. Cleaved ends at the donor site will be joined by cell repair systems like HR. On the other hand, replicative transcription requires the replication of the TE before its transposition. The copy is inserted at the target site, while the original element remains in place, leading to an increase of the genetic material (duplication of the transposon DNA) [84].

In both types of transposition, the transposase cuts the two DNA strands of the target site a few base pairs from each other. Once it has inserted the TE between the two overhangs, the host repair system fills the gaps on the two complementary staggered strands.This introduces direct repeats of a few nucleotides on each side of the newly inserted element. Presence of those direct repeats in a genome are a signature of transposition events, as they stay in place if the TE is transposed to another locus. As a consequence, contrary to site-specific

recombination, transposition is not conservative [168] .

Selectivity and nature of the target site are very different according to the transposase used. Some elements are inserted in quite specific target sites, whereas others show very little obvious patterns of target selection, while not being totally random either. Regarding the mechanism of target selection, some transposases directly interact with specific DNA structures, while others use intermediate proteins making the link between them and specific DNA sequences [140]. Transposition is used by some phages which can integrate in many places in their host, like bacteriophage Mu. The term *illegitimate recombination* is sometimes used to refer to transposition, as this recombination process is independant of any DNA sequence homology [86].

As a conclusion, bacterial genome evolution relies on a wide variety of mechanisms. On the one hand, horizontal gene transfer, with conjugation, transformation or transduction, results in the acquisition of new genes (from another bacteria or the environment). Those new elements (except extra-chromosomal ones) are integrated in the host genome via recombination mechanisms. The latter, together with other mutation mechanisms, reshuffle the genomes, potentially modifying the amount of DNA. All these mechanisms allow the bacterial genomes to evolve, and acquire new traits. On the other hand, vertical transfer consists in the transfer of genetic information, including those changes, from parent to offspring, through cell division (see 1.3.3). Rapid bacteria multiplication makes the spread of those changes even quicker.

Now that we understand the mechanisms by which a single genome evolves, the next step would be to study the evolution of a given population of genomes. Indeed, in everyday life, we are witnesses of the effects of genome evolution. For example, some epidemics are caused by a bacterial species which is normally commensal or even mutualist. From there, multiple questions can be raised: how did they become pathogens? Are these changes reversible? Can we prevent them? The answers to these questions all require a common initial step: finding the differences between the genomes of the pathogen and non-pathogen individuals. For that, one needs to compare the different genomes. This is from where comes the last part of our chapter: comparative genomics.

3 Comparative genomics

In the previous chapters, we saw what is a bacterium (beyond the "little bug making ill" preconception), how its phenotype is regulated by its genome content and organisation (chapter 1), as well as the mechanisms by which its genome can evolve over time (chapter 2).

We now want to understand the evolution of not only one bacterium, but a whole population. To do so, we need to compare the different individuals, and in particular their genomes. But what does "comparing genomes" mean?

3.1 Retrieving the bacterial genome

Concretely, the bacterial genome is one or several DNA molecule(s), made of thousands to millions of nucleotides linked together (see figure 1.7). In practice, the bacterial genome is represented by a sequence of letters based on an alphabet depending on the type of molecule: A, T/U, C and Gs for DNA/RNA, or 20 different letters for proteins (see the outside ring of figure 1.10). This binary information is saved in computer files, which are used for the analyses.

Before starting to compare the genomes, it can be important to understand how they were generated, as an analysis can sometimes depend on the method used to obtain the genomic sequence. First of all, how is the molecular genetic material of a bacterium converted into a sequence of letters?

3.1.1 DNA sequencing

The first step to obtain a file with the full sequence of a genome is DNA sequencing. It consists in "reading" the DNA molecule to determine the order of its nucleotides, and store this information into computer file(s). As sequencing is a whole field on its own, I will just touch upon the subject, to introduce the notions needed for the analyses done in this thesis. Likewise, as this is what I use for my PhD subject, I will only deal with DNA sequencing, but I must mention that sequencing can also be done with other biochemical molecules like RNA (after a reverse-transcription to DNA) or proteins.

Bacteriophage ΦX_{174} ssDNA genome was the first full DNA genome to be sequenced. Its

almost 5400 bases were deciphered by Sanger et al in 1977 [158], using a method based on DNA replication (see figure 1.13). The results were manually analysed to recover the initial sequence. Later, a new version of this method was developed, introducing fluorescent signals which can be detected and directly analysed by a computer. The latter was implemented in the first automated DNA sequencer: ABI 370 [171]. In 1995, the first genome of a free-living organism was sequenced: bacterium *Haemophilus influenzae*, with its almost 2 Mbp circular chromosome [63]. This marked the beginning of automated whole-genome sequencing, which culminated with the sequencing of the first human genome in 2001 [43].

Ten years later, 454 Life Sciences company implemented a new sequencing technique in a highly parallel manner, marking the beginning of a new era in the sequencing world [156]: *Next Generation Sequencing (NGS)*. In order to be able to analyse millions of sequencing reactions at the same time, these High-Throughput methods require a preparation step (a.k.a. library preparation), consisting in a random fragmentation of the extracted DNA, followed by an amplification of this sheared DNA. The proper sequencing step then depends on the technology (see figure 3.1). Most of the latter still rely on DNA replication. However, instead of a posteriori analysing partially replicated DNA fragments, the latter are fully replicated and the system detects and stores the signals emitted by each new base incorporated. This signal can be, for instance, a light produced by the release of a phosphate (pyrosequencing [156]), a proton released (Ion Torrent [151]) or the wavelength of a fluorescent nucleotide (Illumina [70]). These second generation methods allow much higher throughput at much lower-cost and have progressively supplanted Sanger-like methods [109]. However, they rely on much shorter reads (a few hundreds of nucleotides compared to the almost 1000 bp of Sanger method) and have an intrinsically higher error rate, caused by the additional library preparation step (errors during amplification, nonuniform coverage of amplified sheared regions, difficulties to amplify regions with high GC%, ...) [51] [151]. For these reasons, Sanger technologies remain in use (as of 2021) for small sequencing projects.

Trying to solve the above-mentioned drawbacks of NGS technologies while still keeping a lower cost than Sanger methods, *third-generation sequencing* (TGS) methods emerged in the last ten years. The latter are distinguishable from NGS methods by the fact that they sequence 1) in real-time (NGS technologies mark a short pause after each base incorporation) 2) a single-molecule which does not need to be amplified (more uniform coverage over the sequence and less GC biased), and 3) produce long reads [51]. These "long-read sequencing technologies" (LRS) are nowadays dominated by Pacific Biosciences (PacBio) [151] and Oxford Nanopore technologies, and produce reads of tens of kbp in average, and up to the current record of 2.3 Mb for Nanopore (although at the cost of a high error rate for the latter) [97] [142].

For now, LRS has not supplanted NGS technologies, which still remain, in spite of huge improvements, cheaper, faster and more accurate (see figure 3.1). Moreover, NGS technologies like Illumina, which already dominates by far the market as of 2021, propose alternatives with "synthetic long reads", using a system of barcodes to pool their short reads into longer fragments [125]. As of 2021, most genomes are still sequenced by NGS, sometimes associated with partial Long-Read sequencing [4]. Thus, most of the genomes I used for my PhD project have been sequenced by NGS methods.

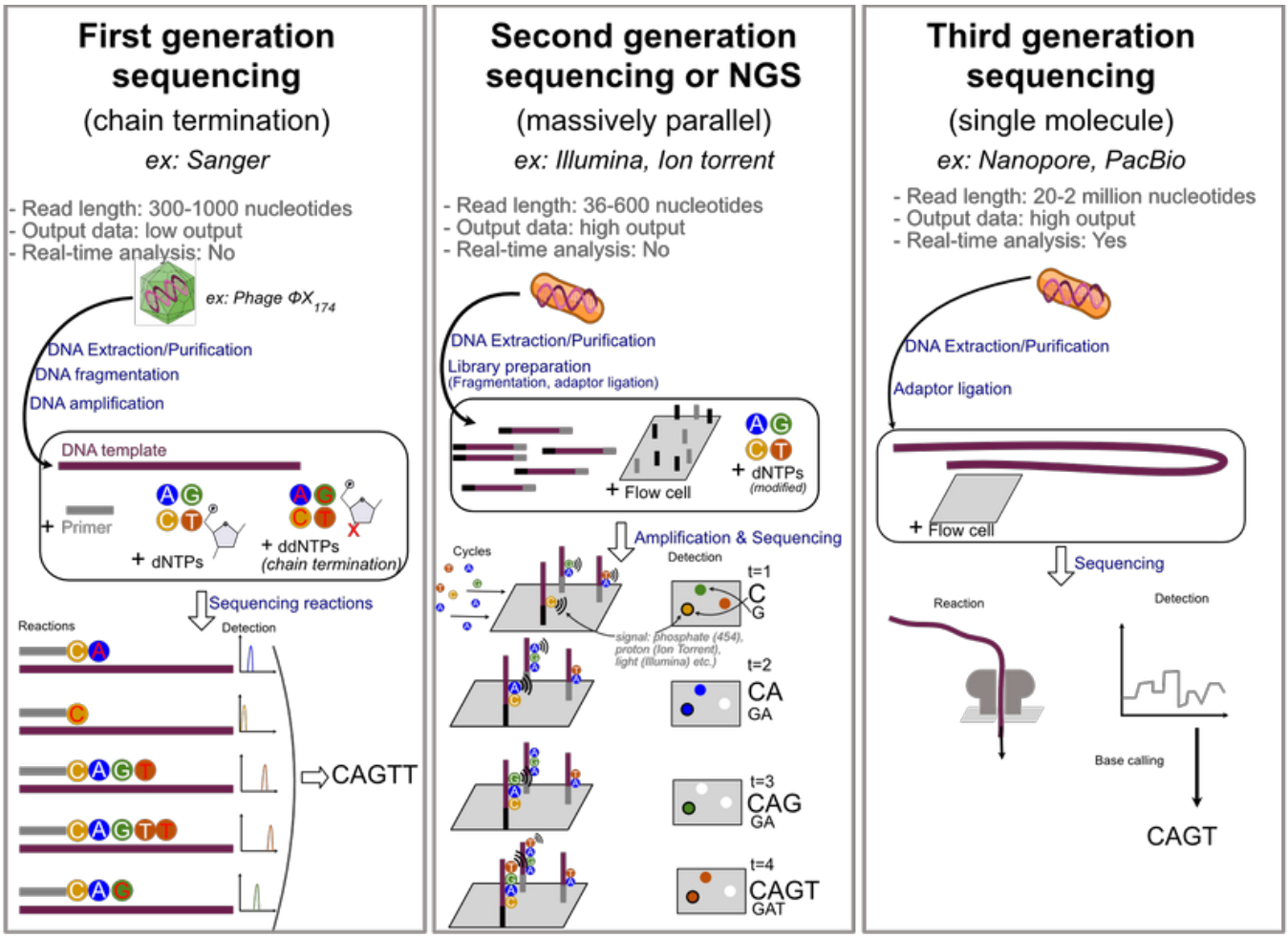


Figure 3.1: Comparison of first, second and third generation sequencing. Figure inspired from [141]. **Sanger-sequencing:** uses ddNTPs (dideoxynucleotides), which lack the hydroxyl groups required to bind to the next nucleotide (see 1.7): the random integration of a ddNTP prematurely ends replication. The resulting set of nested truncated sequences (all starting with the primer, but randomly ending by the insertion of a ddNTP) is analysed to recover the initial sequence.

3.1.2 Assembly

By the time I am writing this thesis, it is not (yet) possible to sequence accurately a full genome in a single read: an assembling step is required to reconstruct the original sequence. This step is often compared to solving a giant puzzle. However, this "puzzle" has some particularities: pieces are overlapping (sequences are randomly sheared and amplified, so that a single nucleotide can be sequenced multiple times), some parts can be missing (parts not read by the sequencer), some can be wrong (sequencing errors), and some can come from other puzzles (contamination). On top of these peculiarities, the difficulty also varies according to the complexity of the organism. One of the biggest challenges is the number of

repeat sequences in the genomes, the latter being indistinguishable while trying to assemble the pieces, especially when repeats are longer than the reads (see figure 3.2) [51]. Most of the time, often due to the aforementioned reasons, the output of the assemblers is not a complete genome, but a collection of smaller sequences called *contigs*. The set of contigs is called a *draft genome*. If the contigs from a draft genome are ordered, we call the result a *scaffold*.

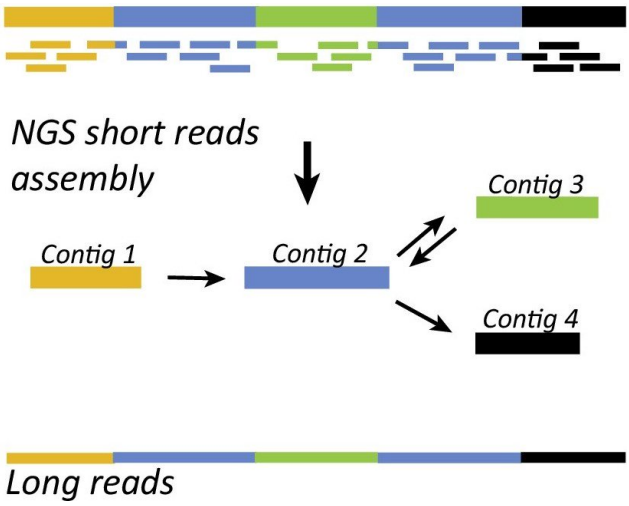


Figure 3.2: Example showing the resolution of a repeated genome region by short-read assembly or long reads. Sequencing a region with two nearly identical repeats (blue) separated by a unique sequence will generate reads corresponding to the upstream region (yellow), the repeats, the sequence between (green), and the downstream region (black), and some reads will overlap the boundaries. Assembly programs cannot assign reads falling in the repeats to unique positions and will assemble those reads into a single contig. The sequence between the repeats cannot be assigned to a unique position either, as it can be placed either upstream or downstream of the 'blue' region. Due to this ambiguity, the sequences upstream of, between, and downstream of the repeats will be assembled into separate contigs. Similar problems arise with structural variants that involve repetitive regions. Figure from [51].

We saw in the previous part that there exist a wide range of sequencing technologies, each one with its own advantages and drawbacks (read length, error rate, error type etc.). Likewise, there are many assemblers, each one adapting its method in such a way that some tools perform better than others on a particular technology, but are less efficient than the latter on other technologies. For example, Canu is specialized in long-read technologies, providing a method able to handle high-noise sequences [104], whereas Velvet is designed for short read sequencing data [206]. Tools like Unicycler are designed for hybrid assemblies, combining short and long-read inputs [201]. SPAdes, first released in 2012 for bacterial genomes assembly based on NGS reads like Illumina or Ion torrent [11], now also provides hybrid assembly methods combining Nanopore, PacBio or Sanger reads with the short-read sequences.

We can classify assemblers into two main categories: *de novo* assemblers and *reference based* assemblers. De novo tools take a set of reads as input and output a complete or draft genome. On the other hand, reference based assemblers take, in addition to the reads, a reference genome as input. This reference is used as a backbone or template to assemble the reads. While using inherently different algorithms to assemble the reads, both types of assemblers share the same computational background: graph data structures that they traverse to extract sequences longer than the reads.

All genomes that I used during my thesis are genomes already assembled by others and registered on the Genbank database (see 3.1.4). Information on the sequencing technology and assembly pipeline used is sadly seldom given, but the vast majority was probably sequenced using Illumina and assembled with SPAdes like assemblers [11].

3.1.3 Annotation

Once the DNA sequences are assembled, the next step, called annotation, consists in identifying genes in these sequences.

One can distinguish two stages in annotation. First, a syntactic annotation (also called structural annotation) identifies the genes encoding proteins. It is based on its definition (presence of start and stop codon, and a length between the two of a multiple of three nucleotides), but also on other criteria like presence of transcription promotors, Shine-Dalgarno sequences and codon usage. Prodigal is a widely used tool for prokaryotic gene prediction [93]. Other features like genes coding for tRNA, rRNA, or CRISPRs are searched with specific prediction software using sequence similarity or structural information [111] [131].

Once genes have been identified, its functional annotation can be performed. This is the most time-consuming step. Basically, it consists in comparing each identified CDS to a database of known proteins, using methods described in 3.2. The reference databases can range from custom databases of private proteins to publicly available databases like the Universal Protein Resource (UniProt, which provides a huge amount of protein sequences with functional annotations), or Pfam (which provides protein families represented by multiple alignments generated using HMM, briefly described in 3.2.2) [182] [62]. Both are used by Prokka, one of the most employed softwares for prokaryotic genome annotation [163].

3.1.4 Databases of bacterial genome sequences

The main databases of publicly available bacterial genomes are housed by the National Center for Biotechnology Information (NCBI). Its major database, GenBank, daily receives original submissions (annotated or not) from individual laboratories and sequencing centers from all over the world [159]. Created in 1982 with a few hundreds of sequences, it now (December 2021) contains more than 1 million bacterial DNA sequences, from more than 65.000 different species, and is still growing very fast. Together with the EMBL Nucleotide Sequence Database (from the European Nucleotide Archive, ENA) and DDBJ (DNA Data Bank of Japan), GenBank participates in the International Nucleotide Sequence Database Collaboration (INSDC) aiming at providing DNA sequences available for free [99] [68] [6]. Daily data exchange between the three partners ensures worldwide coverage and synchronicity.

Also widely used for bacterial comparative genomics, RefSeq database is a curated subset of Genbank introduced in 2000, providing a unique record for each organism [150]. All genomes of Refseq have standardized syntactic and functional annotation, either propagated from the Genbank submission, or calculated by a NCBI annotation pipeline. As of today (December 2021), there are more than 231.000 bacterial genomes in refseq, from more than

40.000 named species. Due to the wide use of NGS methods, only a bit more than 20.000 are completely assembled genomes.

The NCBI assigns a unique identifier (taxid) to each species, and an accession number to each original sequence.

Even if the number of bacterial species sequenced is huge, it must be mentioned that it is biased towards bacteria easily cultivable, and/or with a medical or economic interest. As illustrated by figure 1.14, they represent a very small fraction of the bacterial world.

We now know how we can get DNA sequences of genomes, either starting from their biological strain, or directly downloading them from public databases. This is an essential step, but it is only the beginning, as we still do not know how to compare these sequences. Incidentally, if a functional annotation is needed, comparing sequences is already a required step while retrieving the bacterial genome sequences! But how can we compare sequences, identify and quantify differences, and understand the evolution history leading to these differences? This led to the arrival of a new field in bioinformatics: comparative genomics.

3.2 Comparing genomic sequences

Once we have bacterial genomes in computer files (either from sequencing 3.1.1 or by downloading existing ones in public databases 3.1.4), we need to define methods to compare them. As already mentioned, a bacterial genome is represented by a sequence of letters. From a computational point of view, the most intuitive idea coming to mind while speaking of comparing genomes is thus to compare the strings (or bits, depending on their representation). This chapter will give an overview of the main methods used to compare genome sequences, starting from the less complicated: comparing two genome sequences.

3.2.1 Pairwise comparisons

Optimal alignments

A pairwise alignment is a way of arranging two sequences in order to highlight similarity regions. Intuitively, it can be represented by placing the two sequences one above the other, such that each residue of the first sequence is above either a residue of the second sequence or a gap (represented by a $-$) that has been added to the latter, and conversely (see figure 3.3). With this definition, there exist many different ways to align two given sequences, i.e. to arrange the successive columns of nucleotides and/or gaps. On top of the alignment itself, one can also define a scoring system per column of the alignment. For example one can define that a column composed of a residue and a gap scores -1 point, a column with two identical letters scores $+1$ point and a column with a mismatch (i.e. two different letters) scores 0 point. The sum of scores of all pairs of characters gives a score for the whole alignment (see figure 3.3).

One can also define more complex scoring systems. Substitutions scores can be represented by a similarity matrix, where the cell (i,j) contains the score to attribute when

character i is aligned with character j. The most famous, used to align proteins, is the family of BLOSUM matrices [90]. A gap-penalty scoring system can be associated to this matrix, defining the cost of a gap, potentially varying according to its situation (after other gaps or isolated for example). However, given two sequences and one same scoring scheme, many alignments are possible, some of which having higher scores than others (see figure 3.3). We define the alignment with the best score as the *optimal alignment*.

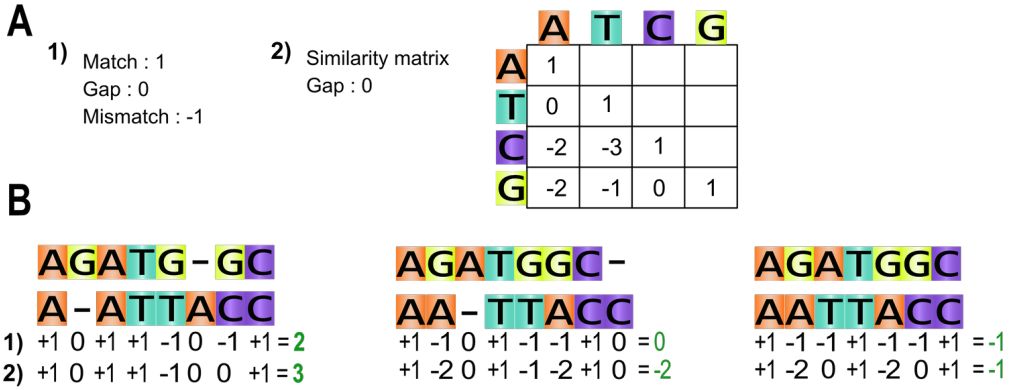


Figure 3.3: A. Two different scoring systems: 1) fix match/mismatch score. 2) with a nucleic similarity matrix. B. Example of three possible pairwise alignments for sequences "AGATGGC" and "AATTACC", with the two different scoring systems.

In 1970, Needleman and Wunsch published an algorithm (today called NW for Needleman-Wunch algorithm) to compute, for a given pair of sequences, the best global alignment score and, at the same time, construct the corresponding alignment(s) [132]. This algorithm needs to score all the positions of the first sequence with all the positions of the second one. Thereby, if the sequences have sizes n and m respectively, the algorithm needs to compute $n \times m$ local scores. To compute the global score, all the intermediate scores are stored in a matrix of size $n \times m$, called the dynamic programming matrix.

However, biological sequences can share common substrings while being very different at the other positions. For example, a Mobile Genetic Element of genome A transmitted to genome B by HGT can be inserted in a region of B which is very different from that of A. In that case, the alignment done by the NW algorithm will give a very bad score and the HGT event can be invisible in the alignment. To take into account such cases, Smith and Waterman presented a variation of NW algorithm, SW, that guaranties to find at least one of the best *local alignment(s)* [172]. The main idea is based on the fact that the scores cannot be negative: bad alignments are scored 0. Thanks to this modification, local alignments are bounded by regions of null scores in the alignment matrix.

These two dynamic programming methods guaranty to reach the best possible score (and consequently the best alignment based on this scoring system), but are time and memory consuming for large DNA or amino acid sequences because they require quadratic time to compute.

Alignment approximations

For large scale analysis, performing alignment by dynamic programming is very expensive in terms of computation time and space. To tackle this problem, fast approximation methods (heuristics) have been developed. During the alignment process, most of the time is spent on filling the dynamic programming matrix. The key idea behind fast heuristics is to quickly define which slice(s) of the sequences are close enough to be aligned to each other or, in other words, select which cells of the matrix "deserve" to be filled. Thus, the not promising alignment slices (corresponding to too different slices of sequences) can be skipped and the time can be spent on parts that will have a good local alignment score.

Most of the computationally efficient algorithms, like FASTA and its extension BLAST developed in the 1990's [120] [3], are based on a *seed and extend* strategy. Before performing any alignment, these algorithms search local high score for words of fixed length k that are called *k-mers*. The first step consists in extracting all the k-mers from a query sequence. Then, each k-mer is used as a *seed*: it is mapped on the second sequence, only allowing matches or substitutions. Each mapped position is scored using a similarity matrix like the BLOSUM matrix, and only the best positions are kept. The best hits are then extended on both sides as long as the score keeps increasing. Finally, if the extended hit has a sufficient score, a local exact alignment is performed. Given that it performs SW computations only from the best *local* score(s), this process does not guaranty to find THE best alignment of the two given sequences. However, it allows the possibility to compare large genomes, where it was impractical using pure SW.

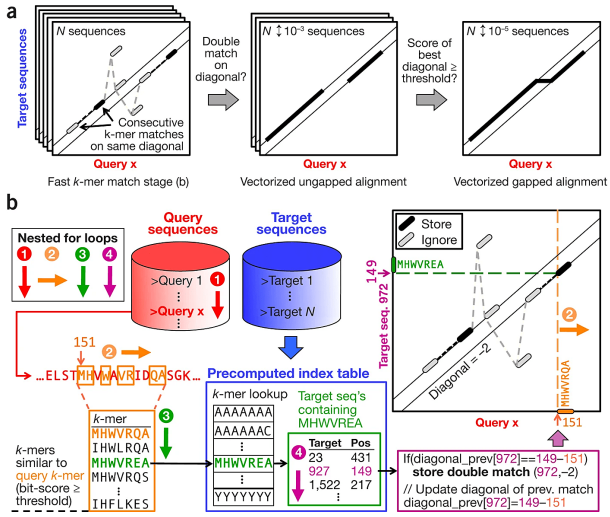


Figure 3.4: Overview of the MMseqs2 algorithm (Figure from [176]).

Such methods continue to be developed and recent works like MMseqs2 and DIAMOND are now outperforming BLAST-like algorithms [176] [22]. The main idea remains the same as for the previous heuristics: fast seeding, extending, and running a SW-like algorithm on the resulting slices. MMseqs2 uses multiple techniques to improve sensibility while keeping the same computation time (see figure 3.4).

The k-mers used as seeds are spaced k-mers (k-mers including "joker positions" which can

map to any nucleotide), which gives the possibility to include errors in the middle of the seed (in addition to the substitutions), and to use bigger k-mers (which are thus more specific to their sequence). Two consecutive shared k-mers lying on the same diagonal (the gap between the shared k-mers is the same in both sequences) are paired and extended to generate pre-alignments. Finally the pre-aligned diagonals are linked using a very well engineered SW algorithm (see figure 3.4a). Concretely, to speed-up the k-mer sets comparisons, most tools index them in a Hash table (briefly, each k-mer is associated to a binary value, which is further affected to a memory slot). It is easier to compute k-mer intersections using hash values because the computer can directly check their presence/absence at the right memory slot instead of permorming multiple comparisons. To be even faster, the authors of DIAMOND index spaced k-mers of both sequences and merge the indexes efficiently to find common k-mers. It is faster than checking independently all the entries.

As for the previous algorithms, these heuristics do not guaranty to find the optimal alignment. Both MMseq2 and recent DIAMOND versions are showing very good performances to generate alignments. They are at least 10 times faster than the other methods for the same accuracy. However, the computation time is still prohibitive to make a all vs all comparison of tens of thousands of genomes.

Pairwise comparison without alignment

Since the alignment process needs to fill a matrix (or a submatrix for heuristics), the time complexity will always be proportional to $n \times m$. To bypass this limit, new algorithms have been developed to compare sequences without the need of an alignment step.

The main idea is that two sequences for which we can obtain a very good alignment have a high probability of sharing multiple substrings. Alignment-free methods are based on this observation, which they take on the other way around: two sequences sharing a high number of substrings (of fixed size k for better algorithm properties) have a high probability to be very similar. In this way, comparing the k-mer content of the two sequences should be a good proxy for sequence alignment. However, generating efficiently the sets of all k-mers of a sequence is not that easy: even if the algorithm is linear in time, it requires a lot of memory (although still a lot less than the full dynamic programming matrix). To limit the memory requirement, *sketching softwares* have been developed: they take advantage of that time property to be fast and only store a subset of the k-mers to be compact in memory. Tools like Mash or sourmash carefully select the subset of k-mers to truly represent the original datasets [135] [186]. If two initial datasets A and B share p% identical k-mers (p represented by the Jaccard index J), they construct subsbets S_A ans S_B such that the probability of collision (meaning of getting the same k-mer) between any k-mer of S_A and any k-mer of S_B is roughly similar to p (see figure 3.5). Other tools like Hyperminhash and Dashing propose memory optimizations [205] [10] and tools like BinDash propose construction time optimizations [207].

However, they produce a degree of divergence that is not linear with the accumulation of sequence substitutions, and tend to over-estimate long genetic distances. Very fast in practice they can be used to quickly estimate a similarity metric between two sequences.

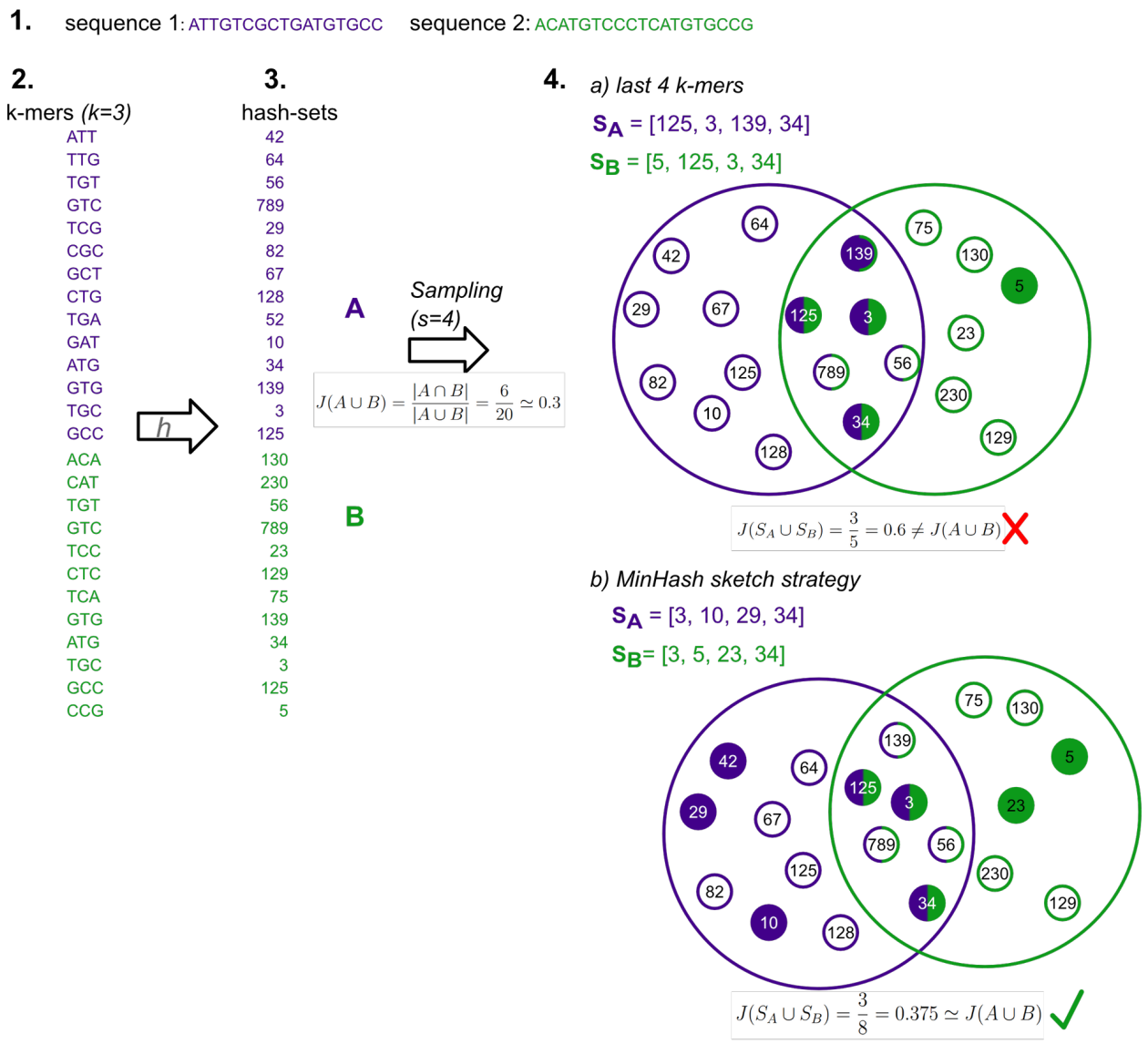


Figure 3.5: Overview of the strategy of sketching softwares. **1.** The two input sequences are decomposed into **2.** their constituent k-mers. The latter are passed through a hash function h to obtain **3.** two sets A and B of 32- or 64-bit hashes, depending on the input k-mer size. The Jaccard index J is the fraction of shared hashes out of all distinct hashes. As storing all hashes requires too much memory, subsets of both hash-sets, S_A and S_B, must be selected. **4.** Two different sketch strategies: little circles represent A and B, and filled circles represent S_A and S_B. Strategy *a)* takes the last 4 hashes of each hash-set, and leads to a biaised subset. Strategy *b)* takes the four smallest hash values, corresponding to MinHash strategy. Because S(A∪B) is a random sample of A∪B, the fraction of elements in S(A∪B) that are shared by both S(A) and S(B) is an unbiased estimate of J(A,B). Figure adapted from [135].

3.2.2 Other comparison methods

In part 3.2.1, we saw how to obtain the optimal alignment between two sequences given a scoring system. It is possible to extend these exact pairwise alignment algorithms to produce an optimal alignment over n sequences ($n > 2$). Similarly to pairwise alignment, we can represent it with a matrix of n rows. One can then define a score for the alignment of the n nucleotides (or gaps) of column i, and sum all local scores to get the full MSA score. However, the score matrix requires as many dimensions as the number of sequences to align, and could only be computed in exponential time (the exponent being n). As this is not applicable in practice, MSA tools using approximations have been developed.

Most of the current tools have similar approaches to solve the MSA problem. They first determine all pairwise distances (or good approximations of the latter) of the n sequences. Then, they add sequences in the whole alignment one by one, starting from the closest ones. This method, called progressive alignment, is performed by tools like Clustal-Omega [184], Mafft [100] and some versions of MUSCLE [56]. Very recent work (late 2021, not yet published) on MUSCLE v5 introduced parameter perturbation (bootstrap) to improve alignments and seems to be better than previous methods on MSA benchmarks [55].

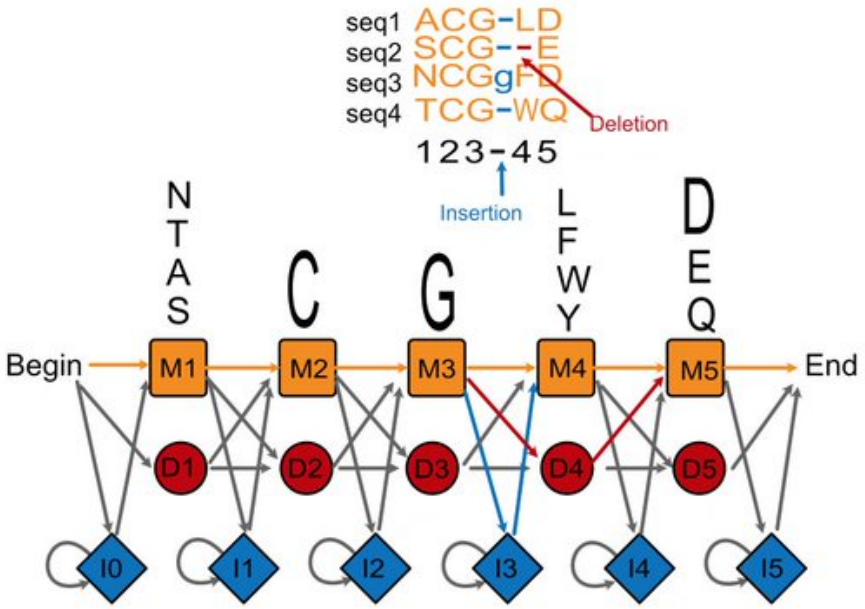


Figure 3.6: Example of a MSA and its corresponding profile-HMM. The boxes in orange are the match states (M). In the M state the probability distribution is the frequency of the amino acids in that position. The row of blue diamond shapes are insert states (I) which are used to model highly variable regions in the alignment. The row of red circular shapes are delete states (D). These are called silent states since they do not match any residues. The final probabilistic model conveys the estimation of the observed frequencies of the amino acids in each position, as well as the transitions between the amino acids derived from the observed occupancy of each position in a multiple sequence alignment. In this model, we show all possible transitions, but with the example above, grey arrows never happen. Figure from https://www.ebi.ac.uk/training/online/

MSA can also be used to build profile-HMMs, which are used to search for homology in databases. Compared to classical MSA, their particularity is their ability to capture position-specific changes in amino-acid sequences. Indeed, protein sequences do not evolve uniformly: some parts of a protein evolve faster than others, and are thus less conserved. Profile-HMMs represent this particularity by providing a position-specific scoring scheme instead of a global alignment score. They are a sort of generalisation of consensus sequences that allow insertions and deletions. They associate each position of the sequence to three different types of states: a match state (the probability of finding an amino-acid or another), a deletion state (probability of matching no residue), and an insertion state between two positions (probability of having an insertion). All states are linked by transition probabilities, following a first-order Markov chain: each transition from a state to another on the profile only depends on the result of the preceding transition, and not on past transitions. An example is given in figure 3.6.

Softwares like HMMER use profile-HMMs to perform profile-sequence alignments [54]. This allows to align distantly related sequences, by identifying conserved domains. As already mentioned in 3.1.4, Pfam is the biggest public database of profile-HMMs, which are used to search homology for protein functional annotation.

3.3 Comparing a whole set of genomes

Once we know how to compare genomic sequences, we could imagine that comparing a full set of genomes is quite straightforward: comparing their respective sequences. However, even if many bacteria have their genome in a single replicon, a significant number have several replicons. In that case, which ones should be compared together? Even within a group of genomes having a single replicon, we will see that comparing their whole genomic sequences is not always consistent with expected relationships, mostly due to the effects of HGT. Finding a good way to compare individuals is far from being simple.

3.3.1 Back to the definition of a bacterial species

To start, let's go back to the definition of a species. In general, it is defined as a group of organisms able to reproduce between each other, typically by sexual reproduction. However, for bacteria, as they reproduce asexually (see Bacterial reproduction), the concept of species is more complicated, and is still up for debate in the microbiologist community, as already mentioned in chapter 1.4. When the first "animacules" were observed through a microscope in the 17th century, they were considered as a *single species* of pleomorphic individuals. Since then, the concept of bacterial species has progressively evolved in parallel to the development of new laboratory techniques, which allowed the retrieval of novel information. In this part, I will quickly go over the main steps shaping the notion of bacterial species, in order to understand the current situation and its new challenges.

In 1872, Ferdinand Cohn first tried to distinguish groups of bacteria according to their microscopic *morphology* (see figure 1.4) [91]. However, the relative simplicity of bacterial shapes was not adapted to describe the wide diversity of bacteria.

In the early 50s, with the development of *pure cultures of bacteria* microbiologists transposed the "reproduction compatibility" criterion of the original definition of a species (i.e. organisms able to *reproduce between each other*) to "*culture compatibility*". Thus, bacteria were considered to belong to the same species if they can culture together and "are accepted by bacteriologists as sufficiently related" [91], this relatedness being based on phenotypic, physiological and biochemical characteristics of the culture. However, even with the many possible combinations of these new criteria, the bacterial diversity was still underestimated.

A little later, microbiologists suggested that bacteria might be better classified by directly considering the source of these observable traits: their DNA. With the *discovery of the structure of DNA*, DNA-DNA hybridization (DDH) became the "gold-standard" to determine the relatedness between strains [198]. Briefly, this method consists in mixing DNA molecules (previously separated as ssDNA) from two individuals, and observing renewed ds-DNA molecules. Indeed, two sequences with a high degree of similarity will tend to bind together, even if they come from two different genomes (see figure 3.7). The percentage of hybrids (two strands from two different organisms) is used to estimate the degree of similarity between the two strains. With the development of *sequencing technologies*, the standard 70% or greater DDH species delineation was progressively replaced, first by rRNA (mostly 16S rRNA) sequence comparison, and later by *whole genome sequence comparison*.

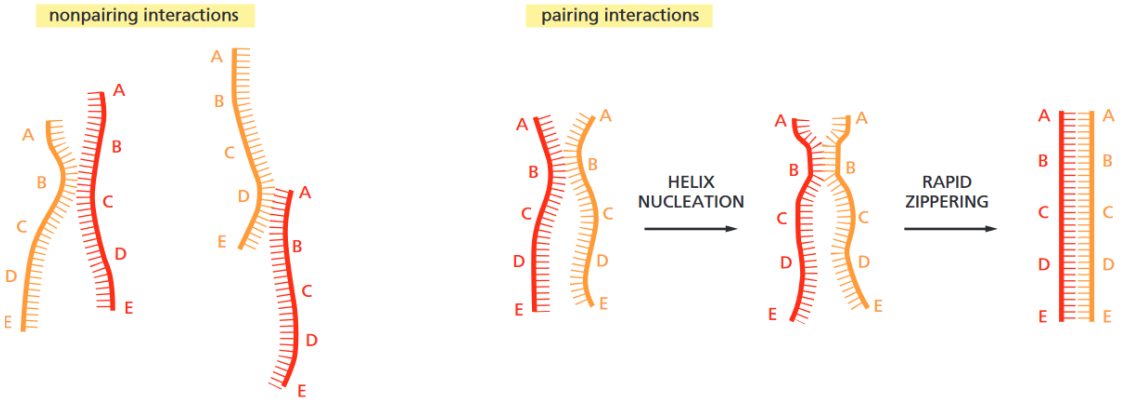


Figure 3.7: DNA-DNA hybridization. DNA double helices can re-form from their separated strands in a reaction that depends on the random collision of two complementary DNA strands. The vast majority of such collisions are not productive, as shown on the left, but a few result in a short region where complementary base pairs have formed (helix nucleation). A rapid zippering then leads to the formation of a complete double helix. Figure from [2].

However, in the early 90s, a study of strains from the genus *Aeromonas* highlighted a puzzling inconsistency between the last two methods: although they observed nearly identical 16S rRNAs, the DNA-DNA re-association values were very low [123]. This lack of congruence between the two methods, suggesting that some DNA regions were not shared by all strains, was confirmed by several studies among which a study of six newly sequenced strains of *Streptococcus agalactiae*. The annotation of these sequences revealed a serendipitous variety of genes, many of which being present only in one strain [180]. Supporting the importance of

HGT, responsible for gene content variability in addition to the expected sequence variability, these observations led to a change of paradigm in comparative genomics: comparing common and specific genes of a set of genomes instead of their whole sequences.

3.3.2 Moving towards the pangenome concept

In the early 2000s, Konstantinidis et al introduced the Average Nucleotide Identity definition, based on the comparison of conserved regions between a pair of strains. [103] They showed that a species can often be defined by a group of organisms within which all pairs of DNA sequences have an *ANI* higher than 94%. In their case, regions are genes, and they are conserved between two genomes if they have a BLAST match of at least 60% overall sequence identity. The ANI value is then the mean of all these BLAST matches. Since then, many other algorithms have been developed to calculate or approximate the average nucleotide identity of the total genomic sequence (not necessarily genes) shared between two strains [204].

Pushing the reflection further, Tettelin et al coined the term *pangenome* (from the Greek $\pi\alpha\nu$, meaning "whole") to describe the complete inventory of genes in their group of *Streptococcus agalactiae* strains [180]. They showed that a bacterial species can be described by its pangenome, which includes a core genome (genes shared by all strains) and a dispensable or accessory genome (divided into genes shared by two or more strains and strain-specific genes) (figure 3.8A). Biologically, core genes are most likely to be essential for survival and growth (aka housekeeping genes), whereas dispensable genes provide supplementary functions with potential selective advantages (ecological adaptation, virulence mechanisms, antibiotic resistance, colonisation of new host...) to some strains [127].

This pangenome definition of a bacterial species led to another important discovery. Although we can describe a species by its gene repertoire, we do not know how many strains are needed to be sequenced in order to fully describe the species. Based on datasets of species for which several sequences were available, Medini et al showed that, in most cases, each new strain added to a dataset increased the gene pool [127]. Although they did their study on less than ten strains per dataset, mathematical extrapolations surprisingly showed that this would still be the case even after sequencing hundreds or even thousands of genomes, leading to the term 'open pangenome' (see figure 3.8B). As an example, a study on 1294 *E. coli* strains found a pangenome of more than 75000 families of homologous genes, 44% of which are singletons (genes present in only one of the 1294 strains). Even if the number of new genes decreases while considering more and more strains, it has been estimated that each newly included strain would still increase the pangenome by 26 genes on average [190].

This behavior is mostly due to Horizontal Gene Transfer, which constantly brings new genes from unrelated organisms (within but also between bacterial species). Tettelin et al showed that the pangenome size p as a function of the number of strains in the dataset n can fit a Heaps' law, such that p is proportional to n^{α} [181]. Finding the parameter α for a given species dataset amounts to estimate the openness of its pangenome, i.e. the species diversity. However, it must be mentioned that these models do not (yet) account for phylogenetic structure and can be affected by sampling biases.

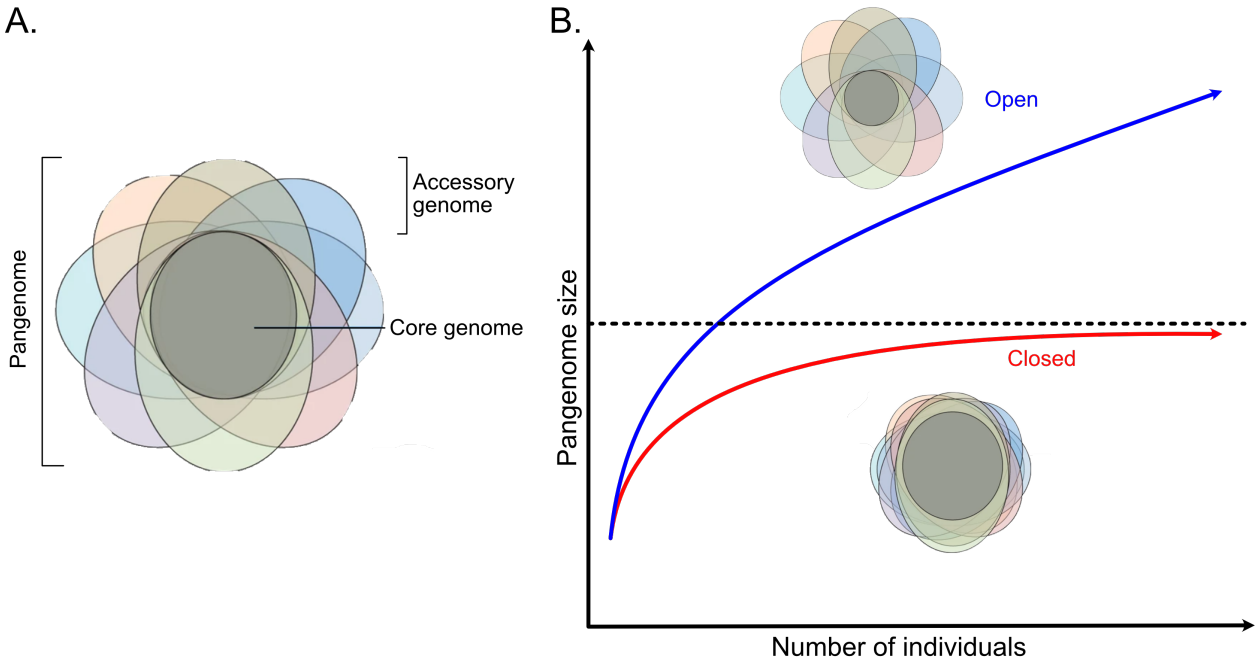


Figure 3.8: Pangenome concept of a bacterial species. A. Composition of a pangenome. B. Pangenomes vary extensively in size and in proportion of core vs accessory (or dispensable) gene content. They can be either open (each new strain adds new genes) or closed (after a certain number of strains, new ones do not add new genes). Figure adapted from [126] and [74].

The term pangenome was further extended to broader taxonomic groups (genus, phylum, and even all bacteria [110]) and to all living organisms (including eukaryotes) [40]. This term is now used to describe *the inventory of sequence entities in a group of organisms*, and not necessarily of a given species. While the first and most intuitive representation of a bacterial pangenome is a set of individual proteins (as most of the bacterial genome codes for proteins, see figure 1.9), this extended definition allows new representations of a pangenome. Hence, instead of being a gene, the sequence entity can be a protein, an arbitrary sequence chunk, a k-mer or else a group of concatenated genes (like operons).

For example, Panseq identifies pangenomic regions by aligning each sequence of the dataset to a reference sequence, using the MUMmer algorithm [108] [106]. Using a reference free approach, Splitmem represents the entire population as a pangenome encoded in a compacted de Bruijn graph [122]. These two softwares define pangenomes as the complete non-redundant set of sequences found in all individuals. This sequence-centric approach is widely used in eukaryotic pangenomics, as the major part of eukaryotes DNA is non-coding, but their intergenic regions fulfil important functions [40]. Even more distant from the classic gene-centric approach used for prokaryotes, Piggy generates pangenomes based only on intergenic regions [185].

From now on, in this manuscript, we will stay with the term pangenome applied to bacterial coding genes.

3.3.3 Pangenome families computation

Mathematically, the pangenome is the union of gene families in a group of bacteria. It is composed of the core-genome, the intersection of these gene families, and the accessory genome, its complement. That being said, I want to raise an ambiguity. Many pangenome studies use the term "genes" in place of "gene families". For example, Tettelin et al found a pangenome of *2713 genes*. In reality, there are a lot more than 2713 genes in eight strains, as the genome of one *Streptococcus agalactiae* strain already contains more than 2100 genes. So, it is important to understand that a bacterial pangenome is made of gene *families*, and not individual genes. In comparative genomics, a gene family is a set of homologous genes, meaning genes with high similarity due to a shared ancestry. So, before any core genome can be computed, one must define these pangenome families.

Here, we are starting from a set of n genomes which have already been assembled (at least as drafts) and annotated (at least syntactically). Thus, we have a set of g genes, g being the sum of genes from all genomes. For each gene, we have its nucleic and protein sequences, as well as all information on its origin (which genome, on which contig and which position on the latter). From there, pangenome computation can be divided into two main steps. First, all pairs of genes are compared. These g^2 comparisons can be represented by a symmetric matrix, which is then used (as it is or after some modifications) to cluster the different genes into homologous families. Figure 3.9 shows this process on a toy example. Table 3.1 (page 76) shows the different methods employed by several pangenome tools.

Comparing genes from all genomes

The first main step of pangenome construction is thus inferring similarity between sequences. There are many different ways to do so. The most intuitive one, which is used by many tools, is to compare all pairs of genes from all genomes. For that, as we saw in the previous part, we can use alignment methods which calculate the optimal alignment score between two sequences, using a deterministic algorithm (like Smith-Waterman). To compute the first bacterial pangenome, made from 6 *Streptococcus agalactiae* strains, Tettelin et al used SW to compare all proteins [180].

In the following years, the first pangenome tools developed, like GET_HOMOLOGS (a pangenome analysis platform accessible to researchers with few computational skills), PanOCT, PGAP or the first version of EDGAR also used the all-against-all comparison method, calling BLAST [41] [64] [208] [18]. In addition to be faster than SW for each alignment, BLAST algorithm is optimized to skip the alignment of too different genes (whereas SW has to align all pairs of genes). However, the number of comparisons is still proportional to the square of the number of genes (and thus of genomes). This makes these tools unable to handle datasets of hundreds of genomes.

To adapt to the increasing dataset size, new tools, like Roary, PIRATE or MetaPGN, introduced a pre-filter step before using these alignment methods [137] [13] [144]. Indeed, as we are looking for genomes from a same species, many genes are nearly identical. Taking advantage of this property, these tools use a quick clustering method, CD-Hit, to remove redundancy [67]. This greedy clustering algorithm first determines a centroid gene (here,

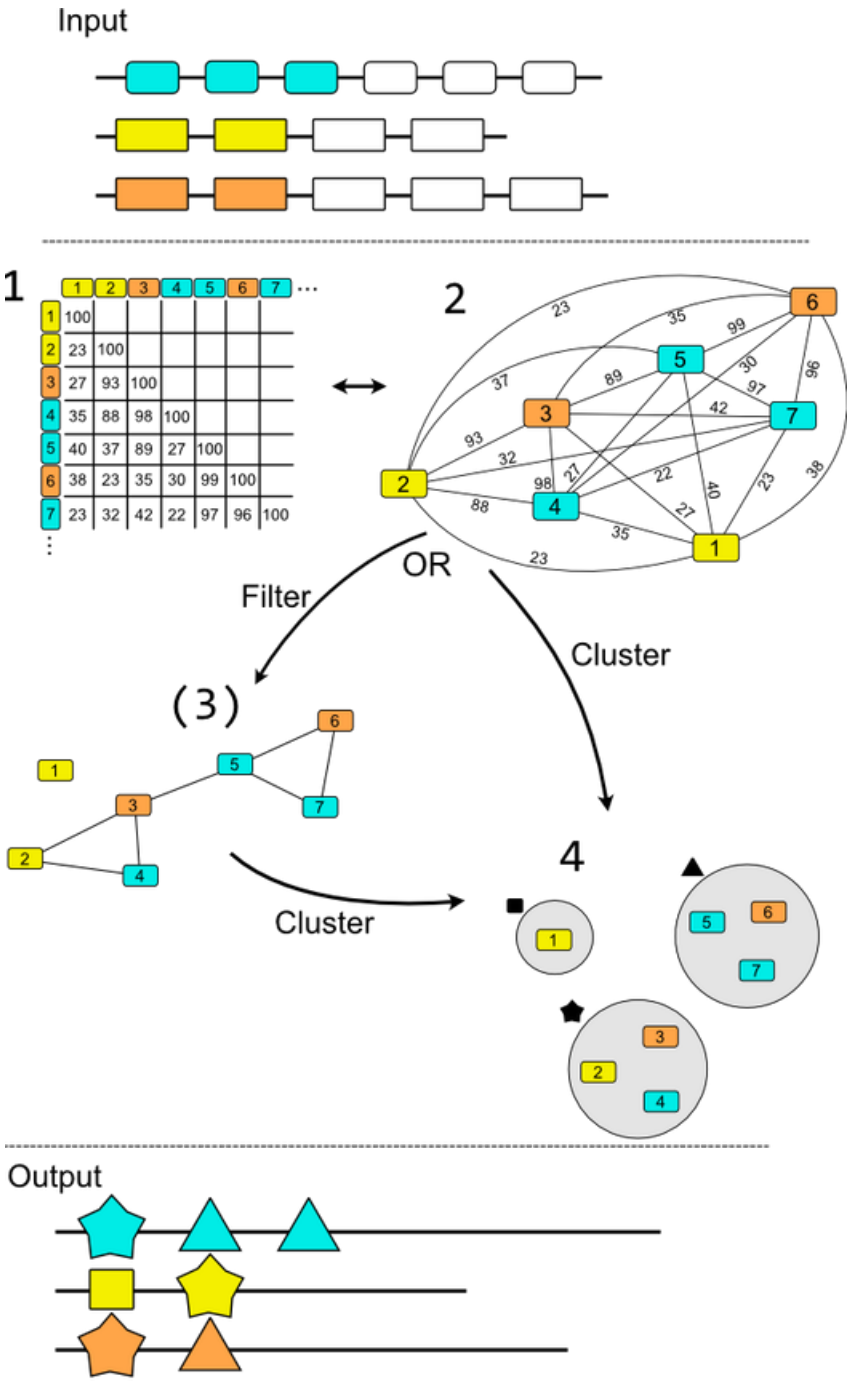


Figure 3.9: General method for pangenome families computation. **1.** Pairwise comparisons of all genes. Results can be represented in a matrix or as a graph (**2.**). **3.** Optional: filtering the resulting graph. **4.** Clustering genes based on similarity to get clusters of homologous gene families. The output gene families are represented by different shapes on the input genomes. Table 3.1 shows the methods employed for each of these steps by several pangenome tools.

the longest sequence), and aggregates all genes within at least $n\%$ similarity from it into a cluster. Genes with lower similarity are considered as new centroids. For the pre-filter, the similarity threshold n is generally fixed at 98 to 100%. This leaves a substantially smaller set of genes (being the set of centroids of the clusters) to be compared all-against-all with the time consuming alignment method. However, when the pangenome is open (which is the case for most species), the higher number of genomes in the dataset, the bigger the pangenome. This implies that new accessory gene families (and consequently new strain-specific genes) will emerge with each new strain added. By definition, these new genes will be different from those previously included, and thus not filtered out by the pre-filtering step, increasing the all-against-all comparisons needed to be computed.

To overcome this new road block, new pangenome tools use ultra fast all-vs-all sequences comparison algorithms. For example, panX uses DIAMOND, and Sonic Panaroid uses MMSeqs2 to search for homology between all pairs of genes [52] [22] [42] [176]. Even if the latter is not strictly speaking a pangenome tool, and not specific to bacteria, it infers similarity between sequences, which is an essential step for the definition of gene families. Given the important increase of speed of these new similarity search algorithms, pangenome computation time using such tools is similar or even faster than the one of computing a pangenome with a pre-filter followed by a slower similarity search method like BLAST. Moreover, the absence of a pre-filter avoids the potential bias introduced by the choice of the representative sequence of such pre-filter clusters. However, a few tools like PIRATE or PEPPAN combine both methods: a pre-filter step using CD-hit or Linclust [176] with a combination of BLAST and faster all-vs-all comparisons tools like DIAMOND on the representative sequences.

Post-process of pairwise comparison scores

The results of this first step can be summarized in a symmetric matrix of size g (see step 1 of figure 3.9): whatever the method used, each pair of genes has a score (alignment score, percentage of similarity/identity, distance between sequences, number of k-mers shared, or any other metric). The value of this score depends on the comparison method, as well as on the parameters used to run it: we cannot directly compare two similarity matrices to compare pangenome tools.

This matrix can also be represented as a graph, where each node is a gene, and the edge between two genes represents their comparison score. Strictly speaking, this corresponds to a weighted graph: each edge has a weight, corresponding to the score between the two nodes it connects. However, other types of graph can be computed from the same similarity matrix.

Some tools choose a threshold above which an edge is added between two genes. A pair of genes with a score lower than this threshold will not be connected in the graph. The resulting graph is thus not weighted, but some of its edges have been filtered. This threshold is most of the time a pre-defined value. For example, to compute the pangenome of their six strains of *Streptococcus agalactiae*, Tettelin et al kept only alignments with a minimum of 50% identity over more than 50% protein/gene length before the clustering step [180]. PGAP, PIRATE and PEPPAN kept the same values as default thresholds to filter their BLAST or Diamond hits. On the other hand, MetaPGN uses more stringent values: it keeps only BLAT

(Blast-like Alignment Tool) hits with identity higher than 95% and more than 90% overlap [101]. Other tools like EDGAR 2.0 do not fix a default value, but use a statistical method to automatically adapt the threshold to the dataset [19].

Sometimes, the similarity threshold is only the beginning of a more complex matrix filtering process. For example, PanOct starts with eliminating very divergent pairs of genes by excluding BLAST hits with less than 20% identity and less than 1% of match length, but then performs many other operations, in order to separate paralogs (see 3.3.4) [64].

Clustering into families

Once the (potentially filtered) similarity graph is built, the next main step of pangenome computation is to group (i.e. cluster) genes into homologous families. Again, clustering methods vary a lot between the different pangenome programs, and also depend on the type of graph previously constructed (weighted or threshold graph for instance). Pangenome computation is a very complicated problem because it is computationally very expansive and also because the "truth" is not known. All algorithms have their advantages and drawbacks, but it is not trivial to determine which one is better. Furthermore, the relevant definition of pangenome and the method to identify it can differ according to the related biological application.

Sometimes, the filtering step is sufficient to determine the gene families. For example, EDGAR or MetaPGN clustering step "only" consists in retrieving the connected components, each one forming a pangenome family. This can be done by transitivity, sometimes called *single-linkage* algorithm: two nodes are in the same cluster if there is at least one path to connect these two nodes. When the matrix is not filtered, or this filter is not considered as enough to determine the homologous families, other clustering methods are applied.

Some tools, like BPGA or PanOct, use greedy clustering algorithms like CD-HIT (explained in 3.3.3) or USearch (same as CD-HIT, but with a predefined order of the sequences instead of taking the longest sequence). These tools must provide a similarity threshold to determine if a new sequence can be assigned to an existing cluster, or create a new one. They have the advantage to be quite fast and intuitive. However, they depend on the choice of the centroid, which can sometimes induce biases.

Instead of agglomerating sequences around a centroid, the Markov CLustering MCL algorithm starts from the complete graph, and tries to detect and split the different communities [57]. This algorithm is based on a weighted version of the similarity graph, where the weight of an edge between two homologous genes i and j is the probability of stepping on node j from node i in random walks. Two processes are alternatively applied to the corresponding Markov matrix P. First, a process of expansion consists in calculating the probabilities for a random walk of size r (r pre-defined). Then, an inflation step is performed on the resulting matrix, in order to enhance the differences between low and high probabilities, and remove the less probable edges. The two steps are repeated until reaching a stable state. Finally, the pangenome families are the resulting connected components. This algorithm is widely used among pangenome tools (PIRATE, Roary, PGAP, panX...). However, it requires an inflation parameter, which is way less intuitive to impose than the similarity threshold of a CD-Hit-

like methods. It is thus very difficult to determine, and the results are more complicated to interpret.

3.3.4 Determine categories of each pangenome family

Once clusters (i.e. homologous families) are defined, they can be classified into core families, accessory families or strain-specific genes, based on their prevalence in the different genomes. The theoretical definition of the core genome, being 'the intersection of genes shared by all genomes' is now much clearer: we can consider as core all families having a member in all genomes. Similarly, families coming from one single genome are strain specific, and the remaining ones are accessory families. But, here again, there are some hidden subtleties, making the core families determination difficult. I will here mention only two main problems, with a few ideas on how to deal with them.

Dealing with paralogs

As already mentioned, two sequences are called homologs if their high similarity is due to a shared ancestor. But this shared ancestor can come from three main phenomena (see figure 3.10). When the two genes come from a speciation event (creation of a new species), they are called orthologs and are transmitted by vertical transfer. On the other hand, xenologs come from an horizontal gene transfer. Finally, paralogs arise from a gene duplication event. As shown in figure 3.10, in addition to the ortholgs, some pangenome families can contain paralogs or xenologs, leading to the presence of several genes for a same genome. Even if, at first sight, it does not seem to be a problem, this raises several questions. Indeed, most of the time, a core genome is computed to consequently build a phylogeny of the genomes in the dataset. To do so, one needs one gene per genome, in order to give even size alignments for each genome to the phylogeny tree inference software. The question is thus on which of the genes should be kept in a genome with paralogs. As all sequences are very similar by construction, a simple way could be to choose any of the two (or more) genes for the genome with paralogs. However, paralogs usually diverge after duplication to take different functions, and their sequence may vary according to this new function. Including paralogs instead of orthologs in phylogenetic analyses could thus lead to mis-interpretation [64].

Although very common in comparative genomics, differentiating paralogs and orthologs remains a difficult problem. This problem can be handled at different steps of the pangenome computation.

A first helpful assumption is that genes which have been duplicated many generations ago have diverged: although they are similar enough to be considered as homologs, the pairwise distance between a paralog and an ortholog should be smaller than between two orthologs. Based on this, tools like PanOct, GET_HOMOLOGS and Pandelos filter their similarity matrix by identifying Reciprocal Best Hits (RBH), and removing unidirectional best hits, as exemplified in figure 3.11. However, very recent duplications can be indistinguishable, even with this method (see blue gene in figure 3.11).

To handle these situations some tools use, instead of or in addition to RBH, a genome

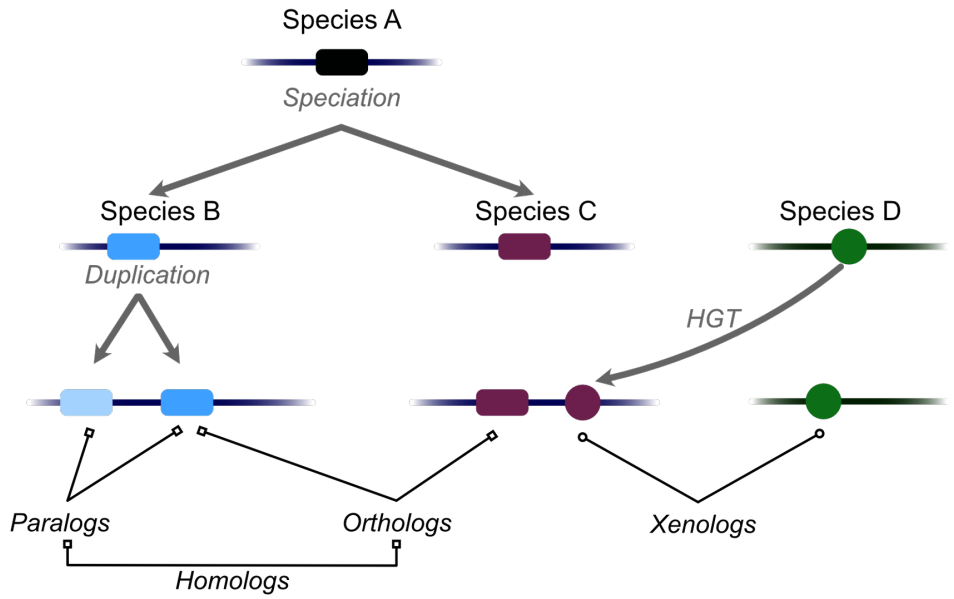


Figure 3.10: Relationships between similar sequences. A pangenome of species B, C and D would contain two different homologous families: squares and circles.

context criterion: two orthologs are more likely to be between the same genes in different genomes, while paralogs can be inserted somewhere else, their different function potentially requiring different regulation paths. This gene co-linearity criterion can be incorporated at different stages of the pangenome computation. PanOct and GET_HOMOLOGS add this condition to their similarity matrix filtering step, to keep only edges between genes having a conserved neighborhood [64] [41]. Other tools, like Roary or MetaPGN post-process the pangenome families obtained after the clustering step, to re-split those having several genes in a same genome [137] [144]. A problem of this approach is that duplications tend to be created in tandem which renders this criterion useless [191].

panX uses a completely different approach to post-process its pangenome families. For each family, a phylogenetic tree is inferred, and branches are pruned according to a given paralogy score. Subtrees define the orthologous families.

Pandelos only post-processes inconsistent families: families which contain two genes belonging to the same genome but are not accounted as paralogs (i.e. not connected by an edge) [21]. It performs the Girvan-Newman algorithm to separate the communities, by removing the edges which are most likely to be "between" two communities, based on the betweenness scores (number of shortest paths passing through this edge) [73]. As such, post-processed families can still have several genes in a same genome.

Dealing with annotations

So far, we have seen that the pangenome is influenced by many aspects: alignment method, similarity parameters (%identity, coverage), clustering method, and the way to deal with paralogs. But we must not forget that, first of all, the pangenome definition is based on

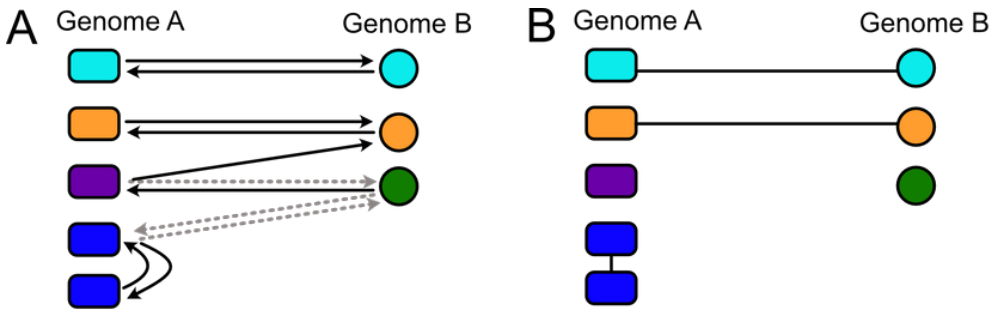


Figure 3.11: Example of Reciprocal Best Hits. A. Hits with a high similarity (obtained by any comparison method, with or without alignment) are represented by arrows. Black arrows represent the best match for each pair of genes. B. A pangenome tool using a filter based on Reciprocal Best Hits would keep these three edges. Cyan and orange genes are likely to be orthologs, but blue genes are rather paralogs.

genes: it is thus annotation-dependant [194]. Hence, inconsistent annotations can greatly impact it, and this will affect the identification of the core genome.

One of the "easiest" ways to limit annotation inconsistency is to make sure that all genomes were annotated with the same pipeline. Indeed, each annotation software has its own criteria regarding the minimum length of a gene, the choice between alternative starts, codon usage etc. To control this variation, some pangenome tools, like MetaPGN and panX include annotation steps.

Independently of the annotation method, the increasing number of draft genomes makes the core genome determination even more difficult. Indeed, poor quality sequencing/assembly can lead to many errors. Genes can be split between two or more contigs and thus detected as two different genes by the annotation tool, corresponding to two parts of the gene. Other genes are only partly assembled, and the assembled part is too short to be recognized as a CDS. Single-base insertions/deletions sequencing/assembly errors can also lead to consider artificial indels as frameshifts, thus missing the gene. This means that some pangenome families which do not have genes in a few genomes might, in reality, be part the core genome. In 2005, Tettelin et al already tried to limit this problem, by running gene-against-genome tblastn to compare genes of a gene family against the genomes missing in it [180]. PIRATE software also performs a DNA search to try to recover miss-annotated genes when they have "almost-core" families. However, if the genome is too fragmented, even the DNA search would fail to recover its gene. Also, this method may result in the inclusion of pseudo-genes in the core genome.

Moving from core to persistent genome

When inferring pangenomes of thousands of genomes, the probability that at least one genome has a problem becomes quite high, and the size of the core genome decreases drastically, sometimes down to zero (see figure 3.12).

To tackle this problem, in very big datasets, the core genome is usually replaced by

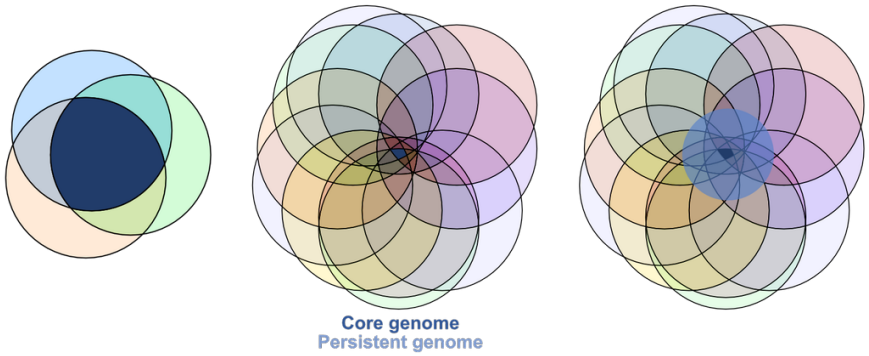


Figure 3.12: Evolution of the core genome size (dark blue zone) with the number of genomes in the dataset (number of circles). When the core genome is too small, one can use the persistent genome (light blue).

the persistent genome, a less stringent notion proposed by Acevedo et al in 2013 to design families with genes present in at least $N\%$ of the genomes [1]. However, as any threshold-based method, the choice of the threshold is difficult, and is most of the time arbitrary. Peppan considers a family in the "relaxed core" if it contains members in at least 95% of the genomes [209]. PPanGGoLiN proposes a statistical method to infer it [71].

Tool	Comparison method	Similarity filtering	Clustering method	Post-process of pangenome families	Publication DOI
Tettelin et al [180] (2005)	SW: all-vs-all proteins TFATSY: proteins vs whole strains FASTA: ORFs vs whole strains	Threshold	CLUSTER (single-linkage)	Synteny	10.1073/pnas.0506758102
GET_HOMOLOGS [41] (2013)	BLAST (all-vs-all) *(optional HMMer step)*	Threshold	Bidirectional Best Hit OrthoMCL COGtriangles	*Optional: split families with different domain architectures (based on HMMer)*	10.1128/AEM.02411-13
PanOCT [64] (2012)	BLAST (all-vs-all)	Threshold RBH Synteny	Greedy hierarchical clustering	RBH Synteny	10.1093/nar/gks757
PGAP [208] (2012)	BLAST (all-vs-all)	Threshold	MCL		10.1093/bioinformatics/btr655
EDGAR [18] [19] (2009, 2016)	BLAST (all-vs-all)	Threshold (calculated)	Connected components	core calculated independently	10.1093/nar/gkw255
Roary [137] (2015)	Prefilte: CD-Hit BLAST on centroids	Threshold *Synteny (optional)*	MCL	Synteny	10.1093/bioinformatics/btv421
PIRATE [13] (2019)	Prefilter: CD-Hit BLAST/Diamond on centroids	Threshold	MCL	If gene missing in only 1 genome: DNA search based on synteny	10.1093/gigascience/giz119
MetaPGN [144] (2018)	Prefilter: CD-Hit Blat: all centroids vs reference	Threshold	Connected components	Synteny	10.1093/gigascience/giy121
panX [52] (2018)	DIAMOND for proteins (all-vs-all) BLAST for rRNA		MCL	Phylogenetic tree to split orthologs/paralogs	10.1093/nar/gkx977
Sonic Panaroid [42] (2019)	MMseqs2 (all-vs-all)	Threshold	In paranoid		10.1093/bioinformatics/bty631
PEPPAN [209] (2020)	Prefilter: linclust BLAST/DIAMOND on references	Threshold	rapidNJ		10.1101/gr.260828.120
pandelos [21] (2018)	Prefilter: k-mer collisions Jacquard similarities	RBH	Connected components	Girvan-Newman to split inconsistent families	10.1186/s12859-018-2417-6
Panaroo [187] (2020)	CD-Hit	Threshold on CD-Hit	graphical representation of CD-Hit clusters based on synteny	- split paralogous vs non-paralogous clusters (#members/genome) - collapse back based on graph context	10.1186/s13059-020-02090-4

Table 3.1: Summary of methods used by different pangenome tools for each main step. This is a non-exhaustive list of tools. Each tool name links to the corresponding publication.

4 Conclusion

This introducing chapter draws to a close. Its six-worded title, *large scale comparative genomics of bacterial genomes*, is now much clearer. Bacteria are very diverse micro-organisms, and their phenotypes result from the expression of their genes, strategically distributed along one or several DNA molecule(s). These molecules are far from being fixed: many mobile genetic elements are exchanged, within and between different replicons. These elements can also be transferred across cells. Constantly varying the gene content of bacterial genomes, horizontal gene transfer is the main mechanism responsible for the emergence of novel functions in genomes. Thus, the methods for whole genome sequence alignments done so far to compare genomes had to be revised to take into account this process. This spurred the apparition of the pangenome concept. However, due to the high number of comparisons required, its computation is far from being trivial, and requires a lot of time. As a consequence, the development of the first tools to generate pangenomes was quickly caught up by the very fast increase of number of bacterial genomes sequenced. The first methods, developed for a few tens of genomes, did not scale up: new heuristics had to be developed.

When I arrived in GEM (Microbial Evolutionary Genomics) team, projects with several hundreds of genomes were becoming more and more common, and the methods used at that time started to reach their limits. Most evolution studies require fundamental blocks of data to be performed: annotated genomes, pan and core genomes, and alignments per gene family. Even if a few tools existed to compute pangenomes of quite big datasets, none were able to build all the fundamental blocks of data mentioned above. The need to use different programs, and develop methods to link them engenders several problems when it comes to simplicity, efficacy and/or reproductibility. Among others, it requires the installation of these different softwares, it depends on multiple versions of each software, and the outputs of a tool potentially need modifications to be compliant with the inputs of the next step. Moreover, with the tools existing at that time, computing the pangenome was becoming too slow for the analysis of thousands of genomes (10h on an example of 1000 *E. coli* genomes). Finally, the increasing proportion of drafts in genome datasets also required changes in the methods, to account for the existence of multiple contigs and missing information.

The challenge was set: propose a standardization for the basis of comparative genomics studies, by developing a tool able to do all steps, accurately and in a reasonable amount of time.

This PhD project gave birth to `PanACoTA` (for PANgenome with Annotations, COre identification, Tree and corresponding Alignments), a new tool now available to the comparative genomics community [146]. In the next chapter, you will find the paper corresponding to the publication of this tool. It describes the method, illustrated by an application on a dataset of almost 4000 *Klebsiella pneumoniae* genomes. During the development of `PanACoTA`, large-scale comparative genomics projects did not take a break. We took advantage of these projects to test and improve the method. Chapter III shows different studies on which the application of *PannaCotta* has been fruitful.

Part II

DEVELOPMENT OF PanACoTA

5 PanACoTA: a modular tool for massive microbial comparative genomics

This paper presents the main part of my PhD, consisting in developing a bioinformatics tool. The development of this new software was motivated by the need of a fast but reliable pangenome computation tool. For this, we had several requirements.

First, we wanted to minimize the annotation inconsistencies (problem stated in chapter 3.3). For that, the tool had to be able to handle large datasets, check the quality of the sequences and filter out genomes not respecting the given criteria, and uniformly (with the same software) annotate the remaining sequences. Then, we wanted a pangenome computation method able to scale to many thousands of genomes in a reasonable amount of time (i.e. less than several hours). Dealing with large-scale datasets, the notion of core genome had to be adapted, and the tool had to give the possibility to use different definitions of pangenome. As most of the comparative genomic studies need a phylogenetic tree, the tool had to output the MSA of all core or persistent genome families. Moreover, we wanted the tool to be modular: one should be able to re-run a step to try new parameters, to start at any step with its own data as an input, and to run other softwares directly from the output of this tool. Finally, the aim was to make this tool freely available to the community.

Based on these requirements, I developed `PanACoTA`, which is presented in the following paper.

Published online 12 January 2021 *NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1* **1**
doi: 10.1093/nargab/lqaa106

PanACoTA: a modular tool for massive microbial comparative genomics

Amandine Perrin [1,2,3,*] **and Eduardo P. C. Rocha** [1]

[1]Microbial Evolutionary Genomics, CNRS, UMR3525, Institut Pasteur, 28, rue Dr Roux, Paris 75015, France, [2]Sorbonne Universite, College doctoral, F-75005 Paris, France and [3]Bioinformatics and Biostatistics Hub, Department of Computational Biology, Institut Pasteur, USR 3756 CNRS, 28, rue Dr Roux, Paris 75015, France

Received September 16, 2020; Revised November 10, 2020; Editorial Decision November 29, 2020; Accepted December 01, 2020

ABSTRACT

The study of the gene repertoires of microbial species, their pangenomes, has become a key part of microbial evolution and functional genomics. Yet, the increasing number of genomes available complicates the establishment of the basic building blocks of comparative genomics. Here, we present PanACoTA (https://github.com/gem-pasteur/PanACoTA), a tool that allows to download all genomes of a species, build a database with those passing quality and redundancy controls, uniformly annotate and then build their pangenome, several variants of core genomes, their alignments and a rapid but accurate phylogenetic tree. While many programs building pangenomes have become available in the last few years, we have focused on a modular method, that tackles all the key steps of the process, from download to phylogenetic inference. While all steps are integrated, they can also be run separately and multiple times to allow rapid and extensive exploration of the parameters of interest. PanACoTA is built in Python3, includes a singularity container and features to facilitate its future development. We believe PanACoTa is an interesting addition to the current set of comparative genomics tools, since it will accelerate and standardize the more routine parts of the work, allowing microbial genomicists to more quickly tackle their specific questions.

INTRODUCTION

Low cost of sequencing and the availability of hundreds of thousands of genomes have made comparative genomics a basic toolkit of many microbiologists, geneticists, and evolutionary biologists. Many bacterial species of interest have now over 100 genomes publicly available in the GenBank RefSeq reference database, and a few have more than ten thousand. This trend will increase with the ever decreasing costs of sequencing, the availability of long-read technologies, and the use of whole-genome sequencing in the clinic for diagnostics and epidemiology. As a result, researchers that would like to use available assemblies are faced with extremely large amounts of data to analyze. Comparative genomics has spurred important contributions to the understanding of the organization and evolution of bacterial genomes in the last two decades (1,2). It has become a standard tool for epidemiological studies, where the analysis of the genes common to a set of strains — the core or persistent genome — provides unrivalled precision in tracing the expansion of clones of interest (3,4). The use of routine sequencing in the clinic will further require rapid and reliable analysis tools to query thousands, and soon possibly millions of genomes from a single species (5). Population genetics also benefits from this wealth of data because one can now track in detail the origin and fate of mutations or gene acquisitions to understand what they reveal of adaptive or mutational processes (6). Finally, genome-wide association studies have been recently adapted to bacterial genetics, to account for variants in single nucleotide polymorphism and gene repertoires (7). They hold the promise of helping biologists to identify the genetic basis of phenotypes of interest. Given the high genetic linkage in bacterial genomes, these studies may require extremely large datasets to detect small effects. More specifically, reverse vaccinology is also a noteworthy application of these pangenomics methods, to identify novel potential antigens among core surface-exposed proteins of a given clade (8).

The availability of large genomic datasets puts a heavy burden on researchers, especially those that lack extensive training in bioinformatics, because their analysis implicates the use of automatic processes, efficient tools, extensive standardization and quality control. Many tools have been recently developed to make rapid searches for sequence similarity with excellent recall rates for highly similar sequences (9–11).

Other tools provide methods to rapidly cluster large numbers of sequences in families of sequence similarity, to get the families common to a set of genomes, to align them, or to produce their phylogeny, four cornerstones of compara-

*To whom correspondence should be addressed. ?Tel: +33 1 45 68 89 83; Fax: +33 1 45 68 87 27; Email: amandine.perrin@pasteur.fr

tive genomics. A number of recent programs have recently been published that include some of these tools to compute bacterial pangenomes (for a review, see (12)). Many of these programs compute alignments and clusters of families using programs that are very fast. They use tools that make some compromises between accuracy and speed, such as DIAMOND (9), USEARCH (13) and CD-HIT (14). The latter is used, among others, by Roary (15), which is currently the most popular tool to compute pangenomes, and Panaroo (16), a very recent tool aiming at reducing the impact of erroneous automated annotation of prokaryotic genomes. BPGA (17), using USEARCH or CD-HIT to cluster proteins, also provides some downstream analyses. PanX (18), which has an outstanding graphical interface, uses DIAMOND to search for similarities among genes.

More recently, SonicParanoid introduced the use of the highly efficient and accurate program mmseqs2 to build pangenomes, and PPanGGOLiN used the same tool to provide a method to statistically class pangenome families in terms of their frequency (19–21) . Some recent programs also use graph-based approaches to further refine the pangenomes, such as PPanGGOLiN and Panaroo (16). For that matter, the analysis of a dataset of 319 *Klebsiella pneumoniae* genomes by both tools provided similar results (16). Some tools, such as PIRATE (22) have also been recently developed to cluster orthologues between distant genomes. However, all these programs lack some or all of initial and final steps that are essential in comparative genomics, including download, quality control, alignment and phylogenetic inference. This spurred the development of PanACoTA (PANgenome with Annotations, COre identification, Tree and corresponding Alignments). To take advantage of the vast amount of genomic information publicly available, one needs six major blocks of operations. (i) Gather a set of genomes of a clade automatically. This requires some quality control, to avoid drafts with an excessive number of contigs. It is also often convenient to check that the genomes are not too redundant, to minimize computational cost and biases due to pseudo-replication. On the other side, it is important to check that genomes are neither too unrelated, to eliminate genomes that were misclassified in terms of bacterial species (or the taxonomic organization of relevance). (ii) Define *a priori* an uniform nomenclature and annotation, without which the calculation of pangenomes and core genomes becomes unreliable for large datasets. (iii) Produce the pangenome, a matrix with the patterns of presence/absence of each gene family in the set of genomes, using an accurate, simple and fast method. (iv) Use the pangenome to identify sets of core or persistent genes. (v) Produce multiple alignments of the gene families of the core or persistent genomes. (vi) Finally, produce quickly a reasonably accurate phylogeny of the set of core/persistent genes. These four collections of data, pangenome, core genome, alignments and phylogenetic tree, are the basis of most microbial comparative genomics studies. At the end of this process, the researcher can produce more detailed analyses, specific to the questions of interest, which often lead to changes such as including/excluding taxa, changing the thresholds of sequence similarity, increasing alignment accuracy, or rebuilding phylogenies using different methods. Such re-definitions can be achieved more efficiently when pipelines are modular and allow to restart the analyses at several key points in the process.

Considering the current availability of pipelines for microbial comparative genomics, we have built one that is modular, easy to setup, uses state-of-the-art tools and allows simple re-use of intermediate results. The goal was to provide a pipeline that allows to download all genomes from a taxonomic group and make all basic comparative genomics work automatically. The pipeline is entirely built in a single language, Python v3, and uses modern methods to facilitate its future maintenance and to limit unwanted behavior. PanACoTA is freely available under the open source GNU AGPL license. Here, we describe the method and illustrate it with an analysis of two datasets of 225 complete and 3980 complete or draft genomes of *K. pneumoniae*. This species is interesting for our purposes because there are many genomes available and it has a very open pangenome (23). The first dataset describes a situation where sequence quality is usually high, and the second illustrates how the method scales-up to a very large dataset where some sequences and assemblies are of lower quality. The procedure is detailed in the Materials and Methods section, whereas the illustration of its use, and how it changes in relation to key options in the two datasets, is detailed in the Results section.

MATERIALS AND METHODS

PanACoTa is implemented in six independent sequential modules, described in the sections below. This allows to start or stop at any step and re-run an analysis with other parameters (see overview in Figure 1 and key parameters in Table 1). It also provides a module `all`, which allows to run all modules in a single-command.

Datasets

The first module `prepare` fetches the compressed non-annotated fasta files assemblies from the NCBI matching a given taxonomy ID using the scripts from `ncbi_genome_download` library (https://github.com/kblin/ncbi-genome-download).

We use two datasets of *K. pneumoniae* genomes to illustrate how PanACoTA functions. DTS1 contains all complete and draft assemblies from the NCBI refseq database on 10 October 2018. DTS2 is the subset of DTS1 containing only the complete genomes (genomes with `assembly_level = Complete Genome`, based on the NCBI summary file).

Quality control procedure

PanACoTA removes assemblies that do not conform with basic requirements in terms of assembly and taxonomy. This is done by the `prepare` module after downloading the genomes, or by the `annotate` module before the annotation step (if the user did not use the `prepare` module).

The first control procedure filters genomes in terms of sequence quality. Since there is usually no standard description of the quality of the sequence assembly in RefSeq genomes, the program infers it from the sequences. First, it

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **3**

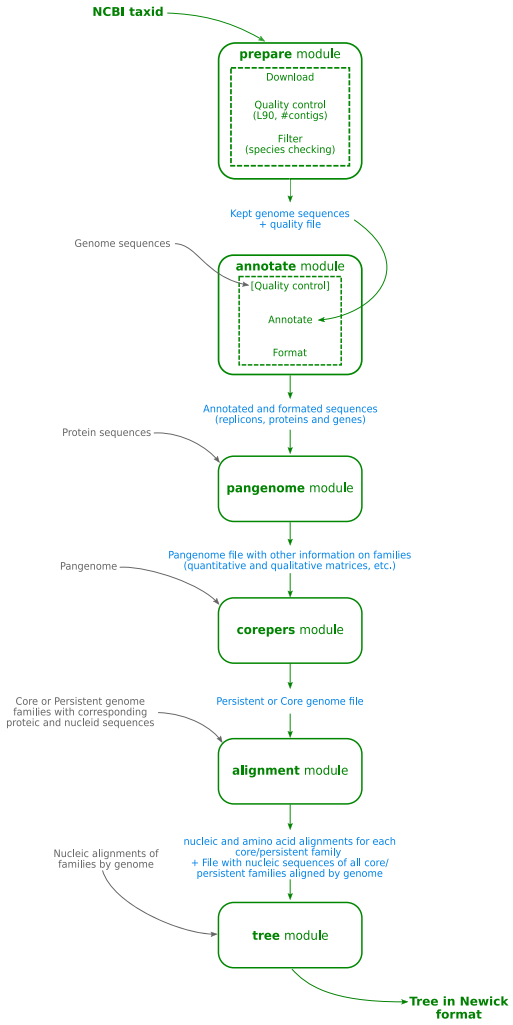


Figure 1. Overview of PanACoTA method.

is common usage to put stretches of 'N' to separate contigs in a same fasta sequence. Hence, PanACoTA splits sequences at each stretch of at least a given number of 'N' to get one fasta entry per contig. Assuming that the user is analyzing genomes from the same species, those genomes should have relatively similar characteristics in terms of number of contigs and length. Hence, PanACoTA calculates the total number of contigs, and the L90 (the minimum number of contigs necessary to get at least 90% of the whole genome). Very high values of these two variables are usually an indication of low quality of sequencing or assembling, resulting in genome exclusion.

The second procedure filters redundant and misclassified genomes. This is done based on the genetic distance between pairs of genomes, as calculated by Mash (24), which can be computed very fast and is accurate for closely related genomes. Mash reduces each genome sequence to a sketch of representative k-mers, using the MinHash technique (25). It then compares those sketches, instead of the full sequences. The Mash distance `D` strongly correlates with alignment-based measures such as the Average Nucleotide Identity (ANI) based on whole-genome sequence comparisons using the blast algorithm (26): $D \approx 1 - ANI$. For ANI in the range of 90–100%, the correlation with Mash distance is even higher when increasing the sketch size. Since pangenomes are typically computed for a single bacterial species, we are here using Mash to discriminate genomes having at least 94% identity. A few recent programs have been published showing slightly more accuracy than Mash, but we found them too slow for the use as a systematic filter when performing millions of pairwise genome comparisons. For example, using 15 cores, FastANI (27) requires around 1h15 to compare all pairs of 200 genomes (40 000 pairwise comparisons), where Mash with a sketch size of 10^6 does the task in less than 3 min. The program dRep (28) uses Mash as a pre-filter and then makes more accurate and time-consuming analyses. This is very useful when comparing draft genomes of very different sizes, like metagenomic assembled genomes, but less so for the analysis of within-species complete genomes. Users requiring a finer grade study of ANI may wish to post-analyze their genomes using these programs.

Bacterial species are usually defined as groups of genomes at more than 94% identity (29), which sets the default threshold for D (`max_mash_dist` = 0.06). On the other extreme, genomes with very high similarity (low Mash distances) provide very similar information. Their exclusion decreases the time required for the analysis and diminishes over-sampling of certain clades. PanACoTA sets `min_mash_dist` to 10^{-4} by default. This represents one point change every 10 genes, which may be close to the sequencing and assembling accuracy of many draft genomes.

The two procedures, quality control and Mash filtering, are linked together. The information on the number of contigs and L90 is useful to chose the genome that is kept between a pair of very similar genomes. In summary, the control procedure works as follows:

1. Genomes with an excessively high number of contigs or L90 are excluded.
2. Genomes are primarily sorted by increasing L90 value, and secondarily by increasing number of contigs to produce a list ordered in terms of quality.
3. The genomes are compared with Mash. For that, the first genome of the ordered list (the one with best quality) is compared to all the others. The ones which do not obey to the distance thresholds are discarded. The procedure then passes to the subsequent genome in the ordered list (if not rejected before), compares it to all remaining genomes, and discards those not respecting the thresholds. The process continues until the ordered list is exhausted.

The output of the `prepare` module is a database with the genomes that passed the two steps of the quality control procedure: 3980 genomes for DTS1 and 225 complete genomes for DTS2 (accession numbers in Supplementary

Table 1. Key parameters for each module of PanACoTA

Module	Key parameters	Short description	Default values
prepare	NCBI species taxid		If user wants to download
	NCBI species		Genomes from NCBI
	- -cutn n	Split contig when there are at least 'n' N in a row	5
	- -l90 x	Discard genome(s) with L90 higher than x	100
	- -nbcont x	Discard genome(s) with more than x contigs	999
	- -min_dist x	Discard genome(s) closer than a Mash distance of x	10^{-4}
	- -max_dist x	Discard genome(s) with a Mash distance higher than x	0.06
annotate	- -l90 x	Discard genome(s) with L90 higher than x	100
	- -nbcont x	Discard genome(s) with more than x contigs	999
	- -prodigal	Use only prodigal instead of Prokka	False
pangenome	-i x	Minimum sequence identity to be considered in the same family	0.8
	-c x	Clustering mode (0 for 'set cover', 1 for 'single-linkage', 2 for 'CD-Hit')	1
corepers	-t tol	Min % of genomes having at least 1 member in a family to consider the family as persistent	1 (core-genome)
	-M	'Multiple persistent genome'	False
	-X	'Mixed persistent genome'	False
align	-c file	File containing core genome	
tree	-s software	Software to infer phylogeny	IQtree

Table S1). PanACoTA also provides a file listing the discarded genomes and why they were discarded.

Annotation

The `annotate` module provides uniform gene annotation. It takes as input a database of fasta sequences, from the `prepare` module or provided by the user. If no information is given on the quality control of those genomes (number of contigs and L90), this quality control is done here (see previous section for more information on the quality control step).

PanACoTA annotates all genomes with Prokka (30). The latter uses Prodigal (31) to identify gene positions. It then adds functional annotations using a series of programs, including BLAST+ (32) to search for homologs in a database of proteins taken from Uniprot and HMMER3 (33) to search for proteins hitting selected profiles from TIGRFAM (34) and PFAM (35). All annotated sequences are renamed using a standard sequence header format. The header of each gene contains 20 characters and provides human readable information on the genome and contig of the gene, its relative position in the genome and if it is at the border of a contig (see Figure 2).

If the user does not need the functional annotation, the module gives the possibility of running only the gene finding part, i.e. only running Prodigal. For very large datasets it is much faster to use this option and annotate a posteriori only one gene per family of the pangenome using Prokka or more complete annotation systems like InterProScan (36). The output of this step consists in five files per genome: the original sequence, the genes, the proteins (all in fasta format), a gff file containing all annotations and a summary information file.

Identification of the pangenome

The `pangenome` module of PanACoTA computes the set of all protein families in the genomes (on the 'Proteins' folder generated by the `annotate` module).

The inference of the pangenome involves comparisons between all pairs of proteins, i.e. its complexity is to the square of the number of genes (and thus of genomes). To generate a reliable pangenome in a reasonable time, PanACoTA calls the MMseqs2 suite (20). The `mmseqs search` module has a very good speed/sensitivity trade-off. In order to reduce time, it uses three consecutive search stages, with increasing sensitivity and decreasing speed. Everything is highly parallelized and optimized on multiple levels. The first step filters up to 99.9% of the sequences by eliminating high dissimilarities, i.e. sequences not having at least two consecutive kmer matches. The second step filters out another 99% of the remaining sequences using an ungapped alignment. This leaves a small amount of sequences to process with an optimized version of the Smith–Waterman alignment, where only scores are calculated, and not the full alignments.

We used the `mmseqs cluster` module included in MMseqs2 suite, with the default `Cascaded clustering` option. This module works in two main steps. It first clusters proteins using linclust (37), a linear time protein sequence clustering algorithm as a prefilter. Then, the representative sequences of this first step are handled by the `mmseqs search` module and clustered. This second step is repeated three times, each time with a higher sensitivity at the `mmseqs search` algorithm module.

PanACoTA uses the `Connected component` mode for clustering, because it has provided results consistent with our previous methods. This mode uses transitive connections to merge pairs of homologous genes. Alternatively, two other clustering modes (`Greedy Set cover`, or `Greedy incremental`) are available in the `pangenome` module. Importantly, the tuning of the options of mmseqs2 allows the sequence similarity analyses to be exceedingly fast or extremely sensitive (20). In PanACoTA the user can change the key parameters `--min-seq-id` and `--cluster-mode`, and re-run the `mmseqs cluster` module to explore their effect on the results. More specific mm-

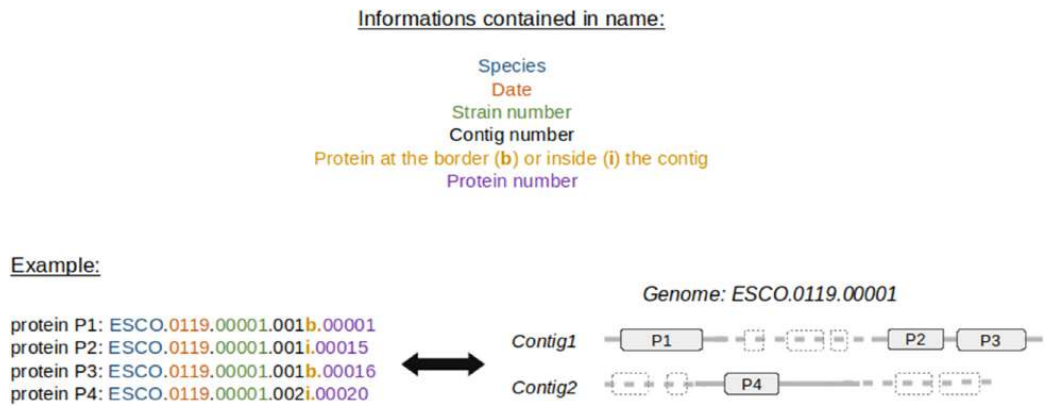


Figure 2. Description of the standard output header format for proteins annotated by PanACoTA.

seqs2 parameters have, for the time being, to be used with the standalone version of the program.

This step outputs files containing one line per family of the pangenome and indicating the gene identifiers, the presence of the gene family (binary matrix), or the number of elements. The latter can be used as input for TreeWAS (38).

Panacota does not take into account synteny between genes in the genomes, which has limited interest in draft genomes. Several programs can do such analyses, e.g. panOCT (39,40), SynerClust (41) or PANINI (42).

Identification of core and persistent genomes

The classification of gene families present in a large number of taxa is done by the `corepers` module using a file generated by the `pangenome` module. In early studies, the pangenome matrix was used to identify the gene families present in all genomes in a single copy: the core genome. However, the increase of the number of genomes in the dataset tends to decrease drastically the size of the core genome. This is because sequencing or annotation errors as well as rare deleterious polymorphism in the populations lead to the rapid decrease of the number of core genes with the increase in the number of input genomes. To overcome this problem, one commonly identifies the persistent genome, which is more robust to rare (true or artifactual) variants. PanACoTA defines three types of persistent genomes (see Figure 3):

1. Strict-persistent: a family that contains exactly one member in at least `N`% genomes ($N = 100$ means it is a core-family). This definition is particularly practical to reconstruct phylogenies without having to handle the existence of multiple copies per genome.
2. Mixed-persistent: a family where at least `N`% of the genomes have exactly one member, and other genomes have either zero, either several members in the family. This definition is intermediate between the other two, i.e. it includes the strict-persistent and is included by the multi-persistent.
3. Multi-persistent: a family with at least one member in `N`% of the genomes. This definition is interesting to analyze patterns of diversification of nearly ubiquitous protein families.

The module `corepers` uses the pangenome instead of a reference genome (whose choice can be questionable). Re-running the module is very fast, because it only requires the re-analysis of the pangenome matrix and can be done multiple times with different parameters.

The output of this module is a file containing the persistent families of proteins.

If the user wants to identify the persistent genome using a statistical approach rather than using fixed thresholds, the gff file generated by `annotate` module is compatible with PPanGGOLiN (21). This software generates the multi-persistent version of the persistent genome (multigenic families are allowed).

Multiple alignments of the persistent gene families

The alignment of the persistent gene families is done by the `align` module using the persistent genome coming from the `corepers` module, or independently provided by the user. When using the strict-persistent genome, all genes are aligned. When using the other definitions of persistent genomes, some genomes can lack a gene or have it in multiple copies and must be handled before phylogenetic inference. When a genome lacks a member or has more than one member (mixed or multi persistent) of a given gene family, PanACoTA adds a stretch of gaps ('-') of the same length as the other aligned genes. Adding a few '-' has little impact on phylogeny reconstruction. For example, it has been showed that adding up to 60% of missing data in the alignment matrix could still result in informative alignments (43). In our experience, when this approach is applied to within-species persistent genomes, it usually incorporates $<1\%$ of gaps. The effect of missing data should thus be negligible relative to the advantage of using the phylogenetic signal from many more genes (i.e. in contrast to using the strict-persistent genome). Alignments are more accurate when done at the

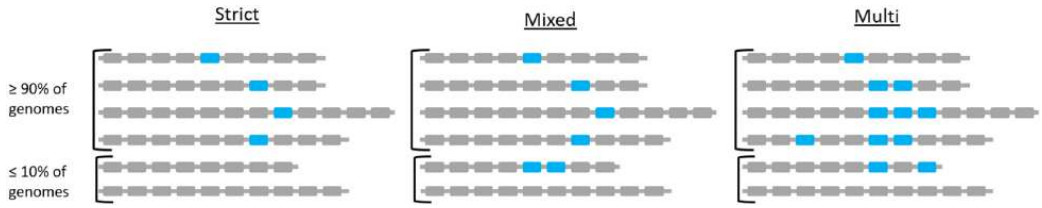


Figure 3. Different types of persistent genomes proposed by PanACoTA, with a threashold of $N = 90\%$.

level of the protein sequence. This has the additional advantage of producing codon-based nucleotide alignments that can be used to study selection pressure on coding sequences. Hence, PanACoTA translates sequences, aligns the corresponding proteins and then back-translates them to DNA to get a nucleotide alignment. This last step constitutes in the replacement of each amino acid by the original codon. Hence, at the end of the process, the aligned sequences are identical to the original sequences.

PanACoTA does multiple sequence alignment using MAFFT (10) as it is often benchmarked as one of the most accurate multiple alignment programs available and one of the fastest (44). It has options that allow to make much faster alignments, at the cost of some accuracy, to handle very large datasets. This loss of accuracy is usually low for very similar sequences as it is the case of orthologous gene families within species, and means that PanACoTA can very rapidly align the persistent genome.

This module returns several output files: the concatenate of the alignments of all families to be used for tree inference, and, for each core/persistent genome family, a file with its gene and protein sequences aligned.

Tree reconstruction

The phylogenetic inference is done with the `tree` module of PanACoTA. It uses as input the alignments of the `align` module or any other alignments in Fasta format.

This is the part that takes most time in the entire pipeline, because the time required for phylogenetic inference grows very fast with the size of the dataset. Even efficient implementations of the maximum likelihood analyses scale with the product of the number of sites and the number of taxa, which is a problem in the case of large datasets (thousands of taxa, with more than ten thousands sites for each one). PanACoTA proposes several different methods to obtain a phylogeny: IQ-TREE (45), FastTreeME (46), fastME (47) and Quicktree (48). According to its needs, the user can choose one of these methods to infer its phylogenetic tree. These trees can be used to build more rigorous phylogenetic inference using methods that are more demanding in computational resources, e.g. by changing the options of IQ-TREE. Whatever the software used, the `tree` module takes as input a nucleotide alignment in Fasta format (like, for example, the output of `align` module), and returns at least a tree in Newick format. According to the software and options used, other output files may be generated, like bootstrap trees for example. IQ-TREE also returns the BIONJ tree from which it started tree search, as well as the pairwise distance matrix corresponding to the output tree. Recombination is known to affect phylogenetic reconstruction (49,50). To tackle this problem, some researchers detect and then remove recombination tracts from genomes before inferring the phylogeny. This can be done outside PanACoTA by modifying the multiple alignments before proceeding to the phylogenetic inference. We have not implemented in PanACoTA the detection or exclusion of recombination tracts. Several studies have shown that removing the identifiable recombination tracts tends to distort phylogenetic inference at a larger extent than simply using all the information in the multiple alignments (51,52). This is probably because available methods miss many events of homologous recombination, leading to biases in phylogenetic inference. When relevant, one can use methods that simultaneously infer recombination and phylogenetic history, altough these tend to be computationally costly.

Implementation and availability

`PanACoTA` was developed in Python3, trying to follow the best practices for scientific software development (53,54). For that, the software is versioned using git, allowing the tracking of all changes in source code during PanACoTA's development. It is freely distributed under the open-source AGPL v3 licence (making it usable by many organizations) and can be downloaded from https://github.com/gem-pasteur/PanACoTA. The software can be installed directly from the git repository, or using pip or conda package-management systems. A singularity image, including all needed dependencies, is also hosted via Docker Hub. By downloading this image, the user can run PanACoTA without installing anything. This is of particular use for running on clusters, where there is usually no root access.

Hosting `PanACoTA` on GitHub allows for issue tracking, i.e. users can report bugs, make suggestions or, for developers, participate to the software improvement. To provide a maintainable and reliable software, we set up continuous integration process: each time a modification is pushed, there is an automatic software installation checking, unit tests are done, and, if necessary, an updated version of the documentation is generated, as well as an update of the docker image on Docker Hub (which can be used as a singularity image as described previously).

As introduced just before, we also provide a complete documentation, including a step by step tutorial, based on provided genome examples, so that the user can quickly get started. It also contains more detailed sections on each

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 7

Table 2. Summary of execution times by (sub)module

MODULE	STEP	DTS1 (3980 genomes)	DTS2 (225 genomes)
prepare	*Downloading*	1 h (5805 genomes)	3 min (266 genomes)
	Quality control	<4 min	∼15 s
	Filter	20 min	∼1 min
annotate	*With Prokka*	5 days	10 h
	With Prodigal	6 h	30 min
pangenome		30 min	1 min
corepers (1 CPU)		1 min	5 s
align	*Strict persistent*	3 h	10 min
	Mixed-persistent	7 h	11 min
tree (IQ-TREE2) (28 CPUs)	*Strict-persistent*	7 h (40 GB RAM)	3 min 10
	Mixed-persistent	24 h (90 GB RAM)	3 min 30

module, aiming at helping users to tune all parameters, in order to adapt the run to more specific needs. This documentation also includes a 'developer' section, addressed to developers wanting to participate in the project.

During its execution, PanACoTA provides logging information, so that user can see real-time execution progress (a `quiet` parameter is also proposed for users needing empty stdout and stderr). This also provides log file(s) to keep track on what was ran (command-line used, time stamp, parameters used etc.).

RESULTS AND DISCUSSION

All execution times mentioned in this section correspond to wall clock time on eight CPUs (except when the number of CPUs is given). A summary of all execution times can be found in Table 2.

Download and preparation of genome sequences

The first module of PanACoTA was used to download all genomes of *K. pneumoniae* using the TaxID 573. It took ∼1 h to download the 5805 *K. pneumoniae* genome sequences (including 266 complete genomes). We used the module `annotate` to make the quality control (L90 < 100 and number of contigs < 999), which took less than 4 min. This step discarded 233 draft genomes, leaving 5572 for further analysis (see Figure 4). When the threshold on the number of contigs was decreased by half (number of contigs < 500), only 52 more genomes were removed (see Figure 4B). To define the best thresholds to the analysis, the user can preview its dataset quality with a 'dry-run' of the `annotate` module. Then, the user can launch the real analysis, from `prepare` or `annotate` with the adapted thresholds.

We removed the very distantly related and redundant genomes using Mash (K-mer size of 21 (default), and sketches of at most 10 000 non-redundant min-hashed *k*-mers). A total of 1592 genomes (including 41 complete genomes) did not respect the distance thresholds (`max_mash_dist` = 0.06 and `min_mash_dist` = $1e^{-4}$). Most (1448) were too similar to other genomes, whereas 144 were too distantly related with the *K. pneumoniae* genomes (Figure 5).

Expert analysis can lead to the definition of narrower ANI values. For example, Kleborate (https://github.com/katholt/Kleborate) (55) defines strong *K. pneumoniae* matches for distances ≤ 0.01 and weak matches between

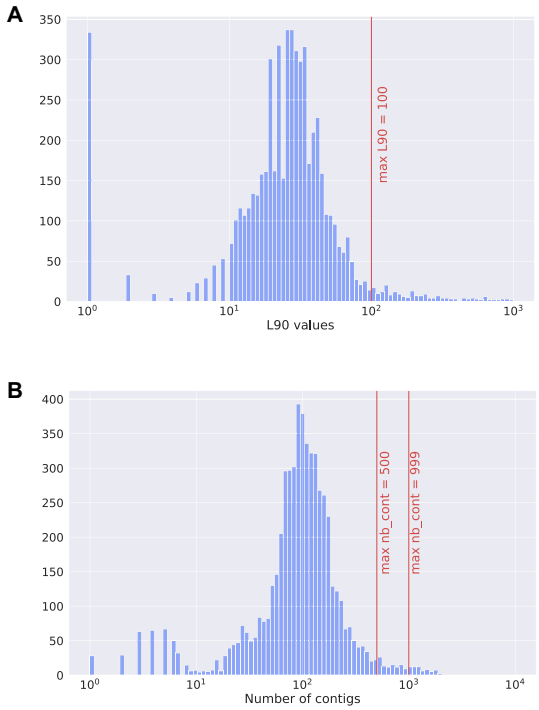


Figure 4. Histograms describing the features of the 5805 *Klebsiella pneumoniae* genomes downloaded from Refseq. (**A**) Distribution of L90 values. (**B**) Distribution of the number of contigs per genome.

0.01 and 0.03. In our dataset, Kleborate would have only removed 22 additional genomes, that it identifies as *K. quasipneumoniae subspecies similipneumoniae*. The default method of Panacota, which is designed for any species, is thus consistent with Kleborate results regarding the specific case of *K. pneumoniae* genomes when starting from the NCBI taxonomy ID.

Three genomes showed an ANI <84% identity, meaning they may not even be from the same genus, which emphasizes the necessity of this kind of analysis before computing a pangenome. They were removed from the analysis (GCF_900451665.1, GCF_900493335.1 and

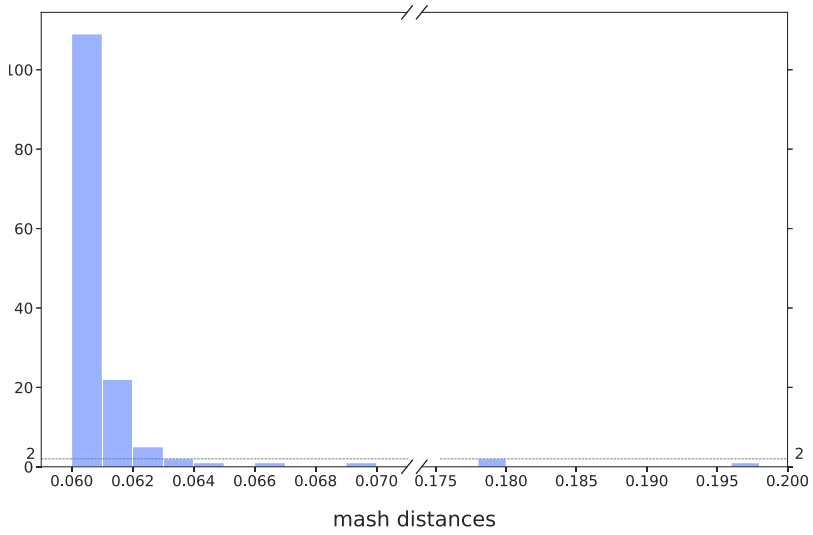


Figure 5. Distribution of Mash distances for the 5572 genomes respecting the L90 and number of contigs thresholds, but having a Mash distance higher than the threshold (0.06).

GCF_900493505.1). A neighbor-joining tree generated from Mash distance matrix with scikit-bio (https://github.com/biocore/scikit-bio) confirmed the gap between those three genomes and the others (see Supplementary Figure S1, where genomes kept in DTS1 are in green, while those discarded are in red).

Finally, these filters left 3980 genomes in the analysis, with an average of 5307 genes per genome, which will be called the reference database DTS1. Among them, there are 225 complete genomes that form the dataset DTS2 (see Figure 6).

We then proceeded to the functional annotation, which is by far the slowest of the first tasks. The annotation of the genomes with Prokka 1.11 took ∼1 min 50 s per genome, i.e. around 5 days for the whole dataset. For comparison, the annotation using only prodigal 2.60 took less than 6 h (annotation + formatting of all 3980 genomes), i.e. 6 s per genome. Assuming that genes from the same pangenome family have similar functions, one can annotate one protein per family at the end of the process and save considerable time.

Building pangenomes

The 3980 DTS1 genomes contain 20 765 062 proteins. It took less than 30 minutes to create the protein database in the MMseqs2 format (Release 11-e1a1c), cluster them (with at least 80% identity and 80% coverage of query and target), and retrieve the pangenome matrices. The DTS1 pangenome has 86607 families. Among them, 35 348 (40%) are singletons (found in a single genome), which is concordant with values observed in *Escherichia coli* (56). The pangenome of DTS2, 1 190 485 proteins, was computed in <1 min. It contains 24 473 families, including 8975 (37%) singletons.

The comparison of these two pangenomes is interesting because it reveals the robustness of the method to changes in sampling size, as summarized in Figure 7. A total of 2147 families contain only members present in both DTS1 and DTS2. Among these, 2122 families are exactly the same in both pangenomes, whereas only 25 were split in the DTS1 pangenome family relative to the DTS2 pangenome. In most of the latter, they are split in two different families of DTS1. This shows that the clustering procedure is quite robust to the addition of a very large number of genomes.

Most important, 22 744 families (that is more than 92% of all DTS2 families) are identical in DTS1 and DTS2 pangenomes. Identical here means that the DTS2 pangenome gene family is included in a DTS1 pangenome gene family, and the other members of this DTS1 pangenome family are only members of genomes not present in DTS2. Furthermore, around half of the remaining families from the DTS2 pangenome are included in a DTS1 pangenome gene family, which contains a few other proteins from DTS2 genomes. Finally, only 187 gene families of the DTS2 pangenome were split into two or three different families of DTS1 pangenome. In other words, 24 286 families (more than 99%) of DTS2 pangenome are subsets of DTS1 gene families. In conclusion, the construction of pangenome families is robust to large variations in the number of input genomes (see Figure 7).

Core and persistent genomes

This part of the analysis is very fast. Using only one CPU, it took around 1 min to generate a core or persistent genome from DTS1 pangenome. PanACoTA provides a core genome and three different measures of persistent genome (see Figure 3). The strict-persistent genome corresponds to cases when the family is present in a single copy in

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **9**

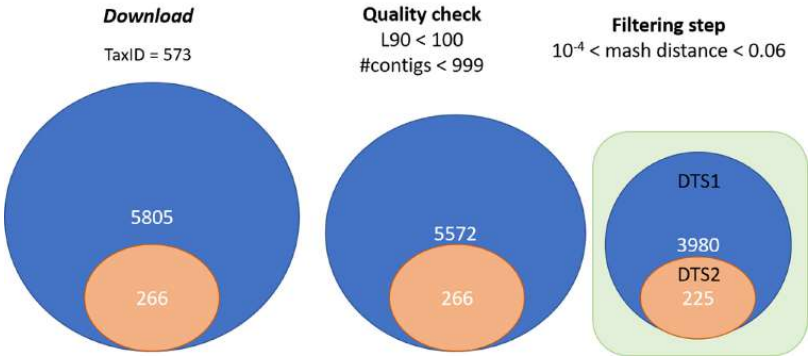


Figure 6. Summary of the procedure to construct DTS1 and DTS2.

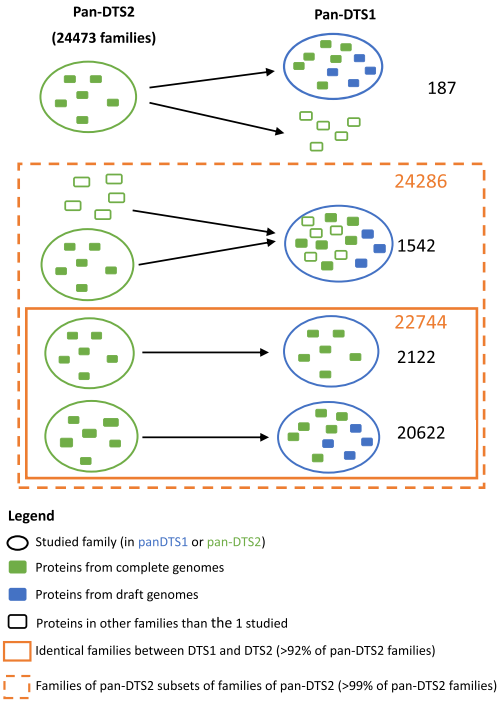


Figure 7. Comparison of the pangenomes generated by PanACoTA for both DTS1 and DTS2.

99% genomes and absent from the others. In DTS2, the set of complete genomes, the difference between the core and strict-persistent genome is appreciable (2238 versus 3295 families), i.e. the persistent genome is 50% larger (see Figure 8). The difference becomes huge when the analysis is done on the much larger (and less accurate) DTS1 dataset, where the two datasets vary by more than one order of magnitude (79 versus 1418 families). In such large datasets of draft genomes the core genome is not biologically meaningful.

The mixed-persistent genome includes the families present in a single copy in 99% genomes and present (potentially in several copies) or absent from the others. It includes the strict-persistent genome. Its size is close to the latter in the small DTS2 dataset, but much larger in DTS1 (see Figure 8). While the mixed-persistent genome is 65% percent of the average genome in DTS1, the strict-persistent is only 27% percent in the same dataset. This shows the relevance of using definitions of the core genome adapted to the dataset in order to build robust phylogenetic trees or to analyze patterns of genetic diversification and natural selection.

Finally, PanACoTA also computes a multi-persistent genome that includes all gene families present in at least 99% of the genomes, independently of their copy number (see Figure 8). Its analysis reveals many genes encoding regulators, transporters and enzymes that are nearly ubiquitous, but often present in multiple copies. As a rule, this definition is interesting to study gene families present in most genomes, but present in very different copy number. On the other hand, it is typically not very useful for phylogenetic inference.

Phylogenetic tree inference

PanACoTA ran mafft v.7.467 using `–auto` option to align all families. For DTS1, it selected the FFT-NS-2 method, while for DTS2, it selected FFT-NS-i method. This was done with both the strict-persistent (1418 families, 3 h) and the mixed-persistent (3441 families, 7h).

PanACoTA used the multiple alignments as input to IQ-TREE multicore version 2.0.6, with the `-fast` option. For the tree based on the alignment of the strict-persistent (1 438 179 positions), it took around 7 h on 28 CPUs and required 38 GB of RAM. For the tree based on the alignment of the mixed-persistent (3 393 006 positions), it took 24 h using 28 CPUs and required 88 GB of RAM.

We wished to understand the differences in phylogenetic inference in terms of the method used to define the persistent genome (strict and mixed persistent). We computed the patristic distance matrix for each tree and a Pearson correlation test showed that they are strongly correlated

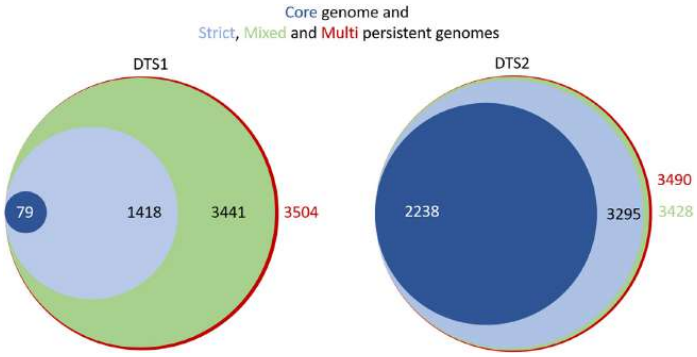


Figure 8. Comparison of the sizes of the core genome and the 3 different types of persistent genomes, for both DTS1 and DTS2. Areas of circles are proportional to the size of the dataset.

(cor = 0.99138, $P < 2.2e^{-16}$). This shows that the distances provided by the two methods are very similar. Hence, if the strict persistent is large enough to generate a phylogenetic tree, it provides adequate distances between genomes. Aligning all mixed persistent families would just take much more time, for a similar result. However, if one is interested in having a robust tree topology, one should use the larger (and computationally costlier) dataset. Indeed, the analyses of Robinson–Foulds distance with R `phangorn` package shows a branch-weighted distance of 0.43 and an absolute distance of 2892 (57). This is because some lineages of *K. pneumoniae* account for a large fraction of the data and these parts of the tree require long informative multiple alignments to produce accurate topologies. Accordingly, the differences in topology between the trees using the DTS2 dataset, which have much larger average branch lengths, show much smaller values of topological distances between the two datasets of persistent genome (RF = 78, wRF = 0.027).

CONCLUSION

PanACoTA is a pipeline for those wanting to test hypotheses or explore genomic patterns using large scale comparative genomics. We hope that it will be particularly useful for those wishing to use a rapid, accurate and standardized procedure to obtain the basic building blocks of typical analyses of genetic variation at the species level. We built the pipeline having modularity in mind, so that users can produce multiple variants of the analyses at each stage. We also paid particularly care with the portability and evolvability of the software. These two characteristics, modularity and evolvability, will facilitate the implementation of novel procedures in the future.

DATA AVAILIBILITY

The two datasets of *K. pneumoniae* genomes used to illustrate PanACoTA were downloaded from the NCBI refseq. Their accession numbers are indicated in Supplementary Table S1. PanACoTA source code is freely available from https://github.com/gem-pasteur/PanACoTA under AGPLv3 license. More information in the last part of Materials and Methods section.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Marie TOUCHON, Matthieu HAUDIQUET and Rémi DENISE for comments, suggestions and bug reports, Yoann DUFRESNE for his help on optimizing some parts of the code, Blaise LI for his help with Singularity, and Bertrand NERON for help on pypi and conda packages.

FUNDING

Agence Nationale de la Recherche Salmo_Prophages [ANR-16-CE16-0029, in part]; Agence Nationale de la Recherche Inception program [PIA/ANR-16-CONV-0005, in part]; Equipe Federation pour la recherche médicale [EQU201903007835, in part].

Conflict of interest statement. None declared.

REFERENCES

1. Vernikos,G., Medini,D., Riley,D.R. and Tettelin,H. (2015) Ten years of pan-genome analyses. *Curr. Opin. Biotech.*, **23**, 148–154.
2. Tettelin,H. and Medini,D. (2020) In: *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, Springer International Publishing.
3. Larsen,M.V., Cosentino,S., Rasmussen,S., Friis,C., Hasman,H., Marvig,R.L., Jelsbak,L., Sicheritz-Pontén,T., Ussery,D.W., Aarestrup,F.M. *et al.* (2012) Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.*, **50**, 1355–1361.
4. Baker,S., Thomson,N., Weill,F.X. and Holt,K.E. (2018) Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science*, **360**, 733–738.
5. Treangen,T.J. and Pop,M. (2018) You can't always sequence your way out of a tight spot. *EMBO Rep.*, **19**, e47036.
6. Sheppard,S.K., Guttman,D.S. and Fitzgerald,J.R. (2018) Population genomics of bacterial host adaptation. *Nat. Rev. Genet.*, **19**, 549–565.
7. Falush,D. (2016) Bacterial genomics: microbial GWAS coming of age. *Nat. Microbiol.*, **1**, 16059.
8. Zeng,L.B., Wang,D., Hu,N.Y., Zhu,Q., Chen,K., Dong,K., Zhang,Y., Yao,Y.F., Guo,X.K., Chang,Y.F. *et al.* (2017) A novel pan-genome

reverse vaccinology approach employing a negative-selection strategy for screening surface-exposed antigens against leptospirosis. *Front. Microbiol.*, **8**, 396.
9. Buchfink,B., Xie,C. and Huson,D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
10. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
11. Bradley,P., den Bakker,H.C., Rocha,E.P., McVean,G. and Iqbal,Z. (2019) Ultrafast search of all deposited bacterial and viral genomic data. *Nat. Biotechnol.*, **37**, 152–159.
12. Kim,Y., Gu,C., Kim,H.U. and Lee,S.Y. (2020) Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotech.*, **63**, 54–62.
13. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
14. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
15. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
16. Tonkin-Hill,G., MacAlasdair,N., Ruis,C., Weimann,A., Horesh,G., Lees,J.A., Gladstone,R.A., Lo,S., Beaudoin,C., Floto,R.A. *et al.* (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.*, **21**, 180.
17. Chaudhari,N.M., Gupta,V.K. and Dutta,C. (2016) BPGA-an ultra-fast pan-genome analysis pipeline. *Sci Rep.-UK*, **6**, doi:10.1038/srep24373.
18. Ding,W., Baumdicker,F. and Neher,R.A. (2018) panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**, e5.
19. Cosentino,S. and Iwasaki,W. (2019) SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics*, **35**, 149–151.
20. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
21. Gautreau,G., Bazin,A., Gachet,M., Planel,R., Burlot,L., Dubois,M., Perrin,A., Médigue,C., Calteau,A., Cruveiller,S. *et al.* (2020) PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.*, **16**, e1007732.
22. Bayliss,S.C., Thorpe,H.A., Coyle,N.M., Sheppard,S.K. and Feil,E.J. (2019) PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience*, **8**, giz119.
23. Holt,K.E., Wertheim,H., Zadoks,R.N., Baker,S., Whitehouse,C.A., Dance,D., Jenney,A., Connor,T.R., Hsu,L.Y., Severin,J. *et al.* (2015) Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E3574–E3581.
24. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
25. Broder,A.Z. (1997) On the resemblance and containment of documents. In: *Procedings. International Conference on Compression and Complexity of Sequences*. IEEE Computer Society, Positano, Salerno, pp. 21–29.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Jain,C., Rodriguez-R,L.M., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, doi:10.1038/s41467-018-07641-9.
28. Olm,M.R., Brown,C.T., Brooks,B. and Banfield,J.F. (2017) DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.*, **11**, 2864–2868.
29. Konstantinidis,K.T. and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 2567–2572.
30. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
31. Hyatt,D., Chen,G.L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
32. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
33. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
34. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
35. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
36. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
37. Steinegger,M. and Söding,J. (2018) Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.
38. Collins,C. and Didelot,X. (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.*, **14**, e1005958.
39. Fouts,D.E., Brinkac,L., Beck,E., Inman,J. and Sutton,G. (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic Acids Res.*, **40**, e172.
40. Inman,J.M., Sutton,G.G., Beck,E., Brinkac,L.M., Clarke,T.H. and Fouts,D.E. (2019) Large-scale comparative analysis of microbial pan-genomes using PanOCT. *Bioinformatics*, **35**, 1049–1050.
41. Georgescu,C.H., Manson,A.L., Griggs,A.D., Desjardins,C.A., Pironti,A., Wapinski,I., Abeel,T., Haas,B.J. and Earl,A.M. (2018) SynerClust: a highly scalable, synteny-aware orthologue clustering tool. *Microbial Genomics*, **4**, e000231.
42. Abudahab,K., Prada,J.M., Yang,Z., Bentley,S.D., Croucher,N.J., Corander,J. and Aanensen,D.M. (2019) PANINI: pangenome neighbour identification for bacterial populations. *Microbial Genomics*, **5**, e000220.
43. Filipski,A., Murillo,O., Freydenzon,A., Tamura,K. and Kumar,S. (2014) Prospects for building large timetrees using molecular data with incomplete gene coverage among species. *Mol. Biol. Evol.*, **31**, 2542–2550.
44. Thompson,J.D., Linard,B., Lecompte,O. and Poch,O. (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*, **6**, e18093.
45. Nguyen,L.-T., Schmidt,H.A., von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
46. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
47. Lefort,V., Desper,R. and Gascuel,O. (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.*, **32**, 2798–2800.
48. Howe,K., Bateman,A. and Durbin,R. (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.
49. Rokas,A., Williams,B.I., King,N. and Carroll,S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.
50. Feil,E.J., Holmes,E.C., Bessen,D.E., Chan,M.S., Day,N.P., Enright,M.C., Goldstein,R., Hood,D.W., Kalia,A. *et al.* (2001) Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 182–187.
51. Hedge,J. and Wilson,D.J. (2014) Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio*, **5**, e02158-14.

52. Lapierre,M., Blin,C., Lambert,A., Achaz,G. and Rocha,E.P. (2016) The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol. Biol. Evol.*, **33**, 1711–1725.
53. List,M., Ebert,P. and Albrecht,F. (2017) Ten simple rules for developing usable software in computational biology. *PLoS Comput. Biol.*, **13**, e1005265.
54. Wilson,G., Aruliah,D.A., Brown,C.T., Chue Hong,N.P., Davis,M., Guy,R.T., Haddock,S.H., Huff,K.D., Mitchell,I.M., Plumbley,M.D. *et al.* (2014) Best practices for scientific computing. *PLoS Biol.*, **12**, e1001745.
55. Lam,M.M.C., Wick,R.R., Wyres,K.L. and Holt,K.E. (2020) Genomic surveillance framework and global population structure for Klebsiella pneumoniae. bioRxiv doi: https://doi.org/10.1101/2020.12.14.422303, 14 December 2020, preprint: not peer reviewed.
56. Touchon,M., Perrin,A., De Sousa,J.A.M., Vangchhia,B., Burn,S., O'Brien,C.L., Denamur,E., Gordon,D. and Rocha,E.P. (2020) Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLoS Genet.*, **16**, e1008866.
57. Schliep,K.P. (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

The tool is now released on https://github.com/gem-pasteur/PanACoTA. The development of the method was interspersed by several comparative genomic studies, which allowed to test the method, and add the subsequent improvements to the tool. The next chapter presents three different applications of `PanACoTA` to which I contributed, as well as their input to `PanACoTA`.

Part III

Applications to comparative genomics studies

Introduction

This part presents three different comparative genomics studies to which I participated, and where the use of PanACoTA (or its developing version) brought interesting results.

The first one corresponds to the study of an outbreak in Wisconsin (chapter 6) [145]. The second one aims at characterizing the genomic diversity of *E. coli* species (chapter 7) [190]. The third one, still in progress, explores an emerging pathogen (chapter 8).

6 *Elizabethkingia anophelis* OUTBREAK IN WISCONSIN

Our lab had recently published a paper on the evolution of *Elizabethkingia anophelis* species when an atypically large epidemic due to this bacterium was declared in Wisconsin. The CDC asked for help on the comparative genomics analysis. I participated to this study by computing and analysing pan and core genomes adapted to different questions. This allowed the discovery of an ICE which was (reversibly) inserted in a DNA repair region, and responsible for the outbreak.

I hereafter include the paper which describes this study. For space purposes, I only added the two most important (about my contribution) supplementary figures. However, all supplementary material (tables and figures) can be seen here: https://doi.org/10.6084/m9.figshare.c.3674146.v5.

nature COMMUNICATIONS

ARTICLE

Received 18 Oct 2016 | Accepted 30 Mar 2017 | Published 24 May 2017

DOI: 10.1038/ncomms15483 OPEN

Evolutionary dynamics and genomic features of the *Elizabethkingia anophelis* 2015 to 2016 Wisconsin outbreak strain

Amandine Perrin[1,2,3,*], Elise Larsonneur[1,2,4,*], Ainsley C. Nicholson[5,*], David J. Edwards[6,7], Kristin M. Gundlach[8], Anne M. Whitney[5], Christopher A. Gulvik [5], Melissa E. Bell[5], Olaya Rendueles[1,2], Jean Cury[1,2], Perrine Hugon[1,2], Dominique Clermont[9], Vincent Enouf[10], Vladimir Loparev[11], Phalasy Juieng[11], Timothy Monson[8], David Warshauer[8], Lina I. Elbadawi[12,13], Maroya Spalding Walters[14], Matthew B. Crist[14], Judith Noble-Wang[14], Gwen Borlaug[13], Eduardo P.C. Rocha[1,2], Alexis Criscuolo[3], Marie Touchon[1,2], Jeffrey P. Davis[13], Kathryn E. Holt[6,7], John R. McQuiston[5] & Sylvain Brisse[1,2,15]

An atypically large outbreak of *Elizabethkingia anophelis* infections occurred in Wisconsin. Here we show that it was caused by a single strain with thirteen characteristic genomic regions. Strikingly, the outbreak isolates show an accelerated evolutionary rate and an atypical mutational spectrum. Six phylogenetic sub-clusters with distinctive temporal and geographic dynamics are revealed, and their last common ancestor existed approximately one year before the first recognized human infection. Unlike other *E. anophelis*, the outbreak strain had a disrupted DNA repair *mutY* gene caused by insertion of an integrative and conjugative element. This genomic change probably contributed to the high evolutionary rate of the outbreak strain and may have increased its adaptability, as many mutations in protein-coding genes occurred during the outbreak. This unique discovery of an outbreak caused by a naturally occurring mutator bacterial pathogen provides a dramatic example of the potential impact of pathogen evolutionary dynamics on infectious disease epidemiology.

[1] Institut Pasteur, Microbial Evolutionary Genomics, F-75724 Paris, France. [2] CNRS, UMR 3525, F-75724 Paris, France. [3] Institut Pasteur, Hub Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, F-75724 Paris, France. [4] CNRS, UMS 3601 IFB-Core, F- 91198 Gif-sur-Yvette, France. [5] Special Bacteriology Reference Laboratory, Bacterial Special Pathogens Branch, Division of High Consequence Pathogens and Pathology, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. [6] Centre for Systems Genomics, University of Melbourne, Parkville, Victoria 3010, Australia. [7] Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia. [8] Wisconsin State Laboratory of Hygiene, Madison, Wisconsin 53718, USA. [9] CIP—Collection de l'Institut Pasteur, Institut Pasteur, F-75724 Paris, France. [10] Institut Pasteur, Pasteur International Bioresources network (PIBnet), Plateforme de Microbiologie Mutualisée (P2M), F-75724 Paris, France. [11] Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. [12] Epidemic Intelligence Service, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. [13] Division of Public Health, Wisconsin Department of Health Services, Madison, Wisconsin 53701, USA. [14] Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. [15] Institut Pasteur, Molecular Prevention and Therapy of Human Diseases, F-75724 Paris, France. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.R.M. (email: zje8@cdc.gov) or to S.B. (email: sylvain.brisse@pasteur.fr).

An outbreak of 66 laboratory-confirmed infections caused by the bacterial pathogen *Elizabethkingia anophelis* occurred in 2015–2016 in the USA states of Wisconsin (63 patients), Illinois (2 patients) and Michigan (1 patient). This was the largest ever documented *Elizabethkingia* outbreak, and the only one with illness onsets occurring primarily (89% of Wisconsin patients) in community settings. Isolates obtained from patients shared a unique genotype as defined by pulsed field gel electrophoresis, and the localized distribution of early cases was suggestive of a point source. A joint investigation by the Wisconsin Division of Public Health, Wisconsin State Laboratory of Hygiene and the Centers for Disease Control and Prevention (CDC) assessed many potential sources of the outbreak, including health-care products, personal care products, food, tap water and person-to-person transmission. The outbreak appeared to wane by mid-May 2016, and a source of infection had not yet been identified by September 2016. The ongoing investigation and updates on this outbreak are described by Centers for Disease Control and Prevention (https://www.cdc.gov/elizabethkingia/outbreaks/) and Wisconsin Department of Health Services (https://www.dhs.wisconsin.gov/disease/elizabethkingia.htm).

E. anophelis is a recently recognized species[1]. Despite recent genomic and experimental work[2–6], virulence factors or mechanisms of pathogenesis by *E. anophelis* are yet to be discovered. Knowledge of the ecology and epidemiology of this emerging pathogen is also in its infancy. All previously reported *Elizabethkingia* outbreaks have been health-care associated[7–9] although sporadic, community-acquired cases have been occasionally reported[10], as has a single instance of transmission of *E. anophelis* from mother to infant at birth[11]. Human infections have varied presentations, including meningitis and septicaemia[12–15]. Strains have been isolated from diverse environments such as hospital sinks (*E. meningoseptica* and *E. anophelis*)[6,7], the mosquito mid-gut (*E. anophelis*)[1] and the space station Mir (*E. miricola*)[16]. Therefore, *Elizabethkingiae* are generally regarded as environmental, and although *E. anophelis* has been recovered from the mid-gut of wild-caught *Anopheles* and *Aedes* mosquitoes[1], there is no indication that mosquitoes serve as a vector to transmit the bacteria to humans. *E. anophelis* is naturally resistant to multiple antimicrobial agents and harbours several genetic determinants of antimicrobial resistance, including multiple beta-lactamases and efflux systems[2,4,6,17,18]. *Elizabethkingia* species are phenotypically very similar, leading to misidentifications that compromise our understanding of the relative clinical importance of each species. Previously reported *E. meningoseptica* outbreaks may in fact have been caused by *E. anophelis,* as this latter species was recently reported to be the primary cause of clinically significant *Elizabethkingia* infections in Singapore[15].

The unique magnitude and setting of the Wisconsin outbreak and its elusive source prompted us to explore the genomic features of the outbreak strain, and compare them to other *Elizabethkingia* strains. We found that the outbreak strain represents a novel phylogenetic sublineage of *E. anophelis* and has unique genomic regions. Furthermore, it displayed exceptional evolutionary dynamism during the outbreak, likely caused by the insertion of the mobile integrative and conjugative element (ICE*Ea*1) into the *mutY* DNA repair gene.

Results

The outbreak is caused by a novel *E. anophelis* sublineage. A phylogenetic analysis was performed with the 69 Wisconsin outbreak isolates (from 59 patients) and 45 comparative strains of *E. anophelis* and other *Elizabethkingia* species (Supplementary Fig. 1a). The tree revealed three major branches, each containing one of the three *Elizabethkingia* species (*E. meningoseptica*, *E. miricola* and *E. anophelis*). The *E. miricola* branch was the most heterogeneous and comprised, in addition to *E. miricola* strains, reference strains of the distinct genomospecies defined by DNA–DNA hybridization[19]: G4071 (genomospecies 2), G4075 (genomospecies 3) and G4122 (genomospecies 4). We, therefore, labelled this branch, which may comprise several species, as the *E. miricola* cluster. The type strain JM-87^T of *E. endophytica* was placed within the *E. anophelis* branch, consistent with a recent report[20]. Eight clinical strains initially identified as *E. meningoseptica* were in fact members of the *E. anophelis* species. Additional discordances found between the phylogenetic position of several strains and their initial taxonomic designation (Supplementary Data 1) underscore the uncertainty associated with species determination for *Elizabethkingia* isolates[20].

The outbreak isolates made up a compact phylogenetic group within *E. anophelis* (sublineage 15 in Supplementary Fig. 1b), indicating that the outbreak was caused by a single ancestral strain. The long branch that separated the outbreak strain from all other sequenced *E. anophelis* strains showed that the outbreak strain is derived from a unique sublineage of *E. anophelis* that had not been previously described. The other *E. anophelis* strains were highly diverse, forming 14 other sublineages. Strains CIP 79.29 and GTC 09686 (sublineage 14) were most closely related to the outbreak strain but had a nucleotide divergence of $\sim$1%. These results show that the currently known sublineages of *E. anophelis* are not close relatives of the outbreak strain.

Phylogenetic diversity and temporal and geographic dynamics. Phylogenetic analysis of the Wisconsin isolates disclosed a highly dynamic outbreak, with a conspicuous genetic diversification into several sub-clusters (Fig. 1, Supplementary Fig. 2). Except for three outliers, all outbreak isolates derived from a single ancestor (node I, Fig. 1). We defined six main sub-clusters (sc1 to sc6, Fig. 1) based on visual inspection of the tree. Whereas sc1 branched off early, sub-clusters sc2 to sc6 shared a common ancestor (node II, Fig. 1).

Several patients were sampled on multiple occasions from 1 to 21 days apart, and from up to four different sites per patient. The cgMLST (core genome multilocus sequence typing) loci of groups of isolates from single patients were always identical, except for one single-nucleotide polymorphism (SNP) observed between isolate CSID 3000515962 and the three other isolates from the same patient. These results indicate that the pathogen population that infected each individual patient was dominated by a single genetic variant. In addition, these results underline the high reproducibility of the sequencing and genotyping processes.

The phylogenetic diversity within the outbreak clade provides an opportunity to estimate the temporal dynamics of the diversification of the outbreak strain. We first tested whether there was a temporal signal, that is, whether the root-to-tip distance was correlated with the date of sampling of bacterial isolates. Bayesian analysis with BEAST using randomized tip dates demonstrated a significant temporal signature (Supplementary Fig. 3), implying that the outbreak strain continued diversifying in a measurable way over the course of the outbreak. We next estimated a mean evolutionary rate of 5.98×10^{-6} nucleotide substitutions per site per year (95% HPD (highest posterior density) = 3.47, 8.61) based on cgMLST genes, corresponding to 24 substitutions per genome per year. This analysis placed Node I, from which all but three (including the hypermutator, see below) infectious isolates were derived, at around July 2015, and the last common ancestor of all outbreak isolates at the end of December 2014 (95% HPD = January 2014, July 2015) (Supplementary Fig. 4). Using an independent whole-genome SNP approach, the evolutionary rate estimate was

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms15483 ARTICLE

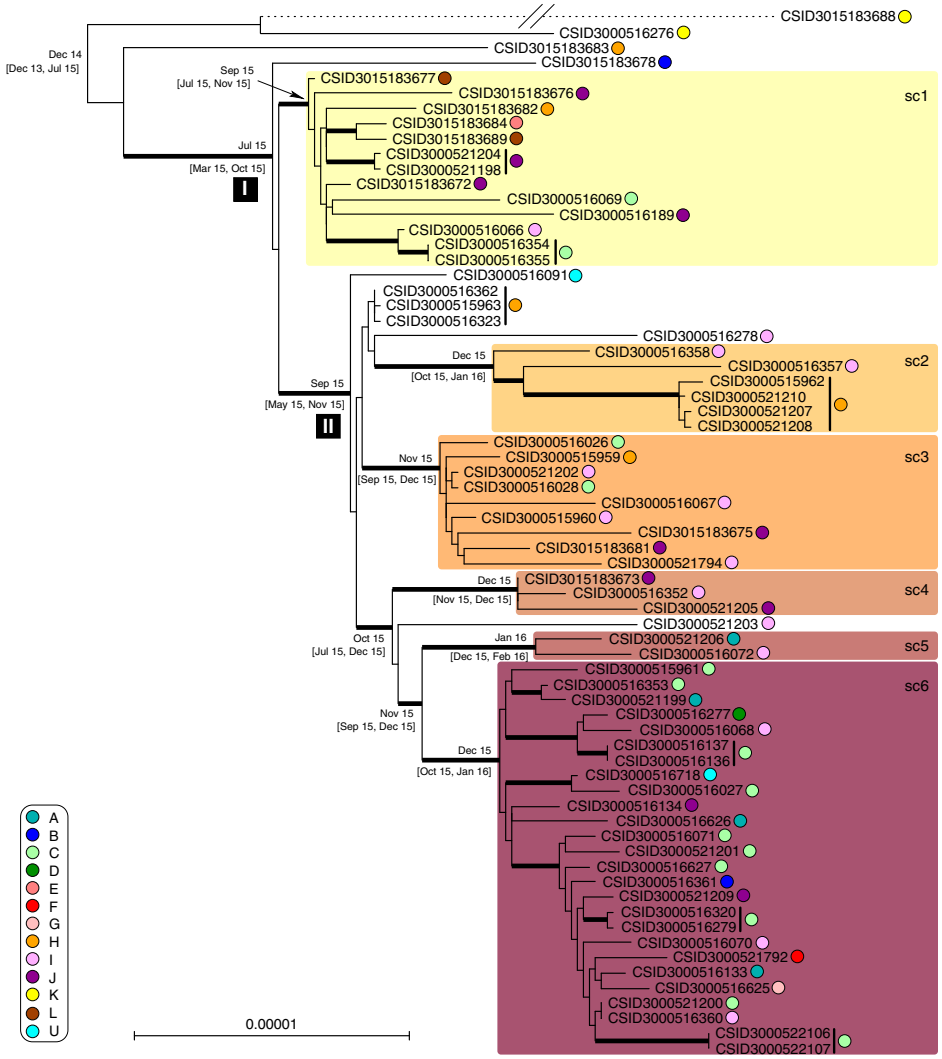


Figure 1 | Phylogenetic tree of the outbreak isolates. Maximum likelihood phylogenetic tree inferred from 3,411,033 aligned nucleotide characters (1,137,011 codons) based on cgMLST data. The tree was rooted based on phylogenetic analyses using epidemiologically unrelated *E. anophelis* strains as an outgroup. Thick branches have bootstrap support > 80% (200 replicates). The scale bar represents substitutions per site. Sub-clusters (sc) 1 to 6 are represented by coloured boxes. Counties A to L (and U for 'unspecified', attributed to the strains from outside of Wisconsin) are represented by coloured circles (see key on the left). Sets of isolates gathered from the same patient are indicated with vertical black lines after the isolate codes. Median Bayesian estimates of the month and year are provided for major internal branches (with 95% HPDs in square brackets). The branching position of the *mutS* isolate CSID 3015183688, denoted by the dashed branch line, was defined based on a separate analysis (using the same methods) and its branch length was divided by 5 for practical reasons.

6.35×10^{-6} nucleotide substitutions per site per year (95% HPD = 3.66, 9.07), and the date of the last common ancestor was estimated at August 2014 (95% HPD = June 2013, June 2015). These two approaches thus provided concordant results and suggested that the initial diversification of the outbreak strain predates the first identified human infection in this outbreak by approximately one year. Because the retrospective epidemiological analysis demonstrates that human cases of *E. anophelis* infection were likely not missed, these results suggest that the strain evolved in its reservoir during an approximately one-year interval before contaminating the source of infection, and that further diversification occurred, either in the reservoir or in the source of infection, as the outbreak was ongoing.

Phylogenetic diversification followed both temporal and geographic trends (Fig. 2). Sub-cluster sc1 appeared first, in multiple locations during the first week, and was later supplemented by the other clusters, with an initial south-east drift of cases during the first 6 weeks. Sc6 appeared later and became the most common of the sub-clusters after February 1, coinciding with concentration of cases in the south-eastern-most

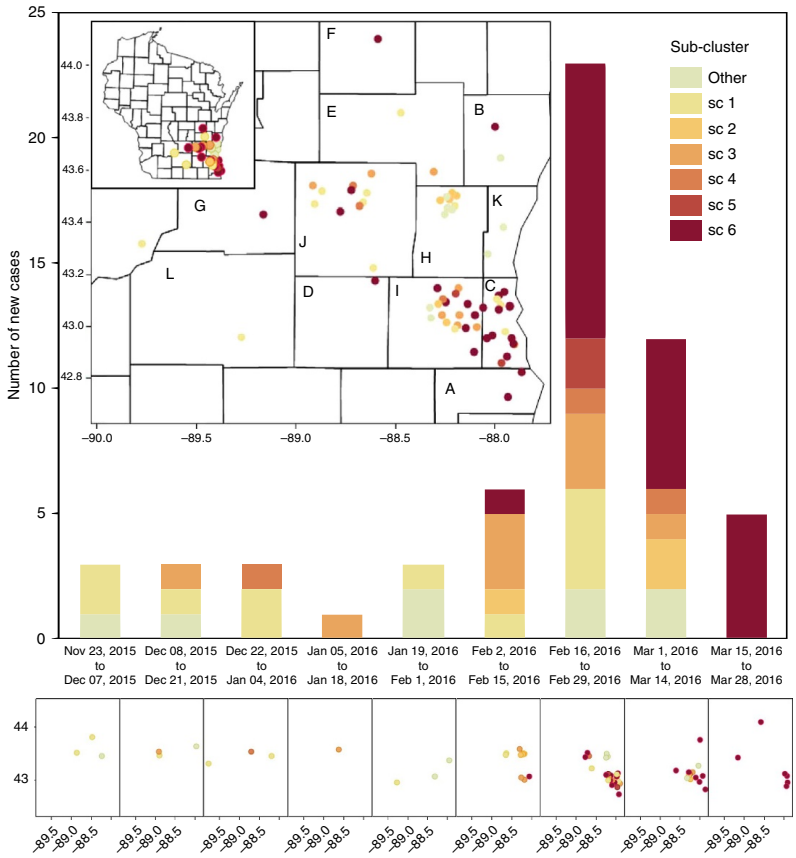


Figure 2 | Temporal-spatial distribution of cases by genetic sub-cluster. Case counts ($n=59$, over the three-state area) are presented in two-week intervals, as indicated below the histogram bars, based on the date of initial positive culture. Genetic sub-cluster colours (see key) correspond to those in the phylogenetic figures. Geographic distribution of Wisconsin cases ($n=56$) is displayed, overall (insets) and by two-week intervals (lower panel). The numbers along the *x* and *y* axis of the maps are longitude and latitude, respectively. Letters inside counties correspond to letters on the lower left key on Fig. 1.

corner of the 12 county outbreak region during the outbreak peak and followed by a wider geographic spread of sc6 after March 1. This is consistent with the relative branching order and estimated ages of sc1 and sc6 inferred from the phylogenetic analysis of genomic sequences (Fig. 1). The fit between the temporal pattern of the outbreak and the evolutionary origins of isolates provides further support to the hypothesis of genomic diversification during the outbreak. In addition, the shift from sc1 to sc6 as the dominant contributing sub-cluster may be indicative of ongoing adaptation or increasing pathogenicity of the outbreak strain.

Mutation spectrum and DNA repair defects. Three atypically divergent isolates were recognized. The isolates CSID 3000516276 and CSID 3015183683 likely represent remnants of early diverged branches. In contrast, isolate CSID 3015183688 was placed at the end of a long branch (Fig. 1), suggesting an acceleration of its substitution rate. This isolate was determined to have a mutation in its *mutS* gene, leading to a hypermutator phenotype (see Supplementary Method 3.1).

Excluding the hypermutator, 247 nucleotide positions (out of 3,411,033 in the 3,408 concatenated cgMLST gene alignments; 0.0072%) were polymorphic among the outbreak isolates. Similar nucleotide variation was demonstrated using the assembly-free approach, which detected 290 SNPs (out of 3,571,924 sites; 0.0081%). We further identified one 2 bp deletion, one 4 bp deletion, one 7 bp insertion, and five 1 bp deletions. This estimated evolutionary rate (5.98×10^{-6} substitutions per site per year within core genes, and 6.35×10^{-6} substitutions per site per year over the entire genome) is exceptionally high for a single-strain bacterial outbreak. We, therefore, analysed the mutational spectrum within the outbreak and compared it with the spectrum of the other *E. anophelis* sublineages, using the assembly-free approach. Strikingly, 253 out of 290 (87%) nucleotide substitutions along the branches of the outbreak tree were G/C->T/A transversions. This is a highly unusual pattern of mutation, and was significantly different from the mutational spectrum in the wider *E. anophelis* tree (11% G/C->T/A; Fig. 3). We noted that the mutational spectrum within the outbreak corresponds to mutations caused by the oxidative lesion 8-oxodeoxyguanosine (8-oxodG), suggesting either mutagenic growth conditions for the strain resulting from a high-oxidative stress environment, or impairment of the base excision repair pathway for 8-oxodG (the 'GO system'), which corrects

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms15483 ARTICLE

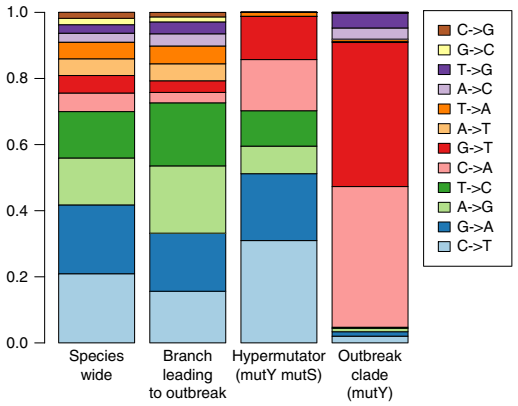


Figure 3 | Mutation spectrum of *E. anophelis* strains by clade. Frequency of each observed substitution mutation, reconstructed from FastML analysis, is shown for different parts of the *E. anophelis* tree.

these lesions[21–24]. We, therefore, inspected the genes that pertain to the GO system, and found that *mutY* was interrupted at position 841 in all outbreak isolates by the insertion of the 62,849 bp Integrative and Conjugative Element ICE*Ea*1 (for integrative and conjugative element 1 of *E. anophelis*, see below) (Fig. 4). This insertion resulted in a premature stop codon truncating the 57 terminal amino acids (aa) of the 342 aa-long MutY protein. MutY is an adenine glycosylase that functions in base excision repair to correct G-A mismatches[25]. Thus, MutY inactivation could explain the large number and atypical pattern of nucleotide substitutions observed within the outbreak. The ICE was not observed at this position in non-outbreak *E. anophelis* strains. Analysis of the mutational spectrum of substitutions on the branch leading to the outbreak strain (before its diversification started) revealed that it was very similar to that of the wider *E. anophelis* species tree (Fig. 3). This indicates that the interruption of *mutY* via insertion of ICE*Ea1* occurred shortly before the last common ancestor of the outbreak isolates.

ICE*Ea*1's integrase is 64% similar to the integrase of CTnDOT, a well-studied ICE[26,27]. We identified a potential integration site (TTT^TT) at position 841 of the *mutY* gene, flanked by inverted repeats in the ICE*Ea*1 and in the wild-type (WT) *mutY* gene (Fig. 4). We provide a model of the insertion of the ICE in a wild-type *mutY* gene (steps A and B, Fig. 4), which explains the position of the ICE in the outbreak strain. Simulating further steps of the ICE's lifecycle suggests that the ICE*Ea*1 insertion should be reversible and that the excision would reconstitute the original and functional *mutY* sequence (steps C and D, Fig. 4).

Evidence for positive selection. The atypical mutation spectrum attributed to the *mutY* truncation resulted in a very high non-synonymous to synonymous substitution ratio (ns/s = 21.4, excluding SNPs present only in the MutS- isolate), with most mutations causing amino-acid sequence alterations in the encoded proteins. Of the 49 nonsense mutations found in *mutS* competent isolates, 45 resulted from transversions unrepaired by the defective *mutY* (for example, GAA- > TAA, GAG- > TAG, and so on), including the *mutS* gene mutation resulting in the hypermutator phenotype of isolate CSID 3015183688. The substitution ratio of SNPs unique to this isolate (ns/s = 3.75) and its overall mutation spectrum (Fig. 3) were different from those of other outbreak isolates, as would be expected due to the high rate of base transition mutations in *mutS*-deficient isolates[28].

Among the 213 inferred protein changes (Supplementary Data 2, non-synonymous and nonsense mutations), some may have had important consequences regarding the virulence or resistance of the outbreak isolates, or on the fitness of the outbreak strain in its reservoir or source. We noted that the serine-83 of DNA gyrase *gyrA*, which is associated with quinolone resistance, was altered in one isolate (CSID 3000521792). Protein changes in the branch leading to node I, from which most outbreak isolates derived, may have contributed to the early adaptation of the outbreak strain to its reservoir or source. They occurred in genes

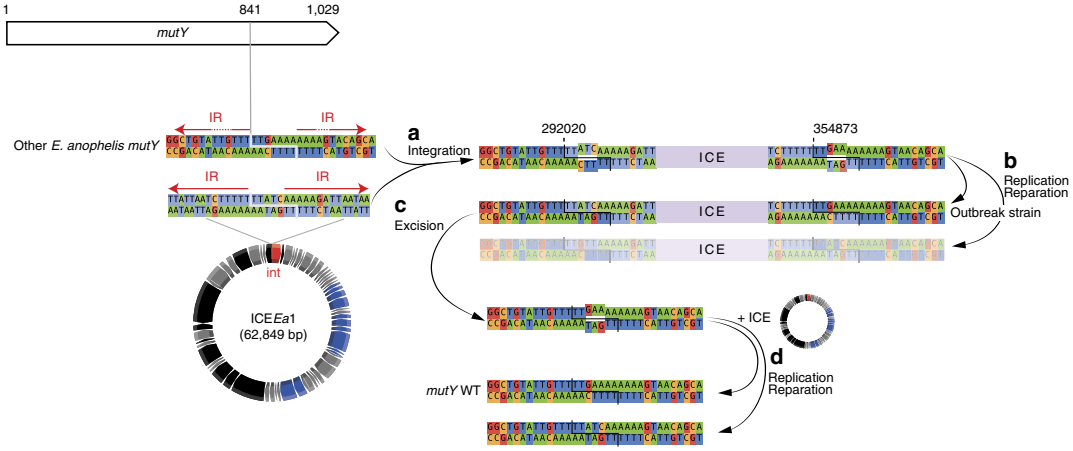


Figure 4 | Excision of ICE*Ea*1 can lead to *mutY* WT in outbreak strains. Here the insertion site is TTT^TT. In both the ICE and the *mutY*, there are inverted repeats (IR, red arrows) separated by ~5–6 nucleotides. Note that the chromosomal IRs are only partially conserved, as denoted by the interrupted arrows. (**a**) Upon insertion of the ICE at that site, this will create two heteroduplexes. (**b**) These will be resolved either by replication or by reparation. One of the two solutions to the heteroduplex resolution leads to the observed outbreak strain sequence. (**c**) If the ICE excises from the outbreak strain sequence, it will produce one heteroduplex. (**d**) The resolution of the heteroduplex left after excision of the ICE will lead to the *mutY* wild-type (WT) gene in one of the two scenarii.

coding for a TonB-dependent siderophore, a peptidase, a two-component regulator, a cysteine synthase and two ABC-transporters (Supplementary Data 2).

To detect positive selection during the course of the outbreak, we looked for genes with multiple parallel mutations. We found 27 genes that had two or more protein-altering mutations (either a non-synonymous or a nonsense mutation leading to protein truncation) that arose independently in separate branches of the tree. Prominent among these were three genes that each had five or six protein parallel mutations (Supplementary Data 2): the *wza* (A2T74_09840) and *wzc* (A2T74_09845) capsular export genes, and the gene A2T74_10040, which codes for a member of the SusD (Starch Utilization System) family of outer membrane proteins involved in binding and utilization of starch and other polysaccharides[29,30]. These observations are best explained by a strong selective pressure to abolish the function of the corresponding gene products. In light of the predominance of sub-cluster 6 towards the end of outbreak, it is interesting to note that the two changes that were specific to this sub-cluster (present in all 26 members of sc6, but in no member of other sub-clusters) were nonsense mutations in the genes *wza* and *susD* (Supplementary Data 2).

Genomic features of the outbreak strain. To define the unique genomic features of the outbreak strain, an analysis of the entire complement of protein families in *E. anophelis* genomes (that is, the *E. anophelis* pan-genome) was conducted (Supplementary Table 1). The *E. anophelis* pan-genome comprised 8,808 protein families, whereas only 3,637 protein families were observed among the 69 outbreak isolates (Supplementary Fig. 5). Further, the core-genome of the outbreak isolates represented 97% of the average number of proteins per genome, and 94% of the outbreak pan-genome. These results underline the strong homogeneity of the gene content of the outbreak isolates as compared with the extensive diversity observed within the *E. anophelis* species as a whole. Four isolates had a 77 kbp deletion affecting 75 genes; these were all from the same patient (Fig. 5; Supplementary Fig. 6; Table 1; Supplementary Data 3).

E. anophelis genomes are well known to harbour multiple genes putatively implicated in antimicrobial resistance. We found (Supplementary Data 4) that the outbreak isolates harboured resistance-associated genes previously observed in other *E. anophelis*[2,4,6,17], coding for multiple efflux systems, class A beta-lactamases, metallo-beta-lactamases and chloramphenicol acetyltransferase. Therefore, the Wisconsin outbreak strain possesses an array of antimicrobial genes similar to other *E. anophelis* strains, consistent with its multiple antimicrobial resistance phenotype (see below).

A search for putative virulence genes led to the identification of 67 genes (Supplementary Data 5). Among these, genes that were highly associated to the outbreak strain as compared with other *E. anophelis* isolates, included a CobQ/CobB/MinD/ParA nucleotide-binding domain protein located on the ICE*Ea*1 element (see below) and five genes involved in capsular polysaccharide synthesis. Capsules are important virulence factors of bacterial pathogens[31]. We, therefore, extended the search for other capsular synthesis associated genes (see Methods) and identified an identical Wzy-dependent capsular polysaccharide synthesis (*cps*) cluster in all outbreak isolates (Supplementary Fig. 7). As previously reported[2], the region of the *cps* locus that encodes for secretory proteins such as Wza and Wzc is highly conserved in *Elizabethkingia*, whereas the proteins involved in generating the specific polysaccharidic composition of the capsule are encoded in a highly variable region (outbreak-specific region 5; Fig. 5). Within the 114 *Elizabethkingia* genomes, 17 different *cps* cluster types were defined based on their gene composition pattern (Supplementary Fig. 7). Remarkably, the Wisconsin strain shared its *cps* cluster (type I) with sublineage 2 isolates, which were associated with an earlier outbreak in Singapore[2,6]. This result suggests that horizontal gene transfer of the *cps* region between *E. anophelis* sublineages may drive the emergence of virulent lineages. The *cps* gene cluster type I has so far only been observed in these two human outbreak *E. anophelis* strains (that is, the Singapore outbreak[2,6] and the Wisconsin outbreak reported here). Altogether with our observation of multiple changes of the *cps* region during the diversification of the Wisconsin outbreak strain these data suggest a possible pathogenic role for the capsular polysaccharide in the outbreak strain.

To identify genomic regions unique to, or strongly associated with, the outbreak strain, we analysed the distribution of the pan-genome protein families within *E. anophelis* and found 13 gene clusters that were conserved among outbreak isolates (present in at least 67/69 outbreak isolates) but absent in most other *E. anophelis* sublineages (Fig. 5, Supplementary Fig. 6). The functional annotations of genes located in these genomic regions suggest they may confer to the outbreak strain improved capacities to tolerate heavy metals, acquire iron, catabolize sugars or urate and synthesize bacteriocins (Table 1; Supplementary Data 3).

Most notably, the integrative and conjugative element ICE*Ea*1 was present in all outbreak isolates but was absent in most other *E. anophelis* strains (region 2 in Fig. 5 and Supplementary Fig. 6). ICE*Ea*1 belongs to the *Bacteroidetes* type 4 secretion system (T4SS-B) class[32]. It encodes the full set of components required for integration/excision and conjugation, including an integrase (tyrosine recombinase), 12 genes coding for the type IV secretion apparatus (including a VirB4 homologue and the type IV coupling protein), a relaxase (MOBP1), an ATPase (virB4) and two genes encoding for RteC, the tetracycline regulator of excision protein (Supplementary Fig. 8). Among its cargo genes (Supplementary Data 3), ICE*Ea*1 carried genes putatively coding for a RND-family cation export system composed of a cobalt-zinc-cadmium efflux pump of the *czcA/cusA* family (which was affected by two distinct non-synonymous mutations during the outbreak), followed by genes with the following annotations: nickel and cobalt (*cnrB*) and mercury (*merC*) resistance, a P-type ATPase associated with copper export (*copA*), a receptor-binding hemin, a siderophore that may allow the bacteria to fix iron from the environment (*hemR*) and a solitary N-6 DNA methylase (MTase) that might be involved in protection from restriction systems. These annotations warrant future research on a possible contribution of the ICE*Ea*1 element to detoxification of divalent cations and to acquire iron from the host during infection. Within the wider *E. anophelis* genome set, the ICE*Ea*1 element was observed in only six non-Wisconsin outbreak isolates: four isolates associated with the Singapore outbreak and strains CIP 60.59 and NCTC 10588 (Supplementary Figs 8 and 9), which were both isolated from patients with severe human infections during the 1950's (Supplementary Data 1). The association of ICE*Ea*1 with virulence deserves further functional investigation. In the six other strains, the ICE*Ea*1 element was inserted in genomic locations distant from *mutY* (Supplementary Fig. 8b). We could not find any other mobile genetic element (that is, prophages, integrons and plasmids) in the genomes of outbreak strains.

Finally, one of the outbreak-associated genomic regions comprises genes for a sodium/sugar co-transporter, a xylose isomerase and a xylose kinase (region 9, Fig. 5, Supplementary Data 3). This region was also present in the mosquito gut isolates Ag1 and R26 (region 9, Supplementary Fig. 6, Supplementary Fig. 9)[11].

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms15483 ARTICLE

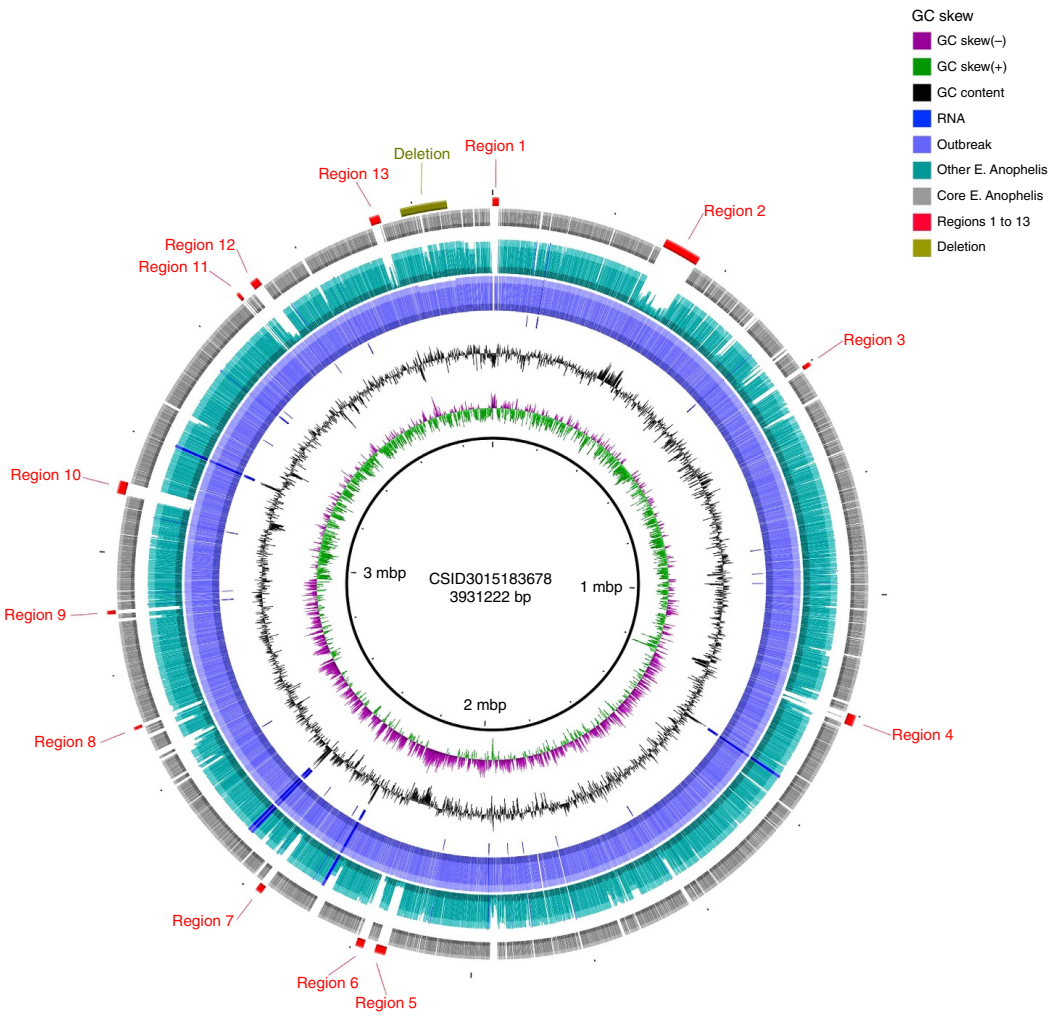


Figure 5 | Circular representation of gene content variation between the outbreak strain and 30 other *E. anophelis* genomes. Circles, from 1 (innermost circle) to 8 (outermost circle), correspond to: Circle 1: scale of the reference genome CSID 3015183678. Circle 2: GC skew (positive GC skew, green; negative GC skew, violet). Circle 3: G + C content (above average, external peaks; below average, internal peaks). Circle 4: non-coding genes (rRNA, tRNA, tmRNA); their positions are also reported in circles 5 and 6. Circle 5: frequency of CSID 3015183678 protein-encoding DNA sequences (CDSs) among the 69 outbreak isolates genomes; note the high conservation, except for a 77 kbp deletion near position 3.8 Mbp. Circle 6: frequency of CSID 3015183678 genes in all other *E. anophelis* genomes, revealing genomic regions containing CDSs with low frequency in the species as a whole. Circle 7: core genes in all 99 *E. anophelis* strains. Circle 8: remarkable genomic regions of the outbreak isolates; specific regions are marked in red, deletion in olive. Functional information about CDSs comprised in these regions is given in Table 1. The figure was obtained using BLAST Ring Image Generator (BRIG)[73]. For more details, see Supplementary Fig. 8.

Antimicrobial susceptibility of outbreak isolates. Antimicrobial susceptibility testing (Supplementary Data 6) revealed a strong homogeneity among outbreak isolates. A low susceptibility against most beta-lactams was found; isolates were resistant against ceftazidime and imipenem, but susceptible to piperacillin, piperacillin-tazobactam and cefepime. Outbreak isolates were also resistant to aminoglycosides (amikacin, gentamycin, tobramycin) and showed low *in-vitro* susceptibility to chloramphenicol, fosfomycin, tetracycline and vancomycin. These phenotypes demonstrate the high level of antimicrobial resistance of *E. anophelis* Wisconsin outbreak isolates, consistent with previous data on other *E. anophelis* isolates[6,14,15,33]. In contrast, outbreak isolates were susceptible to quinolones (ciprofloxacin, levofloxacin) and showed high *in-vitro* susceptibility to trimethoprim-sulfamethoxazole and to rifampicin. Variation in resistance among outbreak isolates was found only for chloramphenicol and for quinolones: first, isolate CSID 3000521792 was resistant to quinolones, consistent with its amino-acid alteration at position 83 of DNA gyrase subunit A (Supplementary Data 2). Second, resistance of isolate CSID 3000516072 to chloramphenicol was decreased compared with other isolates (Supplementary Data 6). Interestingly, CSID

Table 1 | Genomic features associated with the Wisconsin outbreak isolates*.

Name	Start	End	Size (nt)	Size (CDS)	Remarkable features of genomic region
Region 1	3,926,747	10,253	10,564	11	Type I restriction/modification system; DNA-invertase
Region 2	292,287	354,501	62,215	62	ICE*Ea*1; metal resistance, hemin receptor precursor; mercury resistance; enterobactin exporter
Region 3	599,595	606,529	6,935	5	Tetratricopeptide repeat protein
Region 4	1,200,465	1,219,016	18,552	13	CTP pyrophosphohydrolase
Region 5	214,2546	216,0415	17,870	17	Putative polysaccharide synthesis clusters (capsule and LPS)
Region 6	217,9815	219,3156	13,342	13	Putative polysaccharide synthesis clusters (capsule and LPS)
Region 7	2,367,659	2,378,760	11,102	8	Putative deoxyribonuclease RhsC
Region 8	2,705,573	2,710,635	5,063	5	Glycosyl hydrolase, beta-glycosidase and beta-glucosidase
Region 9	2,898,750	2,904,987	6,238	5	Xylulose kinase, xylose isomerase, sodium/glucose co-transporter
Region 10	3,097,180	3,118,179	21,000	21	Transposase; FAD-dependent urate hydroxylase (flavoprotein involved in urate degradation to allantoin)
Region 11	3,477,609	3,483,251	5,643	7	Hypothetical proteins
Region 12	3,506,671	3,521,185	14,515	10	Starch-binding outer membrane protein; Ferrienterobactin receptor precursor; Susd/RagB outer membrane lipoprotein; Nisin biosynthesis protein NisC; Putative lantibiotic biosynthesis protein
Region 13	3,727,334	3,744,334	17,001	15	Transposase, IS200-like
Deletion†	3,779,205	3,856,342	77,138	75	Multidrug resistance protein MdtE and efflux pump membrane transporter BepE; HopJ type III effector protein (found in plant pathogens); ABC-2 family transporter protein; Cytochrome c551 peroxidase precursor; H(+)/Cl(−) exchange transporter ClcA; Sulfite exporter TauE/SafE; Bicarbonate transporter BicA; Vitamin B12 transporter BtuB precursor; Putative transporter YycB; beta-lactamase

CTP, cytidine triphosphate; FAD, flavin adenine dinucleotide; LPS, lipopolysaccharide.
Positions refer to the genome of reference strain CSID 3015183678.
*Present in at least 90% of outbreak genomes and in <20% of the other *E. anophelis*.
†Absent in four *E. anophelis* outbreak genomes (patient 30).

3000516072 had an arginine to leucine alteration at position 164 of the chloramphenicol acetyltransferase, which may impact the function of this chloramphenicol resistance enzyme. As compared with the African isolates, Wisconsin outbreak isolates were more resistant to cefoxitin, amikacin and isepamycin, but less resistant to chloramphenicol, rifampicin and tetracycline. Outbreak isolates differed from the Singapore isolates by their lower resistance level to macrolides and to isepamycin. However, in the absence of interpretative breakpoints for *Elizabethkingia anophelis* antimicrobial resistance, the clinical significance of the above differences is unclear.

Discussion

We defined the phylogenetic diversity and genomic features of a strain of *E. anophelis* that caused an exceptionally large and primarily community-associated outbreak. Our phylogenetic analyses clearly established that the outbreak was caused by a single strain. The phylogenetic analysis showed that the outbreak strain represents a previously undescribed sublineage within *E. anophelis*. The nucleotide distance that separates the outbreak strain from the closest sublineages of *E. anophelis* with available genome data is nearly 1%, similar to the distance that separates, for example, clonal groups of *Klebsiella pneumoniae* with very distinct virulence properties[34,35]. These results raise the possibility that the sublineage to which the outbreak strain belongs may have evolved distinctive virulence or ecological properties, which could have contributed to the atypical size and community occurrence of the Wisconsin outbreak. For example, as xylose is one of the most abundant sugars on Earth, the genes for xylose utilization might provide a growth advantage to the outbreak strain in a reservoir, possibly in the presence of vegetation-derived nutrients. Although it is tempting to speculate on the possible link between the genomic features of the outbreak strain and the magnitude and setting of the outbreak, it is difficult to assess whether the strain has enhanced virulence in humans. The morbidity and mortality potentially attributable to *E. anophelis* infection was confounded by serious co-morbid conditions existing in patients affected by this outbreak. This work nevertheless suggests multiple avenues of research regarding the potential impact of the outbreak strain's unique capsule structure, cation detoxification capacity and sugar metabolism on its pathogenicity.

The phylogenetic position of *Elizabethkingia* strains selected for comparative purposes revealed the need for taxonomic reassignment for a large number of strains, as expected given recent taxonomic changes and the difficulty in differentiating *Elizabethkingia* species based on phenotypic characteristics. We found that several strains initially identified as *E. meningoseptica* are in fact *E. anophelis*. *E. anophelis* can be identified using matrix-assisted laser desorption ionization - time of flight (MALDI-TOF) analysis, but requires updated reference spectrum databases as found here and in a previous work[15]. This further indicates that the clinical importance of *E. anophelis* was previously underestimated, in agreement with results of a recent study[15]. These observations call for more research regarding *E. anophelis* ecology, epidemiology and virulence mechanisms.

Our results highlight important temporal and spatial patterns of the outbreak. They suggest that the bacteria may have been growing in a contaminated reservoir for nearly one year before the first infections occurred. No confirmed *E. anophelis* case could be retrospectively associated with the outbreak before November 2015. This suggests occurrence of either silent propagation resulting in human cases that remained undiagnosed or diversification of the strain in the unidentified source(s) before the initial infection of a patient. Further, the notable evolution of the pathogen during the outbreak, demonstrated by the temporal accumulation of substitutions, suggests that the source must be permissive to strain growth. Alternately, a long incubation period might precede the onset of disease, thus providing a possibility for the isolates to evolve within the patients, but the lack of diversity among multiple isolates from a single patient argues against this possibility. The uniformity of isolates from single patients also shows that although the outbreak strain has diversified, either patients were exposed to sources contaminated by a low-diversity

population, or the colonization and infectious process involves a bottleneck resulting in single clonal infection, even from a multi-contaminated source. This work thus provides a striking additional example of the now well-established power of genomic sequencing to facilitate critical re-examination of epidemiologic hypotheses and outbreak patterns[36,37].

Outbreak isolates differed by a large number of polymorphisms, and the spectrum of mutations among the outbreak isolates was unlike normal variation among other *E. anophelis*. Much less diversity is typically observed during bacterial outbreaks lasting less than one year[37,38]. Because the intra-outbreak diversity was so unusual, we confirmed it by two independent approaches: gene-by-gene analysis (cgMLST) and mapping-based SNP analysis. We identified a probable cause of this atypical mutation pattern: the disruption of the *mutY* gene coding for adenine glycosylase. Anecdotally, one strain further developed a hypermutator phenotype through a disruption of its *mutS* gene, which encodes a nucleotide-binding protein involved in the DNA mismatch repair system.

Beneficial mutations in the outbreak strain could have been selected under conditions encountered in the reservoir or the source, or during colonization or infection. Our results strongly suggest that disruptions of genes encoding proteins involved in polysaccharide utilization or capsule secretion were positively selected. Multiple outbreak isolates had alterations in the starch utilization SusD protein, and/or partial or complete disruption of either the Wza or Wzc polysaccharide transport proteins. The success of sub-cluster 6 during the later weeks of the outbreak might have resulted from the combined effect of complete disruption of both SusD and Wza. How the disruption of these functions could result in a competitive advantage for the outbreak isolates is not immediately apparent. We can speculate that the loss of capsular polysaccharides may facilitate adhesion and colonization, lead to reduced antigenicity or allow the bacteria to disperse more readily due to modified adherence to surfaces. Regardless, our results depict a dynamic outbreak strain that continued evolving while the outbreak was ongoing. One notable outcome of the exceptional genome dynamics of the outbreak strain was the replacement of sub-cluster 1 by sub-cluster 6 as the dominant subtype infecting the patients.

It is likely that the *mutY* phenotype resulted in an increased adaptive capacity of the outbreak strain. For example, the short-term advantage conferred by mutator phenotypes was previously documented in *Pseudomonas aeruginosa* infections among patients with cystic fibrosis[39]. Therefore, the integration of the ICE*Ea*1 in the *mutY* gene was likely favoured by hitchhiking with a positively selected mutation caused by the lack of this repair mechanism. In the longer run, defective DNA repair genes are counter-selected because of mutational load or because they diverge from optimal fitness peaks once the environment is stabilized[40,41]. Based on the structure of the integration site, we hypothesize that the outbreak strain could revert to a functional *mutY* sequence by losing the ICE*Ea*1 through excision, thus recovering a full capacity to repair DNA. This reversible switch of hyper-mutagenesis might have important implications regarding the future survival and possible resurgence of the Wisconsin outbreak strain. We, therefore, urge healthcare and public health systems to establish a laboratory based surveillance for *Elizabethkingia* infections, and to be particularly vigilant for a possible re-emergence of the unique *E. anophelis* strain that caused the Wisconsin outbreak.

Methods

Bacterial isolates. Wisconsin clinical laboratories were asked to submit any confirmed or suspect *Elizabethkingia* isolates to Wisconsin State Laboratory of Hygiene for identification and pulsed-field gel electrophoresis subtyping. Isolates were initially identified as *E. meningoseptica* using conventional biochemical assays and the Bruker MALDI-TOF spectral library. Pulsed-field gel electrophoresis subtyping using an in-house developed protocol, modified after consultation with CDC, was used to determine genetic relatedness among all suspect outbreak isolates. All isolates determined to be *Elizabethkingia* species were submitted to CDC for further characterization. Upon arrival, bacteria were cultivated on heart infusion agar supplemented with 5% rabbit blood agar at 35 °C. The outbreak strain isolates were correctly identified as *E. anophelis* using an expanded MALDI-TOF spectral library, genome sequencing and optical mapping. Conventional biochemical testing was restricted to oxidase, catalase and Gram stain after the MALDI-TOF spectral library provided by the CDC Special Bacteriology Reference Laboratory proved to be a reliable method of identification.

Outbreak isolates (Supplementary Data 1; labelled as Wisconsin outbreak) were primarily derived from blood (54 isolates), and also from sputum (3), bronchial wash (3), pleural fluid (1), synovial fluid (1) and other sites (7) from patients residing in 12 different counties in Southeast Wisconsin, 1 county in Illinois and 1 county in Michigan. Specimen collection dates ranged from November 2015 through March 2016. DNA was extracted using the Zymo Fungal/Bacterial DNA Microprep Kit (Zymo Research Corporation, Irvine, CA). Libraries were prepared using the NEBnext Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA), and sequence reads were generated with the Illumina MiSeq Reagent Kit v2 and MiSeq instrument (Illumina, Inc., San Diego, CA).

For comparative purposes, we included seven isolates stored in the Pasteur Institute's collection (Collection de l'Institut Pasteur or CIP; Supplementary Data 1, isolate names starting with CIP). Strains were cultivated on trypticase soy agar at 30 °C. DNA extraction was performed using the MagNA Pure 96 robotic System with the MagNA Pure 96 DNA and Viral Nucleic Acid small volume kit (Roche Diagnostics). Libraries were constructed using the Nextera XT DNA Library Preparation kit (Illumina, Inc., San Diego, CA) and sequenced on a NextSeq-500 instrument using a 2×150 paired-end protocol.

We also downloaded and included all *Elizabethkingia* genome sequences ($n = 28$ as of 28th April 2016) and sequencing read data sets ($n = 10$ as of 28th April 2016) available in sequence repositories (Supplementary Data 1).

The complete 114 *Elizabethkingia* isolate data set contained 69 Wisconsin outbreak isolates from 59 different patients (one to four isolates per patient, see Supplementary Data 1), 29 historical *E. anophelis* strains, one strain initially classified as *E. endophytica* that has been shown to belong in fact to *E. anophelis*[20], 5 *E. meningoseptica* strains, and 10 strains that belonged to the *E. miricola* cluster (see Results and Supplementary Fig. 1).

Genome assembly and annotation. For each outbreak isolate, an initial assembly was generated using the Celera De Bruijn graph assembler (Celera Genomics Workbench v8, Alameda, California). Isolate CSID 3015183678 was selected as reference for comparative genomics analyses because of its central position in an optical mapping cluster analysis of early outbreak isolates. Its contigs were ordered and oriented based on the *Nco*I optical map to generate a complete circularized genome, which was confirmed based on a PacBio genome sequence[42]. Complete circularized genomes from the other outbreak strain isolates were generated by mapping reads to the reference genome using CLC Genomics Workbench v8 (CLC bio, Waltham, MA), and manually aligned using BioEdit[43]. Indels in the circularized genomes were located using BioEdit's Positional Nucleotide Numerical Summary function.

Assemblies of the seven genomes from the CIP and from publicly available data sets for which only sequence reads were available (see Supplementary Data 1), were generated using SPAdes v.3.6.2 (ref. 44) on pre-processed reads, that is, trimming and clipping with AlienTrimmer v.0.4.0 (ref. 45), sequencing error correction with Musket v.1.1 (ref. 46), and coverage homogenization with khmer v.1.3 (ref. 47).

To obtain uniform and consistent annotations for core and pan-genome analyses, all 114 genome sequences were annotated using PROKKA v.1.11 (ref. 48). The main characteristics for each genome assembly are described in Supplementary Data 1. However, in discussion of the various loci throughout this paper, the locus tags from NCBI's Prokaryotic Genome Annotation Pipeline annotation of reference isolate CSID 3015183678 are used.

Core-genome identification. We built two core-genomes (that is, sets of orthologous proteins present in all genomes compared). The first one contained the proteins common to all *E. anophelis* genomes, while the second one contained the proteins common to all outbreak genomes. Orthologues were identified as bidirectional best hits, using end-gap free global alignment, between the reference outbreak proteome (CSID 3015183678) and each of the 98 other *E. anophelis* proteomes (for the *E. anophelis* core-genome) or each of the 68 other outbreak proteomes (for the outbreak core-genome). Hits with less than 80% similarity in amino-acid sequence or $>20\%$ difference in protein length were discarded. Because genomes from the same species typically show low levels of genome rearrangements at these short evolutionary distances, and horizontal gene transfer is frequent, proteins outside a conserved neighbourhood shared by different strains are likely to be xenologs or paralogues. Thus, for each of the previous pairwise comparisons, the list of orthologues was refined using information on the conservation of gene neighbourhood. Positional orthologues were defined as bidirectional best hits adjacent to at least four other pairs of bidirectional best hits within a

neighbourhood of ten genes (five upstream and five downstream). Finally, only the proteins having positional orthologues in 100% of the compared genomes (all *E. anophelis* genomes or all outbreak genomes) were kept. This resulted in a total of 2,530 proteins for the *E. anophelis* species core-genome, and 3,434 proteins for the Wisconsin outbreak core-genome (see Supplementary Fig. 5).

cgMLST analysis. For the core-genome MLST (cgMLST) analysis, we used two cgMLST schemes (sets of genes present in most isolates and selected for genotyping): one for the *Elizabethkingia* genus, and one for the Wisconsin outbreak isolates. The *Elizabethkingia* cgMLST scheme used was reported previously[2] and contains 1,546 genes. For the novel, Wisconsin outbreak cgMLST scheme, we started from the list of positional orthologues of the outbreak genomes (described above, in the outbreak core-genome part), and added the following conditions to ensure maximum discriminatory potential for genotyping purposes. First, we used the protein-coding sequences (coding DNA sequence, or CDS) having positional orthologues in at least 80% of the outbreak genomes. The use of this lower threshold (instead of 100% for the core-genome), allowed the use of more markers. Next, we removed from this list very small CDS (<200 bp) and genes with closely related paralogues (genes in the same genome with $>80\%$ similarity in amino-acid sequence and $<20\%$ difference in protein length). All genes already present in the *Elizabethkingia* cgMLST scheme were also discarded. These resulted in a set of 1,862 genes for the Wisconsin cgMLST scheme. These protein-coding genes, together with the 1,546 genes of the genus cgMLST scheme, constitute a total of 3,408 loci used for genotyping Wisconsin outbreak isolates of *E. anophelis*. The two cgMLST schemes are implemented in the Institut Pasteur instance of the BIGSdb database tool[49]. Allele sequences and their corresponding numerical designations are publicly accessible at http://bigsdb.pasteur.fr.

Phylogenetic analysis of cgMLST data. CDSs corresponding to the cgMLST schemes loci were aligned at the amino-acid level with MAFFT v.7.245 (ref. 50), back-translated to obtain multiple codon-based sequence alignments, and finally concatenated to obtain supermatrices of characters. This procedure was performed for (i) the entire *Elizabethkingia* sample (114 genomes) with the genus cgMLST scheme (1,546 loci, 554,224 aligned codons), and (ii) the Wisconsin outbreak isolates (69 genomes) by adding the dedicated cgMLST scheme to the genus one (total of 3,408 loci, 1,137,011 aligned codons). For each supermatrix of characters, the phylogenetic analysis was performed using IQ-TREE[51] with the codon evolutionary model being selected to minimize the BIC criterion, that is, $GY+F+\Gamma_4$ and $GY+F$[52] for the *Elizabethkingia* and for the Wisconsin outbreak samples, respectively.

Mapping-based SNP analysis. To assess variation of the entire genome including intergenic regions for phylogenetic analysis of the outbreak isolates, we used a read mapping approach. All read sets were mapped against the same reference outbreak genome sequence (CSID 3015183678) as used for core-genome and cgMLST locus definitions. Read mapping, SNP calling and preliminary filtering were completed using the RedDog phylogenomics pipeline (https://github.com/katholt/RedDog)[53]. Because we were primarily interested in phylogenetic analysis of the conserved regions of the *Elizabethkingia* genomes, SNP sites at which mapping and base calling could be confidently conducted in $<95\%$ of isolates were excluded from further analysis (most of these were located in a 77 kbp region that was deleted in four isolates that were derived from the same patient), as were SNPs located in either putative phage-associated or repeated regions of the reference genome, as detected by Phaster[54] or the nucmer algorithm of MUMmer v3 (ref. 55), respectively. We initially identified 467 SNP loci among the 69 outbreak isolates, and generated an alignment of concatenated SNP alleles at these loci. The spatial distribution of SNPs was visually inspected using Gingr[56]. A $\sim$2 kbp cluster of SNPs was identified (density >0.1, compared with density <0.01 across the rest of the genome), affecting the protease A2T74_14135 in a subset of isolates. Spatially clustered SNPs are typically introduced together via homologous recombination and thus reflect horizontal rather than vertical evolution; hence, this region was excluded from phylogenetic analysis. This yielded a final set of 374 SNPs representing changes that arose within the population of outbreak isolates, within a total core-genome of 3,571,924 bp in size (90.9% of the reference sequence).

The concatenated alignment of these SNP alleles was used to generate a maximum likelihood phylogenetic tree for the outbreak isolates using IQ-TREE[51] (see Supplementary Fig. 2, Supplementary Method 3.2 and Supplementary Data 7). SNPs were mapped back to the tree using FastML v3.1 and the details of each substitution mutation (branch, ancestral allele, derived allele) were extracted from the marginal sequences output file (Supplementary Data 2). The coding effects of the SNPs, inferred using the annotated reference genome, was defined using the parseSNPtable.py script in RedDog and analysed using R.

BEAST analyses. Date estimates of all nodes were derived using BEAST v.2.3.1 (refs 57,58) on the cgMLST supermatrix of aligned nucleotide characters (Supplementary Data 8) with the $GTR+\Gamma_4+I$ nucleotide evolutionary model (one per codon position) and lognormal relaxed-clock model. Constant population size was selected as a tree prior, and BEAST was run with 10^8 chains in order to obtain large effective sampling size values. For comparison, the BEAST analysis was also conducted on the SNP alignment (Supplementary Data 9), using a HKY substitution model and a lognormal relaxed-clock model with constant population size (Supplementary Method 3.3). The significance of the temporal signal in each analysis was assessed using the tip-date randomization technique[59–61] based on 30 samples with reshuffled dates.

Pan-genome analysis. Pan-genomes were built by clustering homologous CDSs into families. We determined the lists of putative homologs between pairs of genomes with BLASTP v.2.0 and used the *E*-values ($<10^{-4}$) to perform single-linkage clustering with SiLiX v.1.2 (ref. 62). A CDS was included in a family if it was homologous to at least one CDS already in the family. SiLiX parameters were set to consider two CDSs as homologs if their aligned part had at least 60% (*Elizabethkingia* genus) or 80% (*E. anophelis*) sequence identity and included $>80\%$ of the smallest CDS. The pan-genomes of *Elizabethkingia* and of the outbreak isolates were determined independently.

Detection of capsular gene clusters. To identify capsular gene clusters, we used our previous approach[2]. In brief, we performed a keyword search of the Pfam database v.29.0 (http://pfam.xfam.org) for protein profiles involved in capsular polysaccharide production such as glycosyl transferases, ABC transporters, Wzx flippase and Wzy polymerase. We then performed a search of these profiles in *Elizabethkingia* genomes using HMMER3 v.3.1b1 (ref. 63) with the *E*-values $<10^{-4}$ and a coverage threshold of 50% of the protein. After the identification of a putative capsular cluster across all genomes, several proteins within the cluster did not match any of the previously selected protein profiles. For completeness, we searched these proteins for known functional domains against the PFAM database using the command hmmscan included in the software HMMER3, and recorded their family and/or annotation (see Supplementary Fig. 7, and regions 5 and 6 of Supplementary Data 3).

Antimicrobial resistance and virulence-associated genes. Acquired antimicrobial resistance genes were detected using HMMER3 v.3.1b1 to screen genome sequences against the ResFams (Core v.1.2), a curated database of antimicrobial resistance protein families and associated profile hidden Markov models with the cut_ga option[64] (Supplementary Data 4). Virulence-associated genes were identified by screening genome sequences against the VFDB 2016 (ref. 65) using BLASTP v.2.0 (minimum 40% identity with *E*-value $<10^{-5}$), as in (ref. 5) (Supplementary Data 5).

Detection of mobile genetic elements. ICEs were identified and classified using MacSyFinder v.1.0.2 (ref. 66) with TXSScan profiles[67]. CRISPR-Cas systems were searched using MacSyFinder v.1.0.2 with Cas-Finder profiles[66] and CRISPR-Finder[68], with default parameters. Integrons were searched using IntegronFinder v.1.4 with –local_max option[69], and prophages using VirSorter v.1.0.3 on RefSeqDB only[70] and PhageFinder v.4.6 (ref. 71).

Antimicrobial susceptibility testing. Antimicrobial susceptibility testing was performed by Kirby Bauer disk diffusion method (http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_6.0_Breakpoint_table.pdf)[72]. As no interpretative criteria exist for *Elizabethkingia*, results were interpreted according to European Committee on Antimicrobial Susceptibility Testing (EUCAST) criteria for *Pseudomonas spp.* We tested a broad range of antibiotics: beta-lactams (piperacillin, cefotaxime, ceftazidime, imipenem, ampicillin, amoxicillin, amoxicillin-clavulanic acid, cephalexin, cefuroxime, cefoxitin, cefepime, cefoperazone-sulbactam, piperacillin-tazobactam), aminoglycosides (streptomycin, amikacin, isepamycin, tobramycin, gentamicin, kanamycin), quinolones (nalidixic acid, ciprofloxacin, pefloxacin, levofloxacin, moxifloxacin), macrolides (erythromycin, clarithromycin, spiramycin, azithromycin) and other classes (chloramphenicol, sulfamethoxazole-trimethoprim, fosfomycin, rifampicin, linezolid, tetracyclin, vancomycin and tigecyclin).

Data availability. Reads for all outbreak isolates and complete genome sequences of outbreak isolates CSID 3015183678, CSID 3015183684, CSID 3000521207 and CSID 3015183681 were submitted to NCBI, associated with project ID PRJNA315668. Reads and draft genome sequences for strains Po0527107 and V0378064 (ref. 2) were submitted to the European Nucleotide Archive and are available under their respective project IDs, PRJEB5243 and PRJEB5242. Reads for strains CIP 78.9, CIP 60.59, CIP 104057, CIP 108654, CIP 79.29, CIP 80.33 and CIP 108653 were submitted to the European Nucleotide Archive and are available under project ID PRJEB14302. In addition, every genome sequence assembled during this study is available in the Institut Pasteur instance of the BIGSdb database tool dedicated to *Elizabethkingia* (http://bigsdb.web.pasteur.fr/elizabethkingia). Supplementary data, tables and high resolution figures are available through FigShare at this link (https://doi.org/10.6084/m9.figshare.c.3674146.v5). We also created a project on microreact, available at this link: https://microreact.org/project/SyaeGCjvg.

References

1. Kampfer, P. *et al. Elizabethkingia anophelis* sp. nov., isolated from the midgut of the mosquito *Anopheles gambiae*. *Int. J. Syst. Evol. Microbiol.* **61,** 2670–2675 (2011).
2. Breurec, S. *et al.* Genomic epidemiology and global diversity of the emerging bacterial pathogen *Elizabethkingia anophelis*. *Sci. Rep.* **6,** 30379 (2016).
3. Chen, S., Bagdasarian, M. & Walker, E. D. *Elizabethkingia anophelis*: molecular manipulation and interactions with mosquito hosts. *Appl. Environ. Microbiol.* **81,** 2233–2243 (2015).
4. Kukutla, P. *et al.* Insights from the genome annotation of *Elizabethkingia anophelis* from the malaria vector *Anopheles gambiae*. *PLoS ONE* **9,** e97715 (2014).
5. Li, Y. *et al.* Complete genome sequence and transcriptomic analysis of the novel pathogen *Elizabethkingia anophelis* in response to oxidative stress. *Genome Biol. Evol* **7,** 1676–1685 (2015).
6. Teo, J. *et al.* Comparative genomic analysis of malaria mosquito vector-associated novel pathogen *Elizabethkingia anophelis*. *Genome Biol. Evol.* **6,** 1158–1165 (2014).
7. Moore, L. S. P. *et al.* Waterborne *Elizabethkingia meningoseptica* in adult critical care. *Emerg. Infect. Dis.* **22,** 9–17 (2016).
8. Balm, M. N. D. *et al.* Bad design, bad practices, bad bugs: frustrations in controlling an outbreak of *Elizabethkingia meningoseptica* in intensive care units. *J. Hosp. Infect.* **85,** 134–140 (2013).
9. Tak, V., Mathur, P., Varghese, P. & Misra, M. C. *Elizabethkingia meningoseptica*: an emerging pathogen causing meningitis in a hospitalized adult trauma patient. *Indian J. Med. Microbiol.* **31,** 293–295 (2013).
10. Hayek, S. S. *et al.* Rare *Elizabethkingia meningosepticum* meningitis case in an immunocompetent adult. *Emerg. Microbes Infect.* **2,** e17 (2013).
11. Lau, S. K. P. *et al.* Evidence for *Elizabethkingia anophelis* transmission from mother to infant, Hong Kong. *Emerg. Infect. Dis.* **21,** 232–241 (2015).
12. King, E. O. Studies on a group of previously unclassified bacteria associated with meningitis in infants. *Am. J. Clin. Pathol.* **31,** 241–247 (1959).
13. Bloch, K. C., Nadarajah, R. & Jacobs, R. *Chryseobacterium meningosepticum*: an emerging pathogen among immunocompromised adults. Report of 6 cases and literature review. *Medicine (Baltimore)* **76,** 30–41 (1997).
14. Frank, T. *et al.* First case of *Elizabethkingia anophelis* meningitis in the Central African Republic. *Lancet (London, England)* **381,** 1876 (2013).
15. Lau, S. K. P. *et al. Elizabethkingia anophelis* bacteremia is associated with clinically significant infections and high mortality. *Sci. Rep.* **6,** 26045 (2016).
16. Kim, K. K., Kim, M. K., Lim, J. H., Park, H. Y. & Lee, S.-T. Transfer of *Chryseobacterium meningosepticum* and *Chryseobacterium miricola* to *Elizabethkingia* gen. nov. as *Elizabethkingia meningoseptica* comb. nov. and *Elizabethkingia miricola* comb. nov. *Int. J. Syst. Evol. Microbiol.* **55,** 1287–1293 (2005).
17. Bellais, S., Aubert, D., Naas, T. & Nordmann, P. Molecular and biochemical heterogeneity of class B carbapenem-hydrolyzing beta-lactamases in *Chryseobacterium meningosepticum*. *Antimicrob. Agents Chemother.* **44,** 1878–1886 (2000).
18. González, L. J. & Vila, A. J. Carbapenem resistance in *Elizabethkingia meningoseptica* is mediated by metallo-β-lactamase BlaB. *Antimicrob. Agents Chemother.* **56,** 1686–1692 (2012).
19. Holmes, B., Steigerwalt, A. G. & Nicholson, A. C. DNA-DNA hybridization study of strains of *Chryseobacterium*, *Elizabethkingia* and *Empedobacter* and of other usually indole-producing non-fermenters of CDC groups IIc, IIe, IIh and IIi, mostly from human clinical sources, and proposals of Chryseobacterium bernardetii sp. nov., Chryseobacterium carnis sp. nov.,Chryseobacterium lactis sp. nov., Chryseobacterium nakagawai sp. nov. and Chryseobacterium taklimakanense comb. nov. *Int. J. Syst. Evol. Microbiol.* **63,** 4639–4662 (2013).
20. Doijad, S., Ghosh, H., Glaeser, S., Kämpfer, P. & Chakraborty, T. Taxonomic reassessment of the genus *Elizabethkingia* using whole genome sequencing: *Elizabethkingia endophytica* Kämpfer *et al.* 2015 is a later subjective synonym of *Elizabethkingia anophelis* Kämpfer *et al.* 2011. *Int. J. Syst. Evol. Microbiol.* **66,** 4555–4559 (2016).
21. Michaels, M. L. & Miller, J. H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J. Bacteriol.* **174,** 6321–6325 (1992).
22. Boiteux, S. & Radicella, J. P. Base excision repair of 8-hydroxyguanine protects DNA from endogenous oxidative stress. *Biochimie* **81,** 59–67 (1999).
23. van Loon, B., Markkanen, E. & Hübscher, U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. *DNA Repair (Amst)* **9,** 604–616 (2010).
24. David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447,** 941–950 (2007).
25. Au, K. G., Clark, S., Miller, J. H. & Modrich, P. *Escherichia coli mutY* gene encodes an adenine glycosylase active on G-A mispairs. *Proc. Natl Acad. Sci. USA* **86,** 8877–8881 (1989).
26. Malanowska, K., Salyers, A. A. & Gardner, J. F. Characterization of a conjugative transposon integrase, IntDOT. *Mol. Microbiol.* **60,** 1228–1240 (2006).
27. Laprise, J., Yoneji, S. & Gardner, J. F. Homology-dependent interactions determine the order of strand exchange by IntDOT recombinase. *Nucleic Acids Res.* **38,** 958–969 (2010).
28. Schaaper, R. M. & Dunn, R. L. Spectra of spontaneous mutations in *Escherichia coli* strains defective in mismatch correction: the nature of *in vivo* DNA replication errors. *Proc. Natl Acad. Sci. USA* **84,** 6220–6224 (1987).
29. Mackenzie, A. K. *et al.* Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant bacteroidetes phylotype bind strongly to cellulose. *Appl. Environ. Microbiol.* **78,** 5935–5937 (2012).
30. Shipman, J. A., Berleman, J. E. & Salyers, A. A. Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. *J. Bacteriol.* **182,** 5365–5372 (2000).
31. Moxon, E. R. & Kroll, J. S. The role of bacterial polysaccharide capsules as virulence factors. *Curr. Top. Microbiol. Immunol.* **150,** 65–85 (1990).
32. Guglielmini, J., de la Cruz, F. & Rocha, E. P. C. Evolution of conjugation and Type IV secretion systems. *Mol. Biol. Evol.* **30,** 315–331 (2013).
33. Hsu, M. S. *et al.* Clinical features, antimicrobial susceptibilities, and outcomes of *Elizabethkingia meningoseptica* (*Chryseobacterium meningosepticum*) bacteremia at a medical center in Taiwan, 1999-2006. *Eur. J. Clin. Microbiol. Infect. Dis.* **30,** 1271–1278 (2011).
34. Bialek-Davenet, S. *et al.* Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. *Emerg. Infect. Dis.* **20,** 1812–1820 (2014).
35. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl Acad. Sci.* **112,** E3574–E3581 (2015).
36. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–474 (2010).
37. Grad, Y. H. *et al.* Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc. Natl Acad. Sci. USA* **109,** 3065–3070 (2012).
38. Zhou, Z. *et al.* Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet.* **9,** e1003471 (2013).
39. Oliver, A. & Mena, A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin. Microbiol. Infect.* **16,** 798–808 (2010).
40. Denamur, E. & Matic, I. Evolution of mutation rates in bacteria. *Mol. Microbiol.* **60,** 820–827 (2006).
41. Söderberg, R. J. & Berg, O. G. Kick-starting the ratchet: the fate of mutators in an asexual population. *Genetics* **187,** 1129–1137 (2011).
42. Nicholson, A. C. *et al.* Complete genome sequences of four strains from the 2015-2016 *Elizabethkingia anophelis* outbreak. *Genome Announc.* **4,** e00563–16 (2016).
43. Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41,** 95–98 (1999).
44. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
45. Criscuolo, A. & Brisse, S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102,** 500–506 (2013).
46. Liu, Y., Schroder, J. & Schmidt, B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29,** 308–315 (2013).
47. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res.* **4,** 900 (2015).
48. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
49. Jolley, K. A. & Maiden, M. C. J. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11,** 1–11 (2010).
50. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
51. Nguyen, L. -T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).
52. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736 (1994).
53. Schultz, M. B. B. *et al.* Repeated local emergence of carbapenem-resistant *Acinetobacter baumannii* in a single hospital ward. *Microb. Genom.* **2,** e000050 (2016).
54. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44,** W16–W21 (2016).
55. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
56. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15,** 524 (2014).

57. Drummond, A. J. & Rambaut, A. BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
58. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29,** 1969–1973 (2012).
59. Duffy, S. & Holmes, E. C. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J. Gen. Virol.* **90,** 1539–1547 (2009).
60. Ramsden, C., Holmes, E. C. & Charleston, M. A. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol. Biol. Evol.* **26,** 143–153 (2009).
61. Firth, C. *et al.* Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.* **27,** 2038–2051 (2010).
62. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12,** 116 (2011).
63. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41,** e121 (2013).
64. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9,** 207–216 (2015).
65. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res.* **44,** D694–D697 (2016).
66. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9,** e110726 (2014).
67. Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6,** 23080 (2016).
68. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35,** W52–W57 (2007).
69. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44,** 4539–4550 (2016).
70. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3,** e985 (2015).
71. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34,** 5839–5851 (2006).
72. EUCAST. Breakpoint Tables for Interpretation of MICs and Zone Diameters, Version 6.0. http://www.eucast.org/clinical_breakpoints (2016).
73. Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12,** 402 (2011).

Acknowledgements

We thank D. Mornico of Institut Pasteur and V. Nyak of CDC for assistance with submission of sequence data to public repositories. We would also like to thank the State Health Departments of Michigan and Illinois for contributing strains and information for the cases outside of the State of Wisconsin. The efforts of laboratory staff in both DHQP and DHCPP are greatly appreciated. This work was supported by Institut Pasteur, French government's Investissement d'Avenir program Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant ANR-10-LABX-62-IBEID), and the Advanced Molecular Detection (AMD) initiative at CDC. O.R. was supported by a fellowship from Fondation pour la Recherche Médicale (grant number ARF20150934077). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Author contributions

This project was designed by K.E.H., J.R.M. and S.B. Specimens and epidemiologic data were collected by K.M.G., T.M., D.W., L.I.E., M.S.W., M.B.C., J.N.-W., G.B. and J.P.D. Whole-genome sequences were generated and assembled by A.C.N., A.M.W., M.E.B., O.R., J.C., D.C., A.C. and V.E. Optical mapping was done by V.L. and P.J. cgMLST analysis was done by A.C. and E.L. Read mapping and SNP analysis was done by A.C.N., K.E.H. and D.J.E. Core and pan-genome analyses were done by A.P. and M.T. Capsular cluster analysis was done by O.R. Analysis of the ICE*Ea*1 integration site in *mutY* was performed by A.C.N. and J.C. Phylogenetic and BEAST analyses were done by D.J.E., K.E.H. and A.C. Antimicrobial susceptibility testing was performed by P.H. and S.B. Additional data analyses and figure creation were done by A.P., A.C.N., A.C., M.T., C.A.G., K.E.H. and S.B. Overall coordination of the study and of manuscript writing was done by S.B. The manuscript was drafted and edited by A.P., E.L., A.C.N., K.E.H., T.M., D.J.E., C.A.G., M.S.W., M.T., E.P.C.R., J.P.D., J.R.M. and S.B. All authors provided final approval of the version submitted for publication.

Additional information

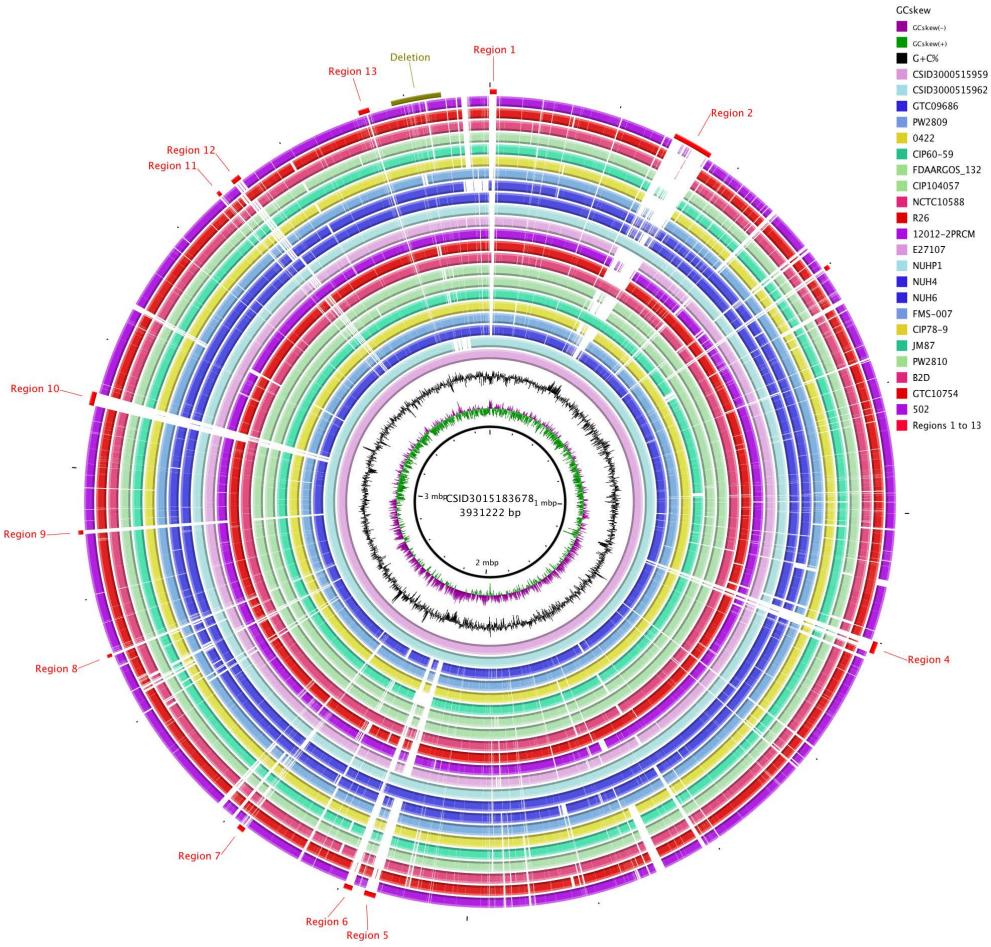
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing interests: The authors declare no competing financial interests.

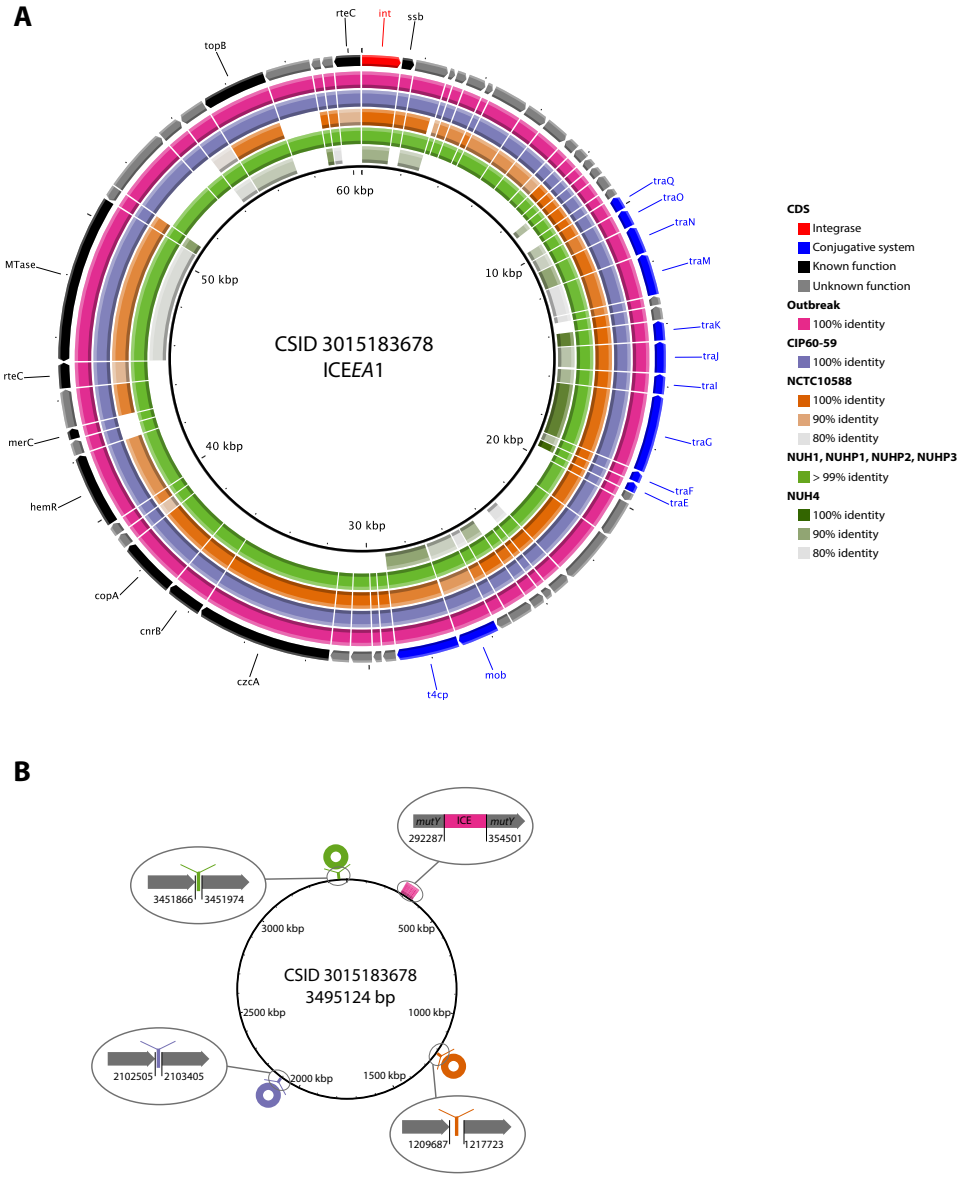
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Perrin, A. *et al.* Evolutionary dynamics and genomic features of the *Elizabethkingia anophelis* 2015 to 2016 Wisconsin outbreak strain. *Nat. Commun.* **8,** 15483 doi: 10.1038/ncomms15483 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Supplementary figure 8 : Circular representation of gene conservation between the reference outbreak isolate E. anophelis CSID 3015183678 and other E. anophelis genomes. Circles are numbered from 1 (innermost circle) to 26 (outermost circle). Circle 1: scale of the CSID 3015183678 genome. Circle 2: GC skew (positive GC skew, green; negative GC skew, violet). Circle 3: G+C content (above average, external peaks; below average, internal peaks). Circle 4: all Wisconsin outbreak isolates except the 4 isolates in circle 5. Circle 5: Wisconsin outbreak isolates CSID 3000515962, CSID 3000521207, CSID 3000521208, CSID 3000521210, which are deleted for a 77 kb region (Deletion, last circle). Circle 6: strain GTC09686. Circle 7: strain PW2809. Circle 8: strain 0422. Circle 9: strain CIP60.59. Circle 10: strain FDAARGOS_132. Circle 11: strain CIP104057. Circle 12: strain NCTC10588. Circle 13: strain R26. Circle 14: strain 12012-2PRCM. Circle 15: strain E27107. Circle 16: strain NUHP1. Circle 17: strain NUH4. Circle 18: strain NUH6. Circle 19: strain FMS-007. Circle 20: strain CIP78.9. Circle 21: strain JM87. Circle 22: strain PW2810. Circle 23: strain B2D. Circle 24: strain GTC10754. Circle 25: strain 502. Circle 26: outbreak-associated genomic regions 1 to 13, and deletion. This representation was performed using BRIG with options: `blastn -F F -e 0.001 -W 10`.



Supplementary figure 10 : Circular representation of the integrative and conjugative element ICEEa1 and gene conservation and ICE localization among representative *E. anophelis* genomes. (A) ICEEa1 content among representative genomes. Circles are numbered from 1 (innermost circle) to 6 (outermost circle). The protein coding genes of ICEEa1 in the reference genome CSID 3015183678 (Circle 6) are compared to: Circle 1: Singapore outbreak strain NUH4; Circle 2: Singapore outbreak strains NUH1, NUHP1, NUHP2 and NUHP3 (all identical). Circle 3: strain NCTC 10588; Circle 4: strain CIP60.59; Circle 5: the 68 other Wisconsin outbreak isolates. On circle 6, genes of the conjugative system are colored in blue, those with a known function in black, and those with unknown function in gray. The figure was obtained using BRIG with option: "`blastp -evalue 0.001 -seg no`", and representing, for each genome, only the proteins resulting from bidirectional best hit (BBH) with 80% similarity or more, and showing synteny (at least 4 syntenic proteins among a radius of 5) with the reference ICE proteins. Identity percentage in the legend is given for deduced amino acid sequences. (B) Relative position on the reference genome of the ICEEa1 element in all the strains that harbor it (same color code as in (A)).

7 THE DIVERSITY OF *E. coli* SPECIES

A collaboration with an Australian team gave us the opportunity to analyse the genomic diversity of *E. coli* species, thanks to a dataset of over one thousand of highly diverse environmental and host-associated strains. I participated to the very beginning (mainly consisting in the preparation of the dataset) and the very end (uploading sequences on ENA platform) of this study.

The first step was the curation of the original dataset, which was made of genomes from many different sequencing, assembling and annotation methods, in order to provide more consistency. This was done thanks to what became the `annotate` module of PanACoTA, which includes a quality control step before the uniform annotation.

Next steps were pan, core and persistent genome computations, which were also done thanks to the embryo of PanACoTA. Before this study, we had tested our pangenome method under construction on datasets of several hundreds of genomes. This study gave us the opportunity to check the behavior of the method at a larger scale: a dataset of thousands of genomes. The test was conclusive: we did not have difficulties in generating the pangenome.

As the core genome of this dataset was too small to infer a reliable phylogeny, we used the persistent genome. Performing MSA on persistent families raised questions on how to handle genomes without any gene, or with several genes in the family. This led to the definition of three types of persistent genomes, now implemented in `corepers` module of PanACoTA (see chapter 5): strict, mixed and multi persistent [146]. The `align` module was also adapted to these new definitions.

PLOS GENETICS

RESEARCH ARTICLE

Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*

Marie Touchon[1]*, Amandine Perrin[1,2], Jorge André Moura de Sousa[1], Belinda Vangchhia[3,4], Samantha Burn[3], Claire L. O'Brien[5], Erick Denamur[6,7], David Gordon[3], Eduardo PC Rocha[1]

1 Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, 25-28 rue Dr Roux, Paris, 75015, France, **2** Sorbonne Université, Collège doctoral, F-75005, Paris, France, **3** Ecology and Evolution, Research School of Biology, The Australian National University, Acton, ACT, Australia, **4** Department of Veterinary Microbiology, College of Veterinary Sciences & Animal Husbandry, Central Agricultural University, Selesih, Aizawl, Mizoram, India, **5** School of Medicine, University of Wollongong, Northfields Ave Wollongong, Australia, **6** Université de Paris, IAME, UMR 1137, INSERM, Paris, 75018, France, **7** AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, 75018, Paris, France

* mtouchon@pasteur.fr

Citation: Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, et al. (2020) Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. PLoS Genet 16(6): e1008866. https://doi.org/10.1371/journal.pgen.1008866

Editor: Xavier Didelot, University of Warwick, UNITED KINGDOM

Received: February 19, 2020

Accepted: May 18, 2020

Published: June 12, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1008866

Data Availability Statement: We have provided the underlying numerical data for all graphs and summary statistics in the Supporting Information files, the S1 Dataset and S2 Dataset. Genome

Abstract

Escherichia coli is mostly a commensal of birds and mammals, including humans, where it can act as an opportunistic pathogen. It is also found in water and sediments. We investigated the phylogeny, genetic diversification, and habitat-association of 1,294 isolates representative of the phylogenetic diversity of more than 5,000 isolates from the Australian continent. Since many previous studies focused on clinical isolates, we investigated mostly other isolates originating from humans, poultry, wild animals and water. These strains represent the species genetic diversity and reveal widespread associations between phylogroups and isolation sources. The analysis of strains from the same sequence types revealed very rapid change of gene repertoires in the very early stages of divergence, driven by the acquisition of many different types of mobile genetic elements. These elements also lead to rapid variations in genome size, even if few of their genes rise to high frequency in the species. Variations in genome size are associated with phylogroup and isolation sources, but the latter determine the number of MGEs, a marker of recent transfer, suggesting that gene flow reinforces the association of certain genetic backgrounds with specific habitats. After a while, the divergence of gene repertoires becomes linear with phylogenetic distance, presumably reflecting the continuous turnover of mobile element and the occasional acquisition of adaptive genes. Surprisingly, the phylogroups with smallest genomes have the highest rates of gene repertoire diversification and fewer but more diverse mobile genetic elements. This suggests that smaller genomes are associated with higher, not lower, turnover of genetic information. Many of these genomes are from freshwater isolates and have peculiar traits, including a specific capsule, suggesting adaptation to this environment. Altogether, these data contribute to explain why epidemiological clones tend to emerge from specific phylogenetic groups in the presence of pervasive horizontal gene transfer across the species.

assemblies have been deposited into the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB34791. The accession number of each genome assembly are indicated in the S1 Dataset. They are all available: https://www.ebi.ac.uk/ena/data/view/PRJEB34791 and https://www.ncbi.nlm.nih.gov/bioproject/PRJEB34791/.

Funding: This work was supported by in-house funding from Pasteur Institute and the CNRS (M.T., A.P., JAM.S. and EPC.R.) and was partially supported by grants from the Fondation pour la Recherche Médicale (https://www.frm.org/) [Equipe FRM 2016, grant DEQ20161136698 to E. D., and Equipe FRM: EQU201903007835 to EPC. R.], by the Laboratoire d'Excellence IBEID (https://research.pasteur.fr/fr/program_project/integrative-biology-of-emerging-infectious-diseases/) [grant ANR-10-LABX-62-IBEID to EPC.R.), by the INCEPTION project (https://research.pasteur.fr/en/program_project/inception/) [grant PIA/ANR-16-CONV-0005 to EPC.R.] and by an Australian Research Council Linkage Grant [grant LP120100327 to D.G., B.V., S.B.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Previous large scale studies on the evolution of *E. coli* focused on clinical isolates emphasizing virulence and antibiotic resistance in medically important lineages. Yet, most *E. coli* strains are either human commensals or not associated with humans at all. Here, we analyzed a large collection of non-clinical isolates of the species to assess the mechanisms of gene repertoire diversification in the light of isolation sources and phylogeny. We show that gene repertoires evolve so rapidly by the high turnover of mobile genetic elements that epidemiologically indistinguishable strains can be phenotypically extremely heterogeneous, illustrating the velocity of bacterial adaptation and the importance of accounting for the information on the whole genome at the epidemiological scale. Phylogeny and habitat shape the genetic diversification of *E. coli* to similar extents. Surprisingly, freshwater strains seem specifically adapted to this environment, breaking the paradigm that *E. coli* environmental isolates are systematically fecal contaminations. As a consequence, the evolution of this species is also shaped by environmental habitats, and it may diversify by acquiring genes and mobile elements from environmental bacteria (and not just from gut bacteria). This may facilitate the acquisition of virulence factors and antibiotic resistance in the strains that become pathogenic.

Introduction

The integration of epidemiology and genomics has greatly contributed to our understanding of the population genetics of epidemic clones of pathogenic bacteria. However, the forces driving the emergence of these lineages in species where most clades are dominated by commensal or environmental strains remain unclear. *Escherichia coli* is a commensal of the gut microbiota of mammals and birds (primary habitat) [1–3], and has been found in host-independent secondary habitats including soil, sediments, and water [4–7]. Yet, some *E. coli* strains produce virulence factors endowing them with the ability to cause a broad range of intestinal or extra-intestinal diseases (pathotypes) in humans and domestic animals [8–13]. Many of these are becoming resistant to multiple antibiotics at a worrisome pace [14, 15].

Studies on *E. coli* were seminal in the development of bacterial population genetics [16]. They showed moderate levels of recombination in the species [3, 17–19], and a strong phylogenetic structure with eight main phylogroups, among which four (A, B1, B2 and D) represent the majority of the strains and four others (C, E, F and G) are rarer [20–22]. Strains differ in their phenotypic and genotypic characteristics within and across phylogroups [2, 3, 23, 24], and their isolation frequency depends on factors such as host species, diet, sex, age [25–27], body mass [28], but also climate [29, 30], and geographic location [31]. Strains of phylogroups A and B1 appear to be more generalists since they can be isolated from all vertebrates [2] and are often isolated from secondary habitats [7, 32–35]. *E. coli* strains able to survive and persist in water environments usually belong to the B1 phylogroup [7, 33, 34]. In contrast, the extraintestinal pathogenic strains usually belong to phylogroups B2 and D [36–38]. Genome size also differs among phylogroups, with A and B1 strains having smaller genomes than B2 or D strains [23].

The phylogenetic vicinity of geographically remote *E. coli* isolates, and the co-isolation of phylogenetically distant strains, supports the hypothesis that strains circulate rapidly across the globe [39, 40]. The genome of the species is also remarkably plastic, since only about half of the average genome is present across most strains of the species and the pan-genome vastly

exceeds the size of the typical genome [41–44]. Interestingly, the rapid circulation of strains and the high plasticity of their genomes have not erased the associations of certain clades with certain isolation sources. These associations might reflect local adaptation to the isolation habitat [16, 45], which would suggest frequent genetic interactions between the novel adaptive changes and the strains' genomic background.

Understanding how the evolution of gene repertoires is shaped by population structure and habitats requires large-scale comparative genomics of samples with diverse sources of isolation representative of natural populations of *E. coli*. Most of the efforts of genome sequencing have been devoted to study pathogenic lineages and very few genomic data are available for commensal strains, especially in wild animals, and environmental strains. Here, we analysed the genomes of a large collection of *E. coli* strains collected across many human, domestic and wild animal and environmental sources in different geographic locations from the Australian continent. This collection is dominated by non-clinical isolates, corresponding to the main habitats of the species. We sought to understand the dynamics of the evolution of gene repertoires and how it was driven by mobile genetic elements. The analysis of the isolation sources in the light of phylogenetic structure and genome variation suggests that rates and mechanisms of adaptation vary with the habitat and the phylogenomic background. This contributes to explain why known epidemiological clones of the species emerge from specific phylogenetic groups, even though virulence strongly depends on the acquisition of virulence factors by horizontal gene transfer.

Results

The large and little known pan-genome of *E. coli*

We sequenced and annotated the genomes of 1,294 *E. coli sensu stricto* strains selected from more than 3,300 non-human vertebrate hosts, 1,000 humans and 800 environmental samples between 1993 and 2015, chosen to represent the phylogenetic diversity of the species (Materials and Methods, Fig 1A, S1 Text). All samples were collected by a single team, spanning a 20 year-period, from different regions in a single isolated continent (Australia). The origin of each strain was accurately characterized and the genomes were uniformly annotated and analyzed using the same bioinformatics processes. The strains were isolated from humans, domesticated and wild animals, representing the primary habitat of *E. coli*, and from freshwater, representing its secondary habitat [3]. Less than 22% of the samples were recovered from clinical situations. A series of controls confirmed that the sequences were of high quality and contained the known essential genes (S2 Text). The genomes varied widely in size from 4.2 to 6.0 Mb (average 5 Mb), but had similar densities of protein-coding sequences (~87%) and GC content (50.6%, S1 Fig and S1 Table).

The pan-genome contains 75,890 gene families, which is over 16 times the average genome size. The core genome is very small (295 genes), a feature typical of comparisons involving many genomes. As a result, we have opted to focus, whenever possible, on the persistent genome. This corresponds to gene families present in at least 99% of the genomes of the sample. This provides some flexibility to account for sequencing or assembling artifacts and to account for the odd genome that may have recently lost a few core genes. The pan-genome families were classified as part of the *persistent* genome (3%), *singletons* (44%, present in a single genome), or *accessory* genome (the remaining) (Fig 1B, S2 Fig). The persistent gene families are a tiny fraction of the pan-genome, but account for half of the average genome (Fig 1B). They were used to build a robust phylogeny of the species (S3 Fig), which was rooted using genomes from other species in the genus (S4 Fig). In contrast, singletons are almost half of the gene families of the pan-genome, but less than 1% of the average genome. As a consequence,

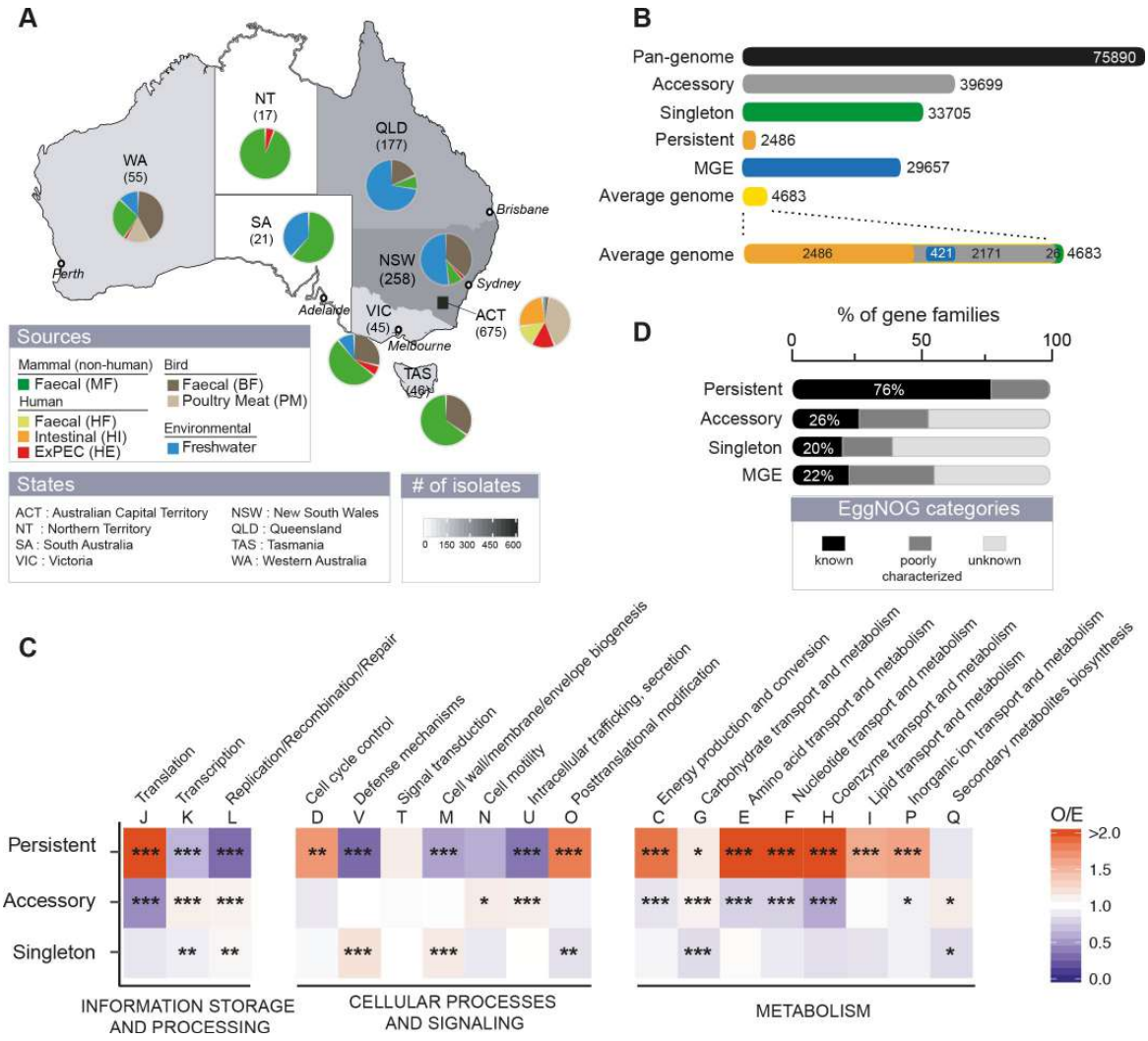


Fig 1. The genetic diversity of Australian *E. coli*. A. Distribution of isolates per region and per source. **B.** The pan-genome is composed of 75,890 gene families, of which 33,705 are singletons (in green, present in a single genome), 2,486 persistent (in gold, present in at least 99% of genomes), the remaining being accessory (in grey). 29,657 gene families (39% of the pan-genome) were related to mobile genetic elements (MGE). **C.** Functional EggNOG categories of pan-genome gene families. The ratio observed/expected (O/E) for the frequency of non-supervised orthologous groups (NOGs, shown as capitalized letters) is reported for all comparisons with a color code ranging from blue (under-representation) to red (over-representation). The level of significance of each Fisher's exact test was indicated ($P > = 0.05$: ns; $P<0.05$: *; $P<0.01$: **; $P<0.001$:***). It was performed on each 2*2 contingency table. Gene families lacking matches to the EggNOG functional categories were discarded. **D.** Percentage of the different EggNOG categories (see insert) in the persistent, accessory and singleton gene families and among genes associated to MGE.

https://doi.org/10.1371/journal.pgen.1008866.g001

the pan-genome is open, as measured by the fit to a Heaps' law model [46], and increases on average by ~26 protein coding genes with the inclusion of a new genome (S2 Fig). Singletons are smaller than the other genes and tend to be located at the edge of contigs (44%). Hence, some of these singletons may result from sequencing and assembly artifacts (S3 Text and S5 Fig). When all the singletons were excluded, the pan-genome still remained open (S2 Fig).

To obtain a better understanding of the functional classifications of genes in the pan-genome, we annotated them using the EggNOG categories (Fig 1C). As expected, the persistent genome over-represented typical housekeeping functions, whereas the accessory genome over-represented cell motility, intracellular trafficking and secretion, carbohydrate transport and metabolism and secondary metabolism. Singletons over-represented defense systems and genes related with the cell envelope. Most singletons (80%) and accessory (74%) gene families, but also a surprisingly high number of persistent gene families (24%), lacked a clear functional assignment as given by the EggNOG database [47] (Fig 1D). Hence, we are still ignorant of the function, or even the existence, of many genes of the species.

Very rapid initial divergence of gene repertoires becomes linear with time

Traditional epidemiological studies of *E. coli* focused on multilocus sequence types (ST) and/or the O-serogroups and H-types (the O:H combination corresponding to the serotype). These epidemiological units regroup strains in terms of sequence similarity in a few persistent genes (ST) or in key traits related to the cell envelope (the LPS structure for the O-group and the flagellum for the H-type). However, it is unclear if these types systematically regroup strains with similar gene repertoires. We identified 442 distinct STs, of which 61% are represented by a single strain. A few STs are very abundant in our dataset: 20 include more than 10 genomes each and encompass 40% of the dataset. STs are usually regarded as very recently diverged strains. Indeed, the intra-ST genetic distances are 10-times smaller those between the other pairs of genomes (0.003 vs. 0.03, Fig 2A). Yet, 6% of intra-ST comparisons have more than 0.01 substitutions per position showing extensive genetic diversity at the genome level (Fig 2B). Some O-groups are abundant, e.g., O8, O2 and O1 (each present in >50 genomes) but almost half of the groups occur in a single genome and 43% of the strains could not be assigned an O-group (even when the *wzm*/*wzt* and *wzx*/*wzy* genes were present). In contrast, most H-types were previously known (87%). We found 311 O:H serotypes among the 726 typeable genomes. Of these, 64% are present in only one genome, 17% are in multiple STs and 7% in multiple phylogroups (e.g. O8:H10). Conversely, half of the 95 STs with more than one genome have multiple O:H combinations, e.g. ST10 has 24. These results confirm that surface antigens and their combinations change quickly and are homoplasic. They also show significant sequence divergence in persistent genes within STs.

We then aimed at assessing if genomes within STs also show extensive variation of gene repertoires. For this, we computed the gene repertoire relatedness (GRR) between genomes (see Methods). Genes from the same gene family are on average 98.3% identical (S2 Fig). Since the threshold to be part of the family is 80% identity, rapid sequence evolution will very rarely lead a gene to be classed apart from its orthologs. As a result, variations in GRR result from gain and loss of genes, not sequence divergence. The GRR values decrease very rapidly with patristic distance (the sum of branch lengths in the path between two genomes in the phylogenetic tree) for closely related strains, as revealed by spline fits (Fig 2A). Similar results were observed when removing singletons, which only account for on average 0.5% of the genes in genomes, suggesting that this result is not due to annotation or sequencing errors (S6 Fig). As a consequence, 85% of the intra-ST comparisons have a GRR lower than 95% (corresponding to ~235 gene differences per genome pair), and some as little as 77% (Fig 2C). These results reveal that even genomes of the same ST can differ substantially in terms of their gene repertoires.

To check if the dataset is representative of the species and can be used to assess its diversity, we compared it with the ECOR collection [48] and the complete genomes available in RefSeq (Materials). All datasets had similar nucleotide diversity (S7A Fig and S1 Table). Using rarefied

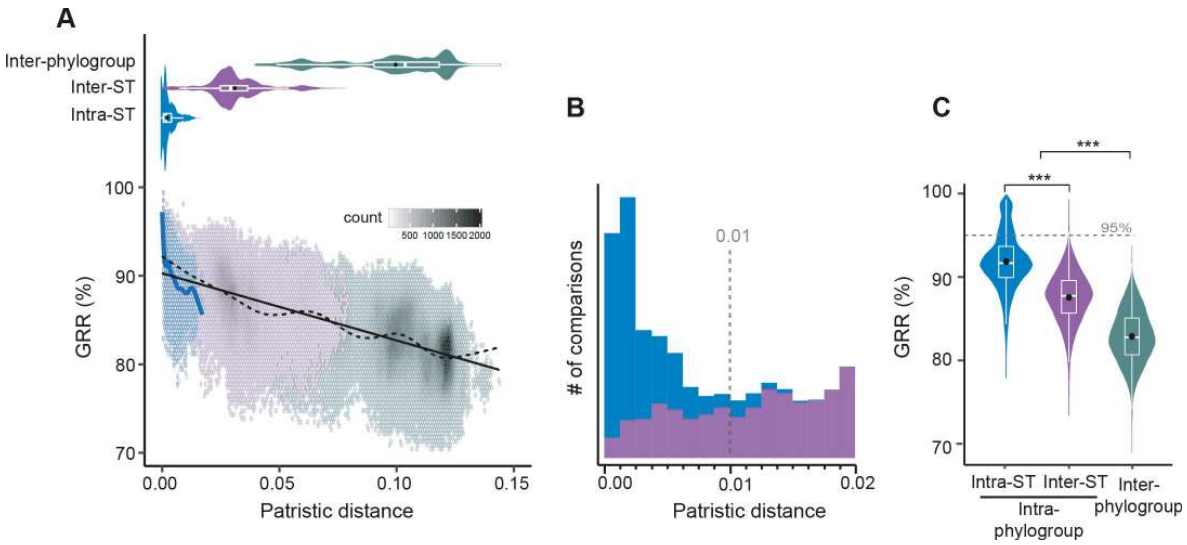


Fig 2. Evolution of Gene Repertoire Relatedness (GRR) with time. A. [Top] Violin plots of the patristic distance computed between pairs of intra-ST (in blue), inter-ST (in purple), and inter-phylogroup (in water green) genomes. [Bottom] Association between GRR and the patristic distance across pairs of genomes. Due to the large number of comparisons (points), we divided the plot area in regular hexagons. Color intensity is proportional to the number of cases (count) in each hexagon. The linear fit (black solid line, linear model (lm)) was computed for the entire dataset (1,294 genomes, $Y = 90.2–75.7^{*}X$, $R^2 = 0.49$, $P<10^{-4}$). The spline fit (generalized additive model (gam)) was computed for the whole (in black dashed line) or the intra-ST (in blue solid line) comparisons. There was a significant negative correlation between GRR and the patristic distance (Spearman's rho = -0.67, $P<10^{-4}$). **B.** Stacked bar plot of the number of intra-ST (in blue) and inter-ST (in purple) comparisons at short evolutionary scales. **C.** Violin plots of the intra-ST, inter-ST and inter-phylogroup GRR (%). (A-B-C*)* All the distributions were significantly different (Wilcoxon test, $P<10^{-4}$), the same color code was used and described in panel A.

https://doi.org/10.1371/journal.pgen.1008866.g002

datasets, to compare sets of same size, ours had the largest pan-genome, partly because of a larger number of singletons (S7 Fig). Our dataset also had the highest α-diversity for the three typing schemes (STs, O-groups, H-types, S1 Table). Since the gene repertoire diversity of *E. coli* in Australia is at least as high as that of ECOR and RefSeq, we studied the variation in gene repertoires beyond the intra-ST level. After the rapid initial drop in GRR described above, the values of this variable decrease linearly with phylogenetic distances (Fig 2A). The average values of GRR given by the regression vary between 90% for very close genomes and 80% for the most distant ones. The variance around the regression line is constant and a spline fit shows few deviations around the regression line. This is consistent with a model where initial divergence in gene repertoires is driven by rapid turnover of novel genes. After this initial process, divergence in gene repertoires increases linearly with patristic distance.

Rates of gene repertoire diversification vary across phylogroups

We used the species phylogeny to study the associations between phylogroups and genetic diversity (Fig 3A). The tree showed seven main phylogenetic groups very clearly separated by nodes with 100% bootstrap support. The 17 phylogroup C strains were all included within the B1 phylogroup and were thus grouped with the latter in this study. For the rest, the analysis showed a good correspondence between the assignment into the known phylogroups—A, B1, B2, D, E, F, and G–and the different clades of the species tree. The tree splits the species initially in a clade with phylogroup B2, F and G on one side and the remaining on the other side. In line with the literature [40], four major phylogroups were very abundant—A (24% of the dataset), B1 (24%), B2 (25%) and D (14%)–whereas the others were rarer. The nucleotide

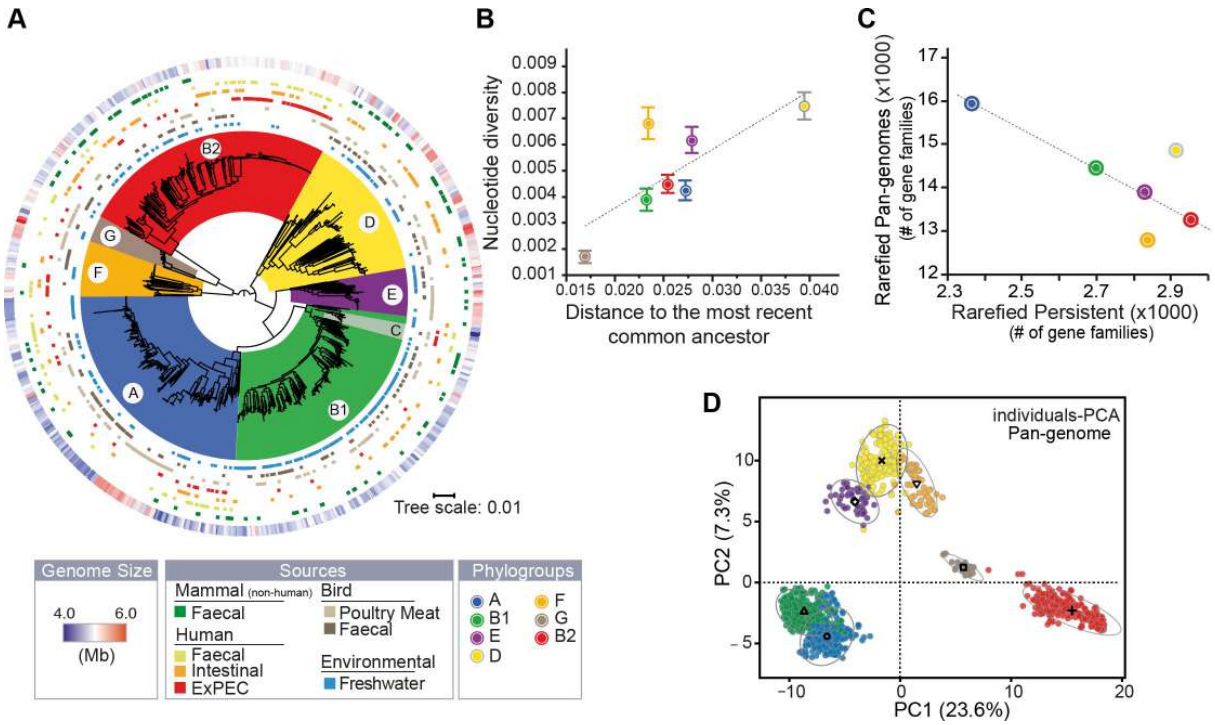


Fig 3. The genetic and ecological structure of Australian *E. coli* population. A. Phylogenetic tree of *E. coli* rooted using the genomes of other *Escherichia* (only shown in S4 Fig for clarity). From the inside to the outside: the 7 main phylogroups (arcs covering the tree), the source of each genome (seven rows), and the size of the genomes (outer row, see insert legend). **B.** Association between the nucleotide diversity per site (Pi, average and s.e) within phylogroup and their distance to their most recent common ancestor (MRCA). In each pylogroup, we averaged the nucleotide diversity (π) obtained for 112 core-genes, and the length branches (from tip-to-MRCA) of the species tree. **C.** Association between the rarefied pan- and persistent-genomes in each phylogroup. We used 1,000 permutations (genomes orderings) of 50 randomly selected genomes (rarefied datasets) to compute the pan- and the persistent-genomes in each phylogroup (ignoring the G group), and then averaged the results. **D.** Principal component analysis of the pan-genome (matrix of presence/absence of each gene family across genomes). Each dot corresponds to a genome in the two first principal components (PC). The ellipse (90%) and barycenter of each phylogroup are reported. The percentages in the axis labels correspond to the fraction of variation explained by the PC. All panels follow the color code of A.

https://doi.org/10.1371/journal.pgen.1008866.g003

diversity of the phylogroups is very dependent on their phylogenetic structure, since some clades have more closely related clusters of strains than others (S8 Fig). Nevertheless, nucleotide diversity, patristic distances, and Mash distances revealed similar trends: the phylogroup D exhibited the highest genetic diversity, followed by F, E, and then by the most abundant groups–A, B1 and B2 –which all have similar levels of diversity (S8 Fig). The phylogroup G was the least diverse, but it is also poorly represented in our dataset (33 genomes from three STs). Overall, genetic diversity is proportional to the depth of the phylogroup, i.e. the average tip-to-MRCA distance, except for phylogroup F which is more diverse than expected (Fig 3B). These results suggest that genetic diversity varies between phylogroups and that within phylogroups it is strongly affected by the time of divergence since the most recent common ancestor.

The sets of genomes of each phylogroup have large and open pan-genomes (S9 Fig and S2 Table). The sizes of these pan-genomes differ widely across phylogroups and are partly correlated to the number of genomes in the phylogroup, explaining why the phylogroup G has the smallest pan-genome (S9 Fig). To control for the effect of sample size, we computed pan-

genomes from 1,000 random samples of 50 genomes for each phylogroup (ignoring the few strains of the G phylogroup, Fig 3C and S2 Table). This revealed larger pan-genomes for phylogroups A, D, and B1 followed by E, B2 and F. Intriguingly, the larger the pan-genome of a phylogroup, the smaller the fraction of its genes that are part of the persistent genome (Fig 3C). This suggests that differences of pan-genome sizes across phylogroups are caused by different rates of gene turnover, which seems to affect, at different extent, both genes present in most strains and genes present in very few.

To quantify the similarities in gene repertoires, we analyzed the GRR values between phylogroups. The smallest values were observed when comparing B2 strains with the rest (S10A Fig). Accordingly, a principal component analysis (PCA) of the presence/absence matrix of the pan-genome shows a first axis (accounting for 23.6% of the variance) clearly separating the B2 from the other phylogroups (Fig 3D). This shows that gene repertoires of B2 strains are the most distinct from the other major phylogroups. The large phylogroups A and B1 are very close in the GRR and in the PCA analyses, showing high similarity in terms of gene repertoires. Interestingly, the phylogroups D and F, which are not close in the species tree, cluster together in terms of gene repertoires. This may explain the conflicting results of our phylogenetic analysis, which places with high confidence the phylogroup D in the same partition of A and B1, and works based on ancestral gene repertoires that place them as a basal group in the tree (not far from F and G) [49]. Hence, phylogroups differ in terms of their gene repertoires and in their rates of genetic diversification, but while some are quite similar (A and B1), others (B2) stand aside from the remaining phylogroups.

Mobile genetic elements drive rapid initial turnover of gene repertoires

Different mechanisms can drive the rapid initial diversification of gene repertoires. Mobile genetic elements encoding the mechanisms for transmission between genomes (using virions or conjugation) or within genomes (insertion sequences, integron cassettes) are known to transfer at high rates and be rapidly lost [50–52]. We detected prophages using VirSorter [53], plasmids using PlaScope [54], and conjugative systems using ConjScan [55] (S11–S13 Figs). These analyses have the caveat that some mobile elements may be split in different contigs, resulting in missed and/or artificially split elements. This is more frequent in the case of plasmids, since they tend to have many repeated elements [56]. Only two genomes lacked identifiable prophages and only 9% lacked plasmid contigs. We identified 929 conjugative systems, with some genomes containing up to seven, most often of type MPF_F, the type present in the F plasmid. On average, prophages accounted for 5% and plasmids for 3% of the genomes (Fig 4A). Together they account for more than a third of the pan-genomes of each phylogroup. We also searched for elements capable of mobilizing genes within genomes: Insertion Sequences, with ISfinder [57], and Integrons, with IntegronFinder [58]. Even if ISs are often lost during sequence assembly, some genomes had up to 152 identifiable ISs representing ~1% of the genome (Fig 4A and S13 Fig). A fourth of the ISs were in plasmids and very few were within prophages. We found integron integrases in 14% of the genomes, usually in a single copy. It is interesting to note that even if the frequency of each type of MGE varies across strains, each of them is strongly correlated with the frequency of the other elements (Fig 4B). Hence, the typical *E. coli* genome has at least one transposable element, a prophage and a plasmid, the key tools to move genes between and within genomes. This means that when genomes are enriched in one type of MGE, they tend to get simultaneously enriched in the remaining types of MGEs.

What is the effect of these MGEs in the dynamics of *E. coli* genomes? First, none of the MGEs gene families is present in more than 99% of the strains (i.e. none qualifies as persistent

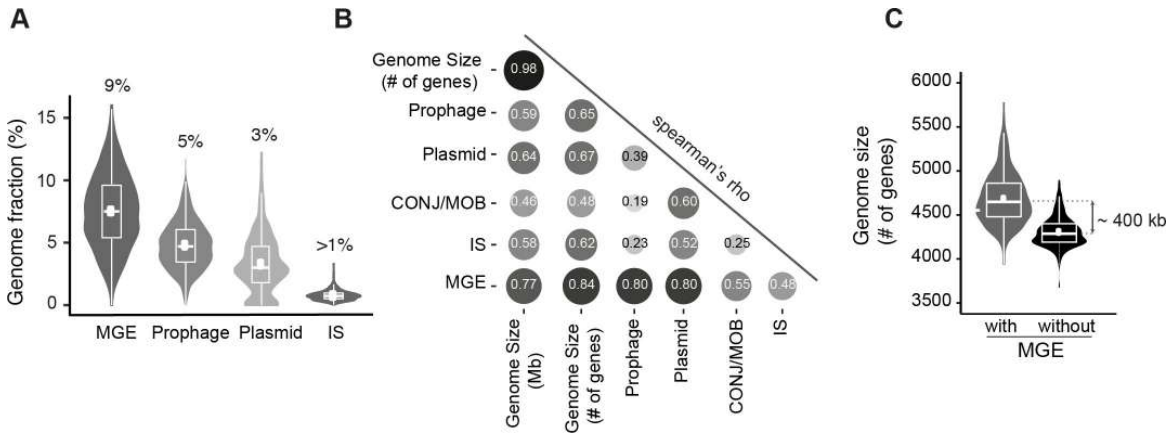


Fig 4. Frequency of mobile genetic elements (MGEs). A. Percentage of genes associated with MGEs per genome (sum in first graph). **B.** Spearman's rank correlation matrix between the number of genes related to MGE and the genome size (in Mb and number of genes). The shades of the grayscale and the size of the circle are proportional to the correlation coefficients. All values are significantly positive ($P<10^{-4}$). **C.** Differences in genome size when MGE genes are included or removed.

https://doi.org/10.1371/journal.pgen.1008866.g004

genes) at the species or at the phylogroup level. Instead, they are systematically at low frequency in the pan-genome, even at the phylogroup level. Hence, these genes rarely rise to high frequency in the species. Second, when we inferred the events of gene gain and loss in the species tree using Count (see Methods), we found that half of the recent gene acquisitions, i.e., those that took place at the level of the terminal branches of the species tree, were MGE genes. Conversely, the acquisitions at the terminal branches correspond to 40% of the MGE genes of the species. Third, the acquisition of MGEs affects the size of the genome. Those identified in this study account for ~8% of the genome size (Fig 4C and S14 Fig), and the number of genes associated with MGEs is strongly correlated with genome size for every type of element (Fig 4B). Fourth, MGEs increase the variability of genome sizes, since removing them decreases the coefficient of variation of the size of gene repertoires by 34% (expected increase of 4% under a Poisson model, Fig 4C). Fifth, the increase in variance in terms of genome size caused by MGEs is amplified by their rapid loss after acquisition (short persistence times in the genome). No MGE-associated gene family is sufficiently frequent to be part of the persistent genome, and most (85%) are present in less than 1% of the genomes. For example, 41% of the IS gene families are singletons (S14 Fig).

These results are consistent with the analysis of the variation in GRR with patristic distance, where some genes have extremely rapid turnover. Here we show that many of them are MGEs. The lack of fixation of MGE-associated genes suggests that the long-term cost of MGEs themselves is significant and/or their contribution to fitness is low (or temporary). But even if most genes associated with MGEs are eventually lost, their cargo genes may be adaptive, remain in the genomes for long periods and eventually become fixed. In conclusion, MGEs have a key role in the initial rapid turnover of genes in genomes because they are aquired at high rates, even if most of their genes are eventually lost.

The smallest genomes have the highest gene turnover

Is the distribution of specific MGEs and their rates of transfer strongly associated with specific traits of genomes, like their phylogroup or isolation source? And if so, is this leading to

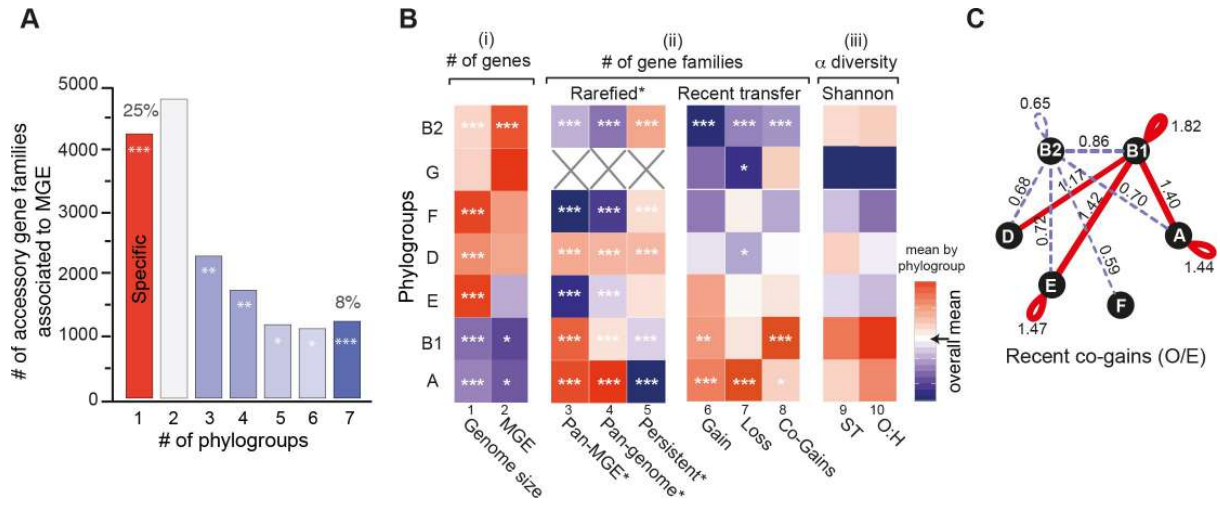


Fig 5. Genetic diversification across phylogroups. A. Number of accessory gene families associated to MGE present in one (i.e., phylogroup-specific) to seven phylogroups. The color code used corresponds to the Z-score obtained for the observed number (O) with respect to the expected distribution (E) (see Methods) for each case with a color code ranging from blue (under-representation) to red (over-representation). The level of significance was reported: |Z-score|: * ([1.96–2.58[), ** ([2.58–3.29[, ***([3.29). **B.** Heatmap where a cell represents the deviation (the difference) of the phylogroup to the rest. All values were standardized by column. The color code ranging from blue (lower) to red (higher), with white (overall mean). The level of significance of each ANOM test was reported: * (P<0.05), ** (P<0.01), *** (P<0.001). **C.** Network of recent co-occurence of gains (co-gains) of accessory genes within and between phylogroups. Nodes are phylogroups and edges the O/E ratio of the number of pairs of accessory genes (from the same gene family) acquired in the terminal branches of the tree. Only significant O/E values (and edges) are plotted (|Z-score|>1.96). Under-represented values are in dash blue and over-represented in red (see Methods).

https://doi.org/10.1371/journal.pgen.1008866.g005

preferential paths of gene transfer within the species? It has been suggested that homologous recombination is much rarer between than within phylogroups [18]. To test if this applies to the transfer of MGEs, we analyzed the distribution of the pan-genome gene families that are part of MGEs (excluding singletons, for the separate analysis of prophages and plasmids, see S15 Fig). There is a small but significant tendency of gene families of MGEs to cluster in a single phylogroup (Z-score>20, see Methods). However, 75% of the phage and plasmid gene families were found in more than one phylogroup and 8% were found in all phylogroups (Fig 5A). Hence, MGEs are key players in genome diversification at the micro-evolutionary scale. Above we showed that they were acquired independently multiple times and most of them have just arrived in their host genome. We now show that they are often transferred across phylogroups.

One might expect more genetic diversity in phylogroups with more MGEs and larger genomes. In apparent agreement with this hypothesis, genomes from phylogroups A and B1 are significantly smaller than the others (Fig 5B, col 1, ANOM tests, $P<10^{-3}$) and have fewer MGE-associated genes (Fig 5B, col 2, ANOM tests, P<0.05). However, these phylogroups also have the largest diversity of gene families associated to MGEs (Fig 5B, col 3, in both the full and rarefied datasets, both ANOM tests, $P<10^{-3}$), *i.e.* they encode fewer but more diverse MGEs. Furthermore, the phylogroups A and B1, in spite of having among the most recent common ancestors of the phylogroups (Fig 3B), have the largest pan-genomes, the smallest persistent genomes, and the largest diversity of STs, and serotypes (Fig 5B, in both the full and rarefied datasets, cols 4,5,9,10, ANOM tests, $P<10^{-3}$). This intriguing pattern suggests that the smallest genomes have the highest turnover of genes, not the lower rates of transfer. To test this hypothesis, we took the quantification of gene gains and losses at the terminal branches of

the species tree, computed with Count (see above), and computed the number of these events per phylogroup. We found that phylogroups A and B1 have the highest number of gene gains and losses per terminal branch (Fig 5B, cols 6–7). Hence, these phylogroups have the smallest genomes but the most frequent events of gene gain and loss.

To study recent gene flow between different phylogroups, we took the genes inferred to be acquired in the terminal branches of the species trees. Among these recently acquired genes we selected the gene pairs from the same gene family (co-gains) that were from the same phylogroup (Fig 5B, col 8) and those corresponding to recent acquisition of the same gene family in two different phylogroups (see Methods, Fig 5C). The results were represented as a graph where the edges represent significantly fewer (dashed lines) or higher (solid lines) number of co-gains than expected by chance. We found that phylogroup B1 has significantly more co-gains of genes with other phylogroups than expected, while the inverse was observed for phylogroup B2. We reached similar results when considering only the co-gains associated with MGEs (S16 Fig). These results are consistent with the separation of the B2 phylogroup from the others in the PCA analysis (Fig 3D). They show that such separation is due to lower rates of transfer in B2, which leads to fewer co-gains within the phylogroup and between this and the other phylogroups. In summary, phylogroups differ in terms of their genome size and in their rates of genetic diversification, the two traits being inversely correlated within the species.

Not everything is abundant everywhere: the interplay between phylogroups and sources in genetic diversification

Frequent horizontal transfer across phylogroups could result in adaptation being independent of the strain genetic background. While we observed that strains from all phylogroups could be isolated in all different sources (Fig 6A), different phylogroups are typically over-represented in some sources and rare in others (Fig 6B). These observations match previous studies [3], and show an association between the phylogenetic structure of populations and the natural habitats of the strains.

How much of the variability in genome size is explained by the source of isolation of the strains? Genome sizes vary significantly across isolation sources. Strains isolated from poultry meat had the largest average genomes, followed by human ExPEC strains. In contrast, strains from wild birds' feces and freshwater had the smallest genomes (Fig 3A and Fig 6C, col 1, ANOM tests, $P<10^{-3}$). We showed above that genome size also varies across phylogroups. To understand the relative role of the two variables, isolation source and phylogroup, we made two complementary analyses. First, we compared the genome size of strains from different sources within each phylogroup. Even if the statistical power was sometimes low, this revealed trends similar to the ones observed across phylogroups (S17 Fig). Second, we used stepwise multiple regressions to assess the effects of phylogroup and the strains' source on its genome size. Both variables contributed significantly, and in almost equal parts, to the statistical model and together explained 36% of the variance ($R^2 = 0.36$; $P<10^{-4}$, S3 Table). We found similar results after removing MGE-associated genes (Fig 6D and S4 Table). We conclude that both isolation source and phylogroup are equally associated with genome size.

Adaptation to a habitat depends on HGT, which is driven by MGEs. This led us to study the distribution of MGEs in relation to isolation sources. There are fewer MGE genes in strains isolated from freshwater and wild birds' feces, which have smaller genome sizes, and more in strains from human ExPEC and poultry meat (Fig 6C, col 2, ANOM tests, $P<10^{-3}$, and S5 Table). We observed similar trends within each phylogroup even if the statistical power was low (S17 Fig). The analysis of the relative contribution of phylogroups and isolation sources to

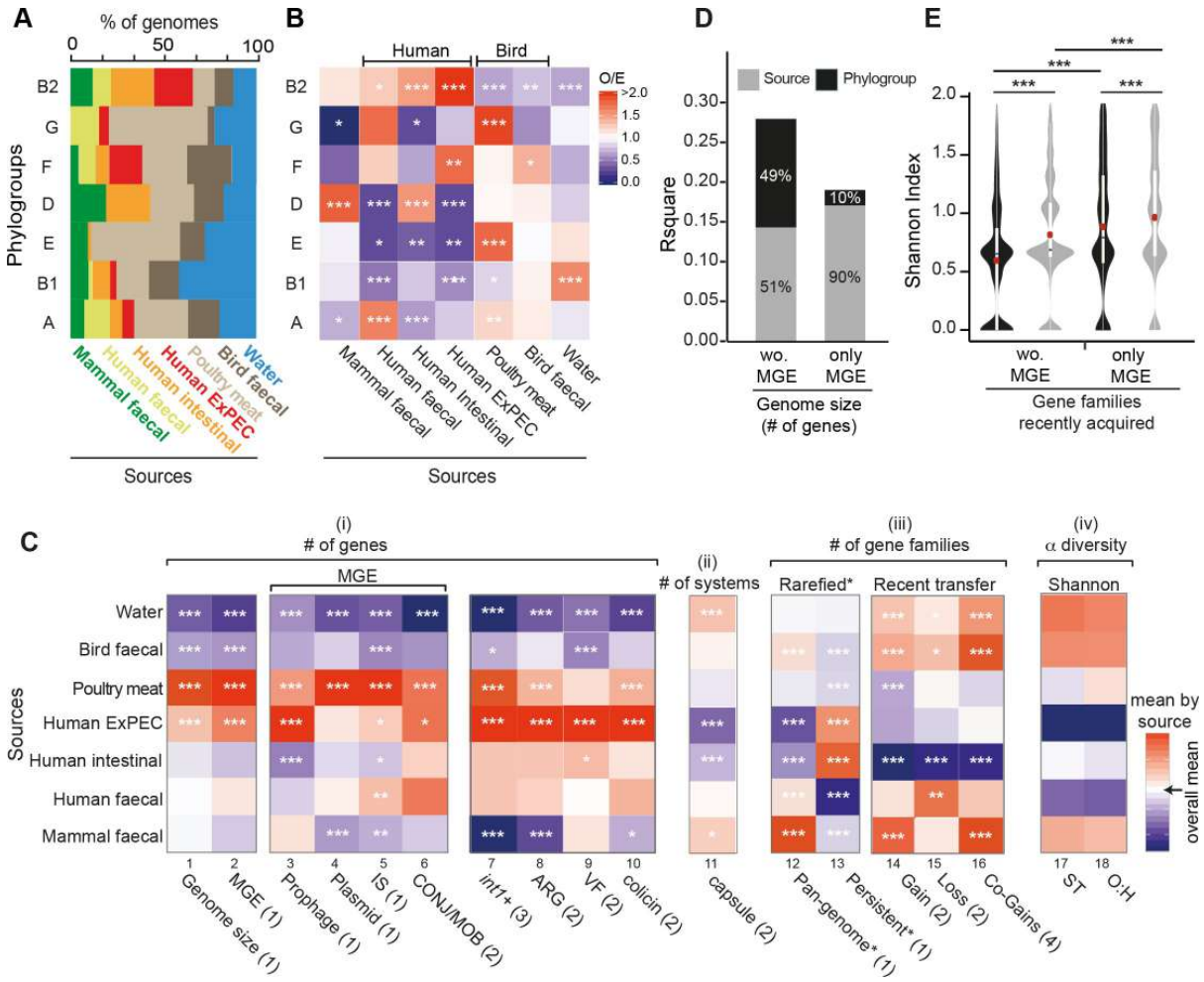


Fig 6. Genetic diversification across sources. A. Distributions of the sources in each phylogroup. **B.** Association between phylogroups and sources. The ratio of the number of observed (O) genomes divided by the expected (E) number was reported for all comparisons with a color code ranging from blue (under-representation) to red (over-representation) (Fisher's exact tests performed on each 2*2 contingency table). **C.** Heatmap showing the associations between isolation sources and a number of traits. Each cell indicates the deviation (the difference) to the overall mean (in white). All values were standardized by column. Tests: standard ANOM (1), non-parametric ANOM tests (2, in presence of deviations from Gaussian distributions), ANOM for proportions (3). We represented the (O/E) ratio of the co-occurrence of gene pairs recently acquired (Co-gains) in each phylogroup with the same color code as in panel B (4). **D.** Contribution of each variable (phylogoup and source) to the variance explained by the stepwise multiple regressions of genome size (for the component of MGEs or the remaining genome) on phylogroup and the isolation source. **E.** Differences in diversity of gene families recently acquired across phylogroups (in black) and sources (in grey) for gene families associated to MGE or the remaining gene families (Wilcoxon tests, red dots (means)). In all panels: the level of significance of each test was reported: * ($P<0.05$), ** ($P<0.01$), *** ($P<0.001$).

https://doi.org/10.1371/journal.pgen.1008866.g006

the number of MGE genes showed that the source of the strain accounted for the vast majority of the explained variance (90%, full model: $R^2 = 0.19$; $P<10^{-4}$, Fig 6D and S6 Table). Accordingly, the number of MGE gene families present in a single source of isolation was higher than expected (Z-score >17, S15 Fig), and nearly one third of these were observed in multiple phylogroups. To quantify this trend, we counted recent independent gains (co-gains, see definition above) of the same gene family (see Methods). This was done for pairs of genomes within the

same source and between different sources. The analysis revealed that co-gains were more frequent than expected within the same isolation source. (Fig 6C, col 16, see Methods). These results suggest that the contribution of MGEs to genome size is primarily driven by the source of the isolate rather than phylogroup membership.

The previous result could arise from preferential co-gains of MGEs in an isolation source relative to a phylogroup, i.e. to frequent transfer of a few MGEs in the multiple isolates from the same type of source. To test this hypothesis, we used the results from Count and built a matrix where for each gene family we indicate the acquisition or not of a gene in each of the terminal branches of the phylogenetic tree. We then compared the clustering of these recent acquisitions by phylogroup and by isolation source using Shannon indexes (see Methods). If the hypothesis is correct, we expected higher clustering (lower diversity) across sources than across phylogroups. We observed slightly higher clustering across phylogroups than across sources, both for MGE and for the other genes (Fig 6E). We conclude that the contribution of MGEs to genome size depends largely on the isolation source but that this does not reflect systematic gains of the same MGE genes in the same source. Instead, the higher frequency of MGEs in genomes of certain sources may result from higher density of MGEs in those habitats (higher infection rates), or from higher probability of acquiring MGEs with adaptive traits at certain sources (higher selection rates).

It is tempting to speculate that the association between the number of MGE genes and isolation sources reflects selection for the acquisition of locally adaptive functions that are transferred by these MGEs. To test this, we searched for the presence of a trait—antibiotic resistance–that has become adaptive only recently and that is frequently transferred by MGEs. We searched for antibiotic resistance genes (ARGs) in our dataset using the reference databases. Many of these ARGs were in integrons (~3 per integron), which is well documented [59], and genomes carrying integrons had more ARGs than the others (Wilcoxon test, $P<10^{-4}$, S18 Fig). Expectedly, integrons and ARGs were more prevalent in human ExPEC and in poultry meat isolates (Fig 6C, cols 7–8) and S5 Table). Similar results were observed in the analyses at the level of each phylogroup (S18 Fig). The clear association of integrons and ARGs with human (or domesticated animals) isolates of *E. coli* independently of the phylogroups' genetic background reinforces the idea that source-specific MGEs provide locally adaptive traits.

Functional differences across phylogroups and isolation sources

Several of the previous results suggest an accumulation of adaptive genes as patristic distances increase. We used a gene-based GWAS to search for functions enriched in phylogroups or in isolation sources (see Methods). The first analysis revealed many gene families (2,754, S2 Dataset) positively and negatively associated with the phylogroups (Fig 7A). While in most cases these associations link a gene family to a phylogroup, the phylogroup A and B1, which are close in the phylogeny (Fig 3A) and in terms of gene repertoires (Fig 3D), have many associations in common (53%). The phylogroup with the largest number of associated genes is B2, which is also in accordance with the PCA analysis that revealed distinct gene repertoires in this phylogroup (Fig 3D). We characterized the functional categories of these associated gene families using EggNOG classification (as previously, S2 Dataset). In general, the categories over-represented are related to genes involved in metabolism (Fig 7B), which is in agreement with previous studies [60, 61].

The genes that were identified in the GWAS often concerned degradation processes, notably aromatic compound degradation (S2 Dataset) [62]. For example, PP (phenylpropionic acid) and HPP/HCI (hydroxyphenylpropionic and hydroxycinnamic acid) degradation

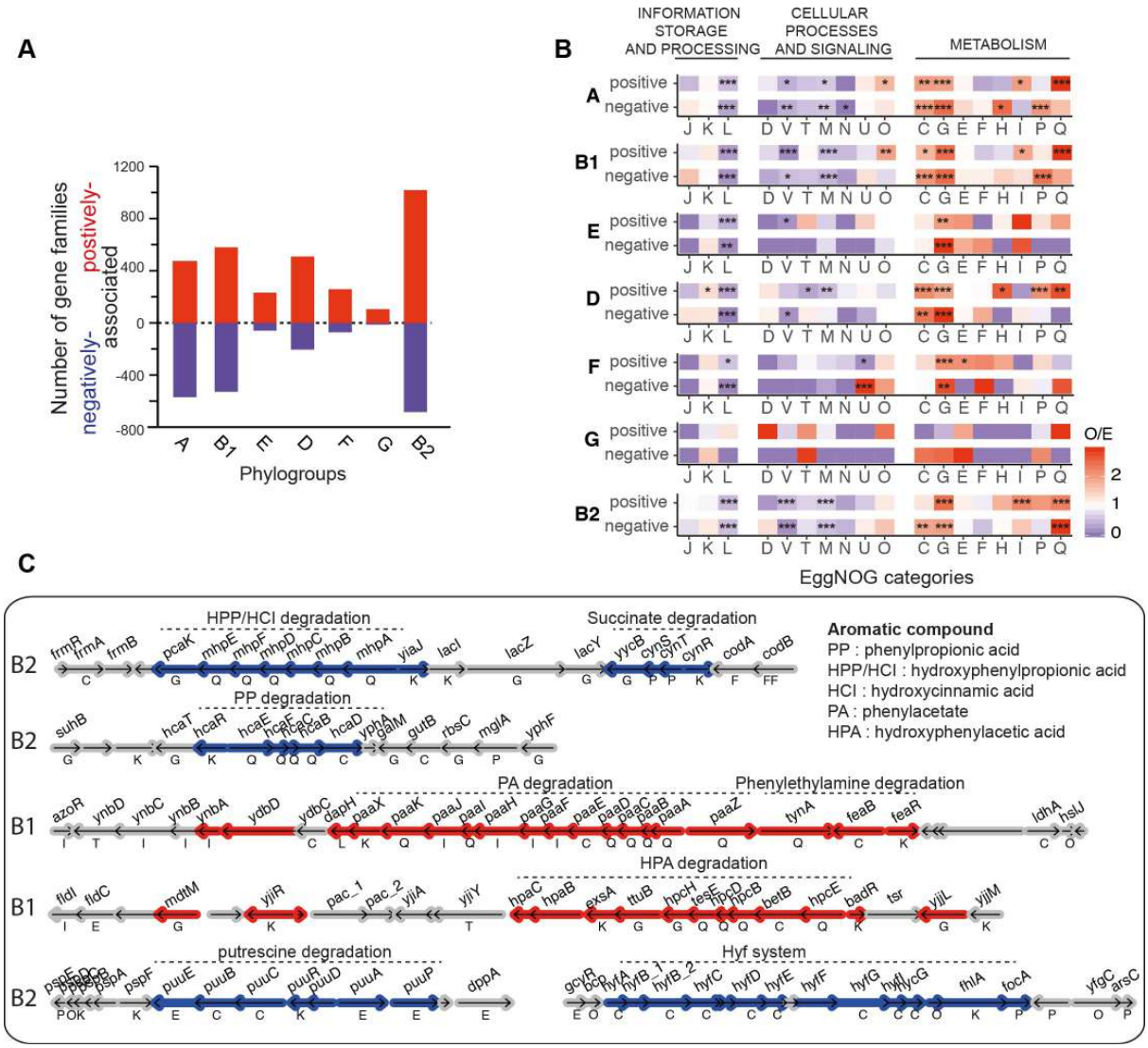


Fig 7. Genetic determinants of each phylogroup. A. Number of gene families positively (in red) and negatively (in blue) associated with each phylogroup. Altogether, they represent 7% of the accessory gene families of the dataset (note that some gene families are associated with several phylogroups). **B**. Observed/expected (O/E) ratios of non-supervised orthologous groups (NOGs, shown as capitalized letters, same code as shown in Fig 1C) in the positively or negatively associated gene families. For example, in phylogroup A there is an over-represents of positive associations in class Q, whereas in class L for the same phylogroup A there is under-represention for both positive and negative associations. The ratio (O/E) was reported for all comparisons with a color code ranging from blue (under-representation) to red (over-representation). The level of significance of each Fisher's exact test was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). It was performed on each 2*2 contingency table. Gene families lacking matches to the EggNOG functional categories (57%) were discarded. **C.** Genomic organization of some regions enriched in genes positively (in red) or negatively (in blue) associated with a phylogroup (indicated on the left). Genes shown in grey are not significantly associated. The name of the gene (when available) is shown above it, its EggNOG functional category (when known) below it.

https://doi.org/10.1371/journal.pgen.1008866.g007

pathways are negatively associated with B2 strains, while PA (phenylacetate acid) and HPA (hydroxyphenylacetic acid) degradation are positively associated with B1 strains (S2 Dataset, Fig 7C). These results are consistent with recent phenotypic tests (growth on specific substrates) [61]. Interestingly, B1 strains are positively associated with genes involved in rhamnose, sucrose, xylose, glycerate, and tartrate degradation pathways, while B2 are negatively associated with traits associated with plant colonization such as the Hyf system (involved in control and pH control), melibiose, cyanate, putrescine, and D-malate degradation pathways (S2 Dataset, Fig 7C). These pathways are involved in alternate carbon source metabolism, and may reflect functional adaptations to different nutritional environments, as proposed previously [63]. These results suggest that B1 strains, contrary to B2, tend to carry traits facilitating adaptation to environmental niches, such as soil and water (where aromatic compounds are highly abundant) or to colonize plants, as previously suggested [64].

The same analysis made at the level of the isolation sources revealed fewer genes (Fig 8A). The different fecal isolates almost lacked associated genes, presumably because this is the most typical and the ancestral environment of the species and it may have adapted to it for a long time. We therefore focused our analysis on genes involved in virulence. The analysis of human ExPEC isolates revealed many associated genes (S2 Dataset), including well-known virulence factors such as ABC-dependent capsule systems, the motility repressor *papX*, the P fimbriae, yersiniabactin, colibactin and multiple type 5a protein secretion systems (Fig 8C). To complement this analysis, we searched specifically for known virulence factors from VFDB [65]. Indeed, they are more prevalent in human strains, and especially in ExPEC isolates (ANOM test, $P<10^{-3}$), while being rare in strains isolated from freshwater and wild birds' feces (ANOM test, $P<10^{-3}$, Fig 6C, col 9). While these virulence factors are more concentrated in phylogroups B2, D, E and F (ANOM test, $P<10^{-2}$) as previously shown [37], the trends regarding isolation sources are conserved within each phylogroup (S19 Fig). In particular, within phylogroup B2, only human strains have a significantly higher average number of virulence factors (S19 Fig) as previously suggested [26].

While virulence factors were associated with human isolates, we oberved associations between certain isolate sources and mechanisms used in antagonistic interactions with other bacteria. This includes overpresentation of type VI secretion systems (T6SSi) in ExPEC, type 5b secretion systems (often associated with contact-dependent inhibition) in poultry meat isolates, and bacteriocins in several isolation sources (S2 Dataset). To detail these results, we searched specifically for colicin gene clusters [66], using BAGEL3 [67] (some of which are also included in VFDB). We found from an average of 2.8 genes in B2 strains to 0.4 in B1 strains. Interestingly, the water isolates have the fewest colicin genes, presumably because free diffusion of these proteins in water makes them inefficient tools of bacterial competition (Fig 6C, col 10 and S19 Fig). Thus, local adaptations resulting from the acquisition of novel genes by HGT, involving antagonistic interactions with other bacteria are associated preferably with certain phylogroups.

E. coli from freshwater are different

E. coli has usually been regarded as a contaminant from animal, mostly human, sources and used to test water quality. Yet, recent data suggests that some strains could inhabit aquatic environments [68]. Given the contrast between the primary and secondary habitats of *E. coli*, respectively guts of endotherms and aquatic environments, this would imply marked differences between the 285 freshwater strains and the others. Indeed, our results show that these strains are systematically different. They are over-represented in phylogroup B1 (43%), a phylogroup under-represented in all other sources of isolation (Fig 6A and Fig 6B). On the other

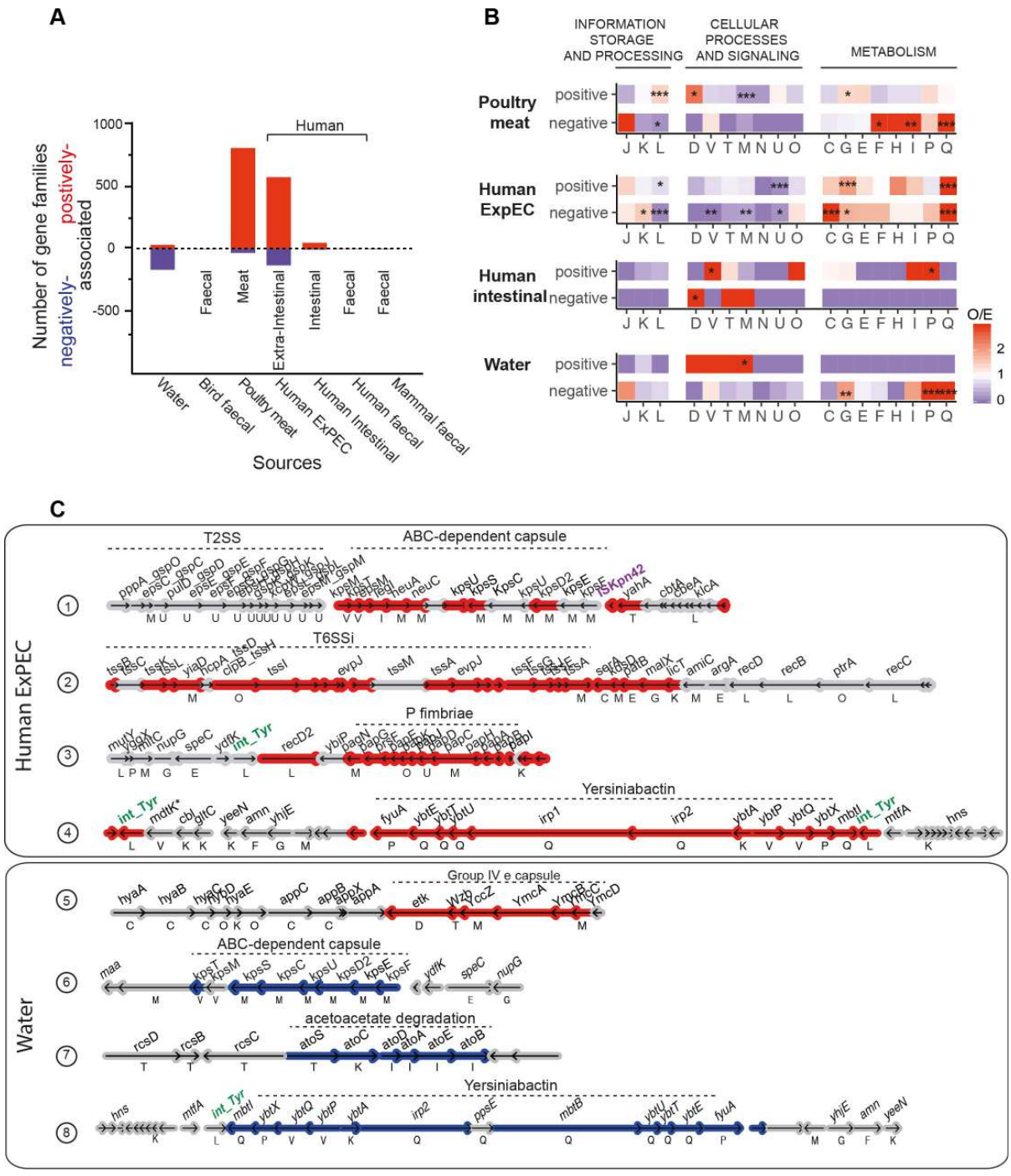

A
Number of gene families
postively-associated
negatively-associated
1000
500
0
-500
Human
Faecal
Meat
Extra-Intestinal
Intestinal
Faecal
Faecal
Water
Bird faecal
Poultry meat
Human ExPEC
Human Intestinal
Human faecal
Mammal faecal
Sources
B
INFORMATION STORAGE AND PROCESSING
CELLULAR PROCESSES AND SIGNALING
METABOLISM
Poultry meat
Human ExpEC
Human intestinal
Water
positive
negative
J K L D V T M N U O C G E F H I P Q
O/E
2
1
0
C
Human ExPEC
T2SS
ABC-dependent capsule
T6SSi
P fimbriae
Yersiniabactin
Water
Group IV e capsule
ABC-dependent capsule
acetoacetate degradation
Yersiniabactin

Fig 8. Genetic determinants of each isolation source. A. Number of gene families positively (in red) and negatively (in blue) associated with each source. **B**. *Observed/expected (O/E) ratios of non-supervised orthologous groups (NOGs, shown as capitalized letters, same as in Fig 1C) in the positively or negatively associated gene families. The ratio (O/E) was reported for all comparisons with a color code ranging from blue (under-representation) to red (over-representation). The level of significance of each Fisher's exact test was indicated (P> = 0.05: ns; P<0.05: *; P<0.01: **; P<0.001: ***). It was performed on each 2*2 contingency table. Only gene families with known functions were considered in this analysis.* Gene families lacking matches to the EggNOG functional categories were discarded.**C**. *Genomic organization of regions enriched in genes strongly positively (in red) or negatively (in blue) associated with a source. Genes shown in grey are not significantly associated. The name of the gene (when available) is shown above it, its functional category (when known) below it.*

https://doi.org/10.1371/journal.pgen.1008866.g008

hand, they are under-represented in B2 (13%), a phylogroup over-represented in strains isolated from humans (this study) and other mammals [2]. The genome size of freshwater strains' is the smallest among all groups of isolates and across phylogroups (Fig 6C, col 1, S17 Fig). Importantly, these strains show average pan-genome sizes in the rarefied dataset, suggesting that adaptation is not exclusively due to genome reduction (Fig 6C, col 12). This is also supported by the high number of gains and losses observed (Fig 6C, cols 14,15), although these genomes have the fewest MGEs and often lack plasmids (Fig 6C, cols 2–6). Consistent with adaptation to this habitat, they have the smallest number of antibiotic resistance genes, virulence factors, and bacteriocins (Fig 6C, cols 7–10, S18 and S19 Figs). In contrast, these strains show the highest diversity of STs and O:H serotypes (Fig 6C, cols 17,18, and S5 Table), and the highest number of capsule systems (Fig 6C, col 12, S20 Fig).

The extreme genomic traits of isolates from water strongly suggest they are not the result of recent fecal contamination from other sources. Instead, they strongly suggest that these strains have changed to adapt to water environments. This change seems to have involved the loss of many genes, and this is apparent from the GWAS analysis, which shows many more negative than positive associations with this isolation source (contrary to all the others) (Fig 8A). Many of them correspond to the virulence factors described above (Fig 8C). The few gene families positively associated with freshwater are over-represented in the EggNOG category M (cell envelope, Fig 8B). Many of these correspond to genes encoding the Group IVe capsular genes (Fig 8C), which contrasts with ABC-dependent capsules that are positively associated with Human ExPEC strains (S2 Dataset). Capsules have been proposed to allow cells to withstand biotic and abiotic challenges, and these results suggest that they are an important component of *E. coli* adaptation to freshwater environments. Overall, these results show that *E. coli* in these environments endured some horizontal gene transfer and important genome streamlining, i.e. a high turnover of gene repertoires that resulted in genomes smaller than the average carrying a few specific adaptations to the environment.

Discussion

Many of the recent advances in the understanding of *E. coli* evolution focused on clinical isolates and placed a lot of emphasis on virulence and antibiotic resistance in a few clinically important lineages [69–74]. Yet, most strains of the species are commensal. Hence, most of the evolution of the species takes place in biotic contexts not associated with pathogenesis. Furthermore, while a lot of attention has been given to the rates of homologous recombination in core genes, it is now clear that the acquisition of novel genes drives the evolution of virulence [12, 42, 75, 76] and antibiotic resistance [77–79] in pathogenic strains as well as that of many other traits in commensal strains [12]. For example, MGEs were recently shown to be more important than point mutations for the colonization of the mouse gut by *E. coli* commensals [80]. Here, we aimed at providing a global picture of the evolution of the *E. coli* genomes with an emphasis on the variation of gene repertoires in strains from a variety of sources (environmental and geographic) across a single continent. This allowed us to study the joint effect of population structure and habitat on the variation of gene repertoires. Our study focused on *E.*

coli isolates from Australia, but its genetic diversity was higher or comparable to other worldwide genome datasets, and its population structure was consistent with previous works [16, 40, 81]. This indicates that what we have observed is likely to be representative of the species as a whole. It also confirms previous reports of the large genetic diversity of the species and of the planetary circulation of all major lineages [39, 45, 82]. Finally, the functional annotation of the pan-genome shows that in spite of over 375,000 papers citing *E. coli* in PubMed in 2019, we are still far from having discovered the full genetic diversity of *E. coli* and from knowing the function of many of its most frequent gene families.

We started our study by quantifying gene repertoire diversification, which we found to follow a two-step dynamic. The very rapid initial diversification, where GRR quickly decreases to ~90%, implicates substantial heterogeneity in terms of gene repertoires for strains that are from the same sequence type and are almost identical in the sequence of persistent genes. Some of the rapid initial divergence of GRR may be due to genome sequencing or assembling artifacts producing singletons and thus inflating pan-genomes. Yet, we have annotated all genomes in the same way. We also confirmed key results by excluding singletons, by showing that singletons represent only ~0.5% of a typical genome, and that many of them have homologs in the databases. The frequency of singletons is only weakly correlated with the number of contigs in draft assemblies, a further sign that they are not just caused by sequencing or assembly issues (S3 Text). Furthermore, our analysis of ancestral genomes showed that a large fraction of well-known MGEs, including phages, ISs and plasmids, were acquired very recently (inferred acquisition at the terminal branches of the phylogenetic tree). Some of these are singletons, whereas others are present across a few genomes of many phylogroups. They contribute directly to the very rapid divergence of gene repertoires between separating lineages. Hence, we do not think that technical issues alone explain the existence of rapid gene repertoire differentiation between recently divergent strains. This raises the question of how much these processes reflect natural selection on incoming genes or high rates of gene loss by drift.

Previous population genetics models applied to other clades observed the existence of genes that have rapid turnover in genomes, i.e. that are rapidly lost after being acquired [83, 84]. Our results show that frequent acquisition of MGEs drives rapid diversification of gene repertoires even between strains that are almost indistinguishable by classical typing schemes. In the present context, this suggests that either many integrations of genetic material are deleterious and get rapidly purged by natural selection or that they are of no lasting adaptive value and get rapidly deleted by genetic drift. The first hypothesis is consistent with the fitness costs associated with the acquisition of many MGEs [85–87], with our observation that most MGEs present in a genome were very recently acquired, and with the abovementioned rapid loss of GRR for small patristic distances. The second hypothesis is consistent with previous works suggesting the existence of mechanistic biases towards gene deletion in bacteria that quickly remove genes without adaptive value from the genome [88, 89]. It is also consistent with the observation that some classes of functions, like defense systems [90] or specific components of the cell envelope [91], are subject to fluctuating selection dynamics and become neutral or slightly deleterious (because costly) after a short period where they are selected for.

After the initial period of rapid GRR decrease with phylogenetic distance, GRR decreases linearly with divergence time, a trend that was not quite clear when we first analyzed this question a decade ago with a much smaller set of genomes [42]. Importantly, this linear decay is not suggestive of the existence of a point beyond which relatedness and gene flow change abruptly. Hence, these results do not suggest incipient sexual isolation within the species from the point of view of horizontal gene transfer. This is confirmed by our analysis that some MGEs are present in many phylogroups and by the finding that many gene families of the pan-genome were recently acquired independently by distantly related strains. An interesting

feature of the comparisons of GRR in function of phylogenetic distances is the large variance around the regression line. This variance may result from very different processes. One of them may be preferential transfer of genes across strains within the same habitat, as observed for the isolates from the same type of source in this work. This type of transfer will lead to pairs of strains with more similar gene repertoires than expected given their patristic distance. Conversely, bacteria shifting from one habitat to another may endure an acceleration of their divergence in terms of gene repertoires. This will be a consequence of selection for different traits, acquired by HGT, and of changes in its preferential gene flow towards strains from the novel habitat.

The rapid evolution of gene repertoires by HGT is consistent with the observation that plasmids, prophages and ISs are almost ubiquitous among *E. coli*. These elements contribute significantly to genome size and even more to the variability of genome size across strains, which supports our previous results [51, 92]. While most MGEs are quickly lost from lineages, or drive the lineage extinct, the large influx of such elements can bring adaptive accessory traits such as antibiotic resistance genes [78] and virulence factors [93, 94]. They also pave the way for cooption processes [95]. The contribution of the MGE genes to genome size across the species is more strongly associated with the isolation source of the strains than with the phylogroup. However, the recent co-acquisition of MGEs by different strains is also associated with the phylogroup. This is consistent with a scenario where the abundance of MGEs in a genome is strongly dependent on the habitat, but their diversity also depends on the phylogroup. Since most MGE genes arrived in the genome very recently, this suggests that habitat exerts a strong constraint on the flow of gene exchanges across the species, in line with the view that bacteria exchange more genes with those they coinhabit with [96, 97].

The adaptive novel genetic information being acquired with MGEs must be integrated in the cell functioning. This need of favorable genetic backgrounds for certain local adaptation processes could explain the observed over-representation of some phylogroups in certain isolation sources. Virulence factors and antibiotic resistance genes provide relevant examples. In our dataset, the plasmids encoding virulence factors are often conjugative and should be able to circulate widely, but the virulent clones often concentrate in a few phylogroups. Selection for antibiotic resistance is expected to be higher in human-associated clones, and especially the virulent ones, because these are the most targeted in the clinic. Hence, they endure stronger selection to keep the ARGs arriving in MGEs. These causal links result in preferential associations of genetic backgrounds with virulence factors and ARGs, and therefore with the frequency of human isolates in a given phylogroup. It remains to be quantified the degree to which these trends are due to epistatic interactions between novel genes and the genetic background and to the availability of specific genes by horizontal transfer in certain sources. In conclusion, these results contribute to explain why epidemiological clones tend to emerge from specific phylogenetic groups even in the presence of massive horizontal gene transfer.

Genetic diversity, created by HGT, recombination, or mutation, affects a species' ability to adapt to novel ecological opportunities. The higher the diversity of gene repertoires in a population, the more likely that one of those genes will prove helpful in the face of environmental challenges such as antibiotics. We observed that the generalist phylogroups, such as A and B1, have larger pan-genomes than specialist phylogroups like B2. This was not expected based on their smaller genome sizes or the lower frequency of MGEs in their genomes. We propose that this reflects the high variability of the environments where they circulate—in terms of conditions, other strains and MGEs—and the associated diversity of local adaptation processes. Phylogroup B1, in particular, is associated with the presence of a number of metabolic traits suggesting interactions with plants. Phylogroup B2 strains, by comparison, have developed specific traits that may let them take advantage of some particular resources, e.g. they are better

adapted to the mammal gut environment [2]. This has resulted in large genomes that are quite different from the other major phylogroups of *E. coli*, as revealed by the phylogeny of the species based on the polymorphism on persistent genes, the PCA analysis of the pan-genome matrix, the GWAS analysis, and the large number of MGEs identified in their genomes. Yet, they are overall more conserved (largest persistent-genome, smaller pan-genomes, fewer recent gene acquisitions). This may explain why it has been suggested that strains from phylogroups A, D and B1 derived from an ancestral B2-like genetic background. The conservation of a larger core genome is consistent with our quantification of genetic exchanges: B2 strains exchange less genetic material with strains from its own and from other phylogroups than the remaining large phylogroups. This has placed it apart in terms of gene repertoires and in terms of preferential habitats. Altogether, these results suggest that the habitat and the phylogenetic structure jointly determine the size of genomes. The results also suggest the hypothesis that the large genomes of some phylogroups, like B2, may be caused by a relative decrease in the rate of gene loss, and not necessarily by an increase in the rate of gene gain.

The integration of information on gene repertoires, population structure and isolation sources sheds some light on the origin of environmental strains. This is illustrated by the identification of genomic traits in freshwater *E. coli* isolates that are very different from the average traits of the species and that suggest adaptation of certain lineages to this environment. For bacteria, freshwater environments are much more nutrient poor than the guts of endotherms, and it's interesting to note that strains associated with this environment have more streamlined genomes. This may represent, at the micro-evolutionary scale, an adaptation similar to that observed in other bacteria adapted to poor nutrient environments that also have small genomes and few MGEs [98, 99]. These results are also consistent with recent studies showing that *E. coli* B1 strains can persist longer in water than strains of the other phylogroups, and that B1 strains isolated repeatedly in water often encode very few virulence factors and antibiotic resistance genes [7, 33, 34]. Interestingly these strains have been shown to be able to grow at low temperatures [7]. The prevalence of B1 isolates has been observed in other environmental samples, such as drinking water and plants [64]. The characteristics observed in freshwater isolates might be general to this environment, since they were observed in strains from the B1 and from other phylogroups (S16–S20 Figs). If some *E. coli* lineages are indeed adapted to freshwater this radically changes the range of environments from where they can acquire novel genes and the selection pressures that shape their subsequent fate. This finding also implies that environmental isolates are not necessarily the result of source-sink dynamics where *E. coli* strains evolve in relation to selection pressures linked to the host and environmental strains are just sinks where such strains find evolutionary dead-ends. Instead, the environment outside the host could have a significant impact on the evolution of *E. coli* subsequently colonizing human hosts.

Materials and methods

Strains

We used different collections of *E. coli* strains recovered in Australia between 1993 and 2015 (for a more detailed description, see S1 Text and S1 Dataset). The subset of strains selected for whole genome sequencing includes: (1) *faecal strains* isolated from various birds (N = 195 strains), non-human mammals (N = 135), and humans living in Australia (N = 93); (2) *clinical strains* isolated during intestinal biopsies of patients with inflammatory bowel disease (N = 172), or corresponding to human ExPEC strains collected from urine or blood (N = 112); (3) *poultry meat strains* isolated from chicken meat products from diverse supermarket chains

and independent butcheries (N = 283); (4) and *freshwater strains* isolated from diverse locations across Australia (N = 285).

Sequencing

Of the 1,304 isolates, 70 were sequenced at Broad institute using the Roche 454 GS FLX system (this was done 10 years ago, detailed in [100]), 70 were sequenced by GenoScreen (Lille, France) using the HiSeq2000 platform. The rest were sequenced at the Australian Cancer Research Foundation (ACRF) Biomolecular Resource Facility (BRF) of the Australian National University, using the Nextera XT sample preparation kit (Illumina) and the Illumina Miseq (paired-end sequencing), as detailed in [101].

Assembling

Paired-end read files were processed and assembled with CLC Genomics Workbench v.9.5.3 (Illumina) using their *de novo* assembly algorithm with default parameters.

All genomes sequenced by the Broad institute were available into the NCBI Assembly (www.ncbi.nlm.nih.gov/assembly/) or SRA (www.ncbi.nlm.nih.gov/sra/) databases. While, the rest of the assemblies was deposited into the European Nucleotide Archive (PRJEB34791). The accession number of each genome is reported in S1 Dataset.

Datasets

We used 4 datasets in this study. (1) The ***Australian dataset*** described above is the main (default) dataset. (2) ***RefSeq dataset***: We retrieved 370 *E. coli* complete genomes from GenBank Refseq (available in February 2018). (3) ***ECOR dataset***: We retrieved 72 draft genomes of the *E. coli* reference (ECOR) collection from DDBJ/ENA/GenBank [48]. Strains in this collection were isolated from diverse hosts and geographic locations and have been used for more than 30 years to represent the phylogenetic diversity of *E. coli* as they have been selected from over 2,600 natural isolates based on MLEE data [17]. (4) ***Outgroup dataset***: We retrieved 65 other closely related *Escherichia* genomes from ENA/GenBank and sequenced 21 others on the Illumina MiSeq platorm (assembled as described above). They belong to Clade I (N = 14), Clade II (N = 2), Clade III (N = 8), Clade IV (N = 2), Clade V (N = 14), *E. fergusonii* (N = 8) and *E. albertii* (N = 38) species. Only five of them were complete, others were draft genomes. In this study, these genomes (called hereafter *outgroup* genomes) were only used to root the Australian *E. coli* species tree. The general genomic features and the sequencing status of these 1,832 genomes are reported in S1 Dataset.

Data formatting

In an attempt to overcome the bias from different annotations all genomes of the four datasets were annotated using Prokka v.1.11 [102] which provided consistency across the entire datasets (with hmmer v.3.1b1, aragorn v.1.2.36, barrnap v.0.4.2, minced v.0.1.6, blast+ v.2.2.28, prodigal v.2.60, infernal v.1.1, ncbi_toolbox v.20151127, and signalp v.4.0). We performed three quality controls on genomic sequences of Australian and outgroup datasets (see S2 Text). A total of 10 *E. coli* draft genomes and one genome from clade V failed at least one of these tests and were removed from further analysis, leading to a final dataset of 1,294 Australian *E. coli* genomes and 87 outgroup genomes. The main characteristics of each draft genome are reported in S1 Dataset.

E. coli typing. ***Phylogroup***. The phylogroup of each *E. coli* genome (from ECOR, RefSeq, and Australian datasets) was determined using the *in silico* ClermonTyping method [20].

Multilocus sequence typing* (*MLST*).** Sequence type (ST) was identified by the MLST scheme of Achtman [10] using mlst v.2.16.1 (https://github.com/tseemann/mlst). We assigned STs for a large majority of genomes, i.e., for 99%, 96% and 97% of the ECOR, RefSeq and Australian genomes resp. ***Serotype. Serotype (O- and H-genotypes) was inferred with the EcOH database [103] using ABRicate v.0.8.10 (https://github.com/tseemann/abricate)). Currently there are 220 *E. coli* O-groups and 53 H-types described in this database. While 99% of Australian genomes had H-group assigned, only 57% had O-group assigned even if *wzm*/*wzt* and *wzx*/*wzy* genes are present. All these results are reported in S1 Dataset.

Nucleotide diversity

The ***nucleotide diversity*** of the three datasets, *i.e.*, ECOR, RefSeq and Australian, was computed from the multiple alignments of 112 core gene families present in all *E. coli* genomes of these three datasets, (see below), using the diversity.stats function from the *PopGenome* v.2.6.1 R package [104]. We also used these 112 core gene families to assess the nucleotide diversity for each phylogroup of the Australian dataset.

ST and O:H diversity

The ***Shannon index*** was computed to assess the diversity of ST and O:H serotypes within each phylogroup and source. For this, we calculated their relative frequency in each group and then applied the function skbio.diversity.alpha_diversity from the *skbio.diversity* v.0.4.1 python package (http://scikit-bio.org/docs/0.4.1/diversity.html).

Mash distances (M)

***Genome similarity*.** Due to the high cost of computing ANI [105] via whole-genome alignment, we estimated genome similarity calculating the pairwise Mash distance (M) between all Australian genomes using Mash v.2.0 [106]. Importantly, the correlation between the Mash distances (M) and ANI in the range of 90–100% has been shown to be very strong, with $M \approx 1\text{-}(ANI/100)$ [106]. All the resulting Mash distances between *E. coli* genomes are well below 0.05, in agreement with the assumption that they all belong to the same species. The median is 0.027 and the maximal value is 0.04 (S4 Fig). **Australian *E. coli* reference genomes**. The Mash distance was strongly correlated to the patristic distance in our dataset (spearman's rho = 0.92, $P<10^{-4}$). We used it to select 100 Australian *E. coli* strains representative of the species' diversity (called hereafter *reference* genomes). Such *reference* genomes were used to root the Australian *E. coli* tree (to drastically reduce the computational time required to build the rooted tree). To select representative genomes, we performed a hierarchical WPGMA clustering from the Mash distance matrix computed with all Australian *E. coli* genomes, and then we cut it off to have only 100 clusters. In each of these clusters, the genome with the smallest L90 was selected. This *reference* dataset contained all the phylogroups and was composed of: 15-A, 10-B1, 13-E, 39-D, 11-F, 10-B2 and 2-G genomes.

Identification of pan-genomes

Pan-genomes are the full complement of genes in the species (or dataset, or phylogroup) and were built by clustering homologous proteins into families. We determined the lists of putative homologs between pairs of genomes with MMseqs2 v.3.0 [107] by keeping only hits with at least 80% identity and an alignment covering at least 80% of both proteins. Homologs proteins were then clustered by single-linkage [108]. We computed independently the pan-genome of each dataset, *i.e.*, ECOR, RefSeq, Australian and of the 87 outgroups with the 100 Australian *E.*

coli reference genomes. Each pan-genome was then used to compute a matrix of presence-absence of gene families. Hence, gene copy number variations were not taken into account in this part of the study. The alpha exponent of Heap's Law was used to infer whether a pan-genome is open or closed [46]. Thus, if α (alpha) < = 1, the pan-genome is open. In contrast, α (alpha) > 1 represents a closed pan-genome. This coefficient was computed using the *heaps* function of the *micropan* v.1.2 R package [109] with n.perm = 1000. Principal component decomposition of the Australian pan-genome, *i.e*, the matrix of presence-absence of protein families was computed using the *prcomp* function from the *stats* v.3.5.0 R package.

The pan-genome of each phylogroup and source was taken from the pan-genome of the species. The pan-genome of the MGE (called Pan-MGE) was also taken from the species pan-genome and contained only genes encoding for MGEs.

Rarefaction of pan-genomes

The number of singletons was strongly correlated to the number of genomes analyzed in each phylogroup (Pearson's correlation = 0.97, $P<10^{-4}$), indicating that the pan-genomes size depend on the number of genomes analyzed. Thus, to compare genetic diversity across datasets (e.g. phylogroups), we rarefied the genome datasets, *i.e*., each pan-genome was constructed with the same number of genomes in each comparison. To do this, 1,000 subsets of X genomes (X depending on the analysis, specified in the results section) were randomly selected for comparison in each group, resulting to datasets called hereafter *rarefied* datasets (S9 Fig).

Identification of persistent-genomes

Gene families that are persistent were taken from the analysis of pan-genomes. A gene family was considered as persistent when it was present in a single copy in at least 99% of the genomes. We found 2,486 persistent gene families when considering the 1,294 Australian genomes, representing 52% of the average genome.

Identification of core-genome

The core genome was taken from the analysis of the pan-genome. A gene family was considered as core if it is present in one single copy in all the genomes. To assess the nucleotide diversity, we built a core-genome with all the genomes of the ECOR, RefSeq, and Australian datasets. It was composed of 112 core gene families. Each gene family was aligned with mafft v.7.222 (using FFT-NS-2 method) [110], and used to compute the average nucleotide diversity (π) in each dataset and within each phylogroup (see above).

Functional assignment of the pan-genome

Gene functional assignment was performed by searching for protein similarity with hmmsearch from HMMer suite v.3.1b2 [111, 112] on the bactNOG subset of the EggNOG v.4.5.1 database [47]. We have kept hits with an e-value lower than 10^{-5}, a minimum alignment coverage of 50% of the protein profile, and when the majority (>50%) of non-supervised orthologous groups (NOGs) attributed to a given gene family pertained to the same functional group (category). The gene families that cannot be classified into any existing EggNOG clusters were grouped into the "unknown" category. Hits corresponding to poorly characterized or unknown functional EggNOG clusters were grouped into the "poorly characterized" category.

Phylogenetic analyses

We built a rooted phylogeny of the species in two steps. **The phylogenetic species tree of Australian *E. coli*** was reconstructed from the concatenated alignments of the 2,486 persistent genes of the 1,294 Australian *E. coli* strains (see S3 Fig for a description of the method). The alignment was done using the corresponding protein sequences with mafft v.7.222 (using FFT-NS-2 method) [110]. Protein alignments are more accurate and produce codon-based alignments that can be used for population genetics analysis. Since at this evolutionary distance the DNA sequences provide more phylogenetic signal than protein sequences, we back-translated the alignments to DNA, as is standard usage. This involved replacing every amino acid in the alignment by the original codon. Hence, the DNA sequence remains unchanged after translation and back-translation. We built phylogenies from persistent genomes to avoid the loss of signal associated with the small core genomes. When a genome lacked a member of a persistent gene family, or when it had more than one member, we added a stretch of gaps ('-') of same length as the other genes for it in the multiple back-translated alignments. Adding a few "-" has little impact on phylogeny reconstruction. For example, Filipski et al [113] showed that adding up to 60% of missing data in the alignment matrix could be informative. In our study, only 0.3% of the genes are missing in the matrix and the effect of missing data should be negligible relative to the advantage of using the phylogenetic signal from 2,486 persistent genes instead of only the one of 295 core genes (S3C Fig). We have not removed recombination tracts from the multiple alignment because this has been shown to amplify errors in determining phylogenetic distances and it usually does not affect the topology of the tree [114, 115]. If determination of the recombination was accurate in our >1,300 genomes dataset, this would have led to the exclusion of almost all the genes. The length of the resulting alignment for the species was 2,298,168 bp. Each tree was computed with IQ-TREE multicore v.1.6.7 [116] under the GTR+F+I+G4 model. This model gave the lowest Bayesian Information Criterion (BIC) among all models available (option–m TEST in IQ-TREE). We made 1,000 ultra-fast bootstraps to evaluate node support (options–bb 1000 –wbtl in IQ-TREE) and to assess the robustness of the topology of each tree [117].

The phylogenetic tree of *Escherichia* genus was inferred from the persistent-genome obtained with the 87 outgroup genomes and the 100 *E. coli* reference genomes (see above) using the same procedure as the species tree. In this case, the persistent-genome is composed of 1,589 gene families, and the resulting alignment of 1,469,523 bp. The genus phylogenetic tree was extremely well supported: all nodes had bootstrap support higher than 95%. Its topology was consistent with a previous study [118] (S4C Fig). Then, we used it to precisely root the species tree (S4D Fig).

The most recent common ancestor of each phylogroup: We identified the node corresponding to the most recent common ancestor (MRCA) for each phylogroup from the rooted species tree using the *findMRCA* function from the *phytools* v.0.6.44 R package. Then, the subtree of each phylogroup was extracted using the *extract.clade* from the *ape* v.5.2 R package [119]. The distance to the MRCA was computed from the length of branches in each subtree. It corresponds to the average depth (distance from the MRCA) of all genomes (tips) within a phylogroup and was inferred using the *depthTips* from the *phylobase* v.0.8.6 R package (https://github.com/fmichonneau/phylobase).

Evolutionary distances

For each pair of genomes, we computed a number of measures of similarity: 1) The **Patristic distance** was computed from the length of branches in the *Australian E. coli* species phylogenetic tree. The patristic distance is simply the sum of the lengths of the branches that link two

genomes (tips) in the tree, and was inferred using the *cophenetic* function from the ape v.5.2 R package [119]. They were computed between all pairs of genomes, of the same ST (*intra-ST*), of different ST (*inter-ST*) within identical phylogroup, or of different phylogroups (*inter-phylogroup*). As expected, we found that the *intra-phylogroup (both intra-ST and inter-ST)* patristic distances were significantly shorter than the *inter-phylogroup* (Wilcoxon test, $P<10^{-4}$). 2) **The Gene Repertoire Relatedness index** (GRR) between two genomes was defined as the number of common gene families (the intersection) divided by the number of genes in the smallest genome [120]. It is close to 100% if the gene repertoires are very similar (or one is a subset of the other) and lower otherwise. 3) **The Manhattan index** between two genomes is the number of different gene families. If two genomes have identical gene content, the corresponding Manhattan index is 0. 4) **The Jaccard index** between two genomes was defined as the number of common gene families (the intersection) divided by the number of gene families in both (the union). The Jaccard index between two genomes describes their degree of overlap with respect to gene family content. If the Jaccard distance is 1, the two genomes contain identical protein families. If it is 0 the two genomes are non-overlapping.

To characterize the genetic diversification of each phylogroup of the Australian dataset, we computed the three different standard indexes: the GRR, the Jaccard, and the Manhattan indexes. All these indexes were highly correlated (S10B Fig). Thus, only analyses with GRR were reported and illustrated in the main text. Note that we always used the matrix of presence/absence of gene families to compute all these indexes, meaning that multiple occurrences were not considered. This downplays the impact of IS on pan-genome size and makes more conservative estimates of GRR divergence.

Reconstruction of the evolution of gene repertoires

We assessed the evolutionary dynamics of gene repertoires of the Australian genomes using Count (downloaded in January 2018) [121] with the Wagner parsimony method. Due to the size of our dataset it was not possible to do the analysis using birth-death models, but our previous analyses revealed very few differences between the two methods in smaller datasets [122]. Wagner parsimony penalizes the loss and gain of individual family members (with relative penalty of gain with respect to loss of 1, option g = 1), and infers the history with the minimum penalty. Thus, from the pan-genome, *i.e.*, the matrix of presence-absence of gene families, and the rooted species tree, Count inferred the most parsimonious gain/loss scenario of each gene family along the tree. At each tree node, Count detailed information about individual families: presence/absence, and family events on the edge leading to the node. Hence, we have reconstructed the gene content of ancestral genome at each node. At each terminal branch, the expected total number of recent acquisitions (HGT) was computed by summing all family-specific gene gains obtained from the edge leading to the tip. Among them, we identified MGE associated genes that were recently acquired in each genome. We applied a similar strategy to identify recent losses.

Distribution of accessory families across phylogroups (or sources)

We counted the number of MGE-associated gene families across phylogroups (Fig 5A) or sources (S15 Fig). We excluded the singletons from this analysis to avoid over-estimation of the number of families specific to one category. To test if some categories over-represented or under-represented these genes, we made 1,000 simulations. In each simulation, we shuffled the phylogroup (or source) assignment of the genomes while keeping the same number of taxa in each category (phylogroups or sources). Thus, the presence of a gene family in a genome and its frequency in the pan-genome remains the same, only the phylogroup (or the source) of

genomes changes. The Z-score obtained for the observed number in the real data with respect to the random distribution (from 1,000 simulations) was reported for each case with a color code ranging from blue (under-representation, Z-score<-1.96) to red (over-representation, Z-score>1.96).

Recent co-occurrence of gains (co-gains) of gene families within phylogroups

We counted the number of recently acquired gene pairs (co-gains) from the same pan-genome gene family (see above) within and between phylogroups. Recently acquired genes were defined as those inferred as acquired in terminal branches using Count. To test if some phylogroups over-represented or under-represented these co-gains, we compared the observed number (O) within each phylogroup to the expectation (E) given by 1,000 simulations. In each simulation, we shuffle the phylogroup assignment of the taxa (same approach as for the accessory gene families) and count the number of co-gains within and between phylogroups. For each phylogroup, we then divided the number observed in the real data (O) by the average number observed in the simulations (E), and computed the Z-score of the observed number (O) with respect to the random distribution (E). We considered an over(under)-representation significant when Z-score>1.96 (Z-score<-1.96). Note that the O and E numbers had to be previously normalized (divided by the total number of gene pairs, i.e. the sum of pairs within and between phylogroups, in the real data, and in each simulation, resp.). We applied the same approach (i) considering only gene pairs encoding for MGEs (similar result as in Fig 5), (ii) for sources (instead of phylogroups, Fig 6).

Network of co-occurrence of gains (co-gains) of gene families across phylogroups

All co-gains (see above) were split into all possible combinations of phylogroup pairs (21 combinations). To test if these co-gains are over- or under-represented between phylogroups, we compared the observed number (O) between each phylogroup to the expectation (E) given by 1,000 simulations with the same strategy as above. As before, we normalized the observed and expected numbers by the total number of co-gains in each simulation, calculated the (O/E) ratio, and the Z-score of each observed value in the real data with respect to the random distribution (E). The network was drawn using the *igraph* v.1.2.2 R package (https://igraph.org/r/) with the circle layout option, where nodes are phylogroups, edges are (O/E) values for which the Z-score is significantly different from zero. The width of the edges is proportional to the (O/E) value and the color is blue for under- and red for over-representation (Fig 5C). We applied the same approach considering only gene pairs encoding for MGEs (S16 Fig).

Gene family diversity

We computed Shannon indexes to assess the diversity of each gene family recently acquired (terminal branches) across phylogroups and across sources (Fig 6E). If diversity is low, this means that acquisitions are clustered by phylogroup or source (depending on the analysis). For this, we calculated the relative frequency of each gene family recently acquired within each phylogroup (vs. each source). It is simply the number of genomes (within a phylogroup) with at least one acquisition divided by the total number of genomes in the phylogroup. We therefore obtained 2 vectors per gene family (one for phylogroups and one for sources) each containing 7 frequencies (for each phylogroup or each source) and then applied for each vector the function *diversity* from the *vegan* v.2.4.6 R package (https://github.com/vegandevs/vegan).

If the index is 0, recent acquisitions of genes of the family are limited to a single group (phylogroup or source). The higher the index, the more scattered the acquisitions of the family's genes are (across phylogroups or sources).

GWAS

We studied the association between the pan-genome, i.e., the matrix of presence-absence of gene families, and different phenotypes (i.e., phylogroups, and sources) using Scoary v.1.6.16 [123]. The method used the rooted species tree to correct for phylogenetic dependency. To correct for multiple comparisons, only gene families with a Bonferroni-adjusted p-value $< 10^{-10}$ were selected. In the case of phylogroups, more stringent thresholds were applied, i.e., p-value 10^{-20}. We used the odds ratio (R) to determine whether the gene is positively (R>1) or negatively (R< = 1) associated with the tested phenotype. Analyses of the whole pan-genome or excluding all singletons produced similar results. A complete list of gene families positively and negatively associated with each phenotype is described in S2 Dataset. The sequence of one gene from each family is also available in the S2 Dataset, to facilitate the use of these results by the community.

Statistics

All basic statistics were performed using R v 3.5.0, or *JMP*-13. (i) **Analysis of means**: We used **ANOM** to compare group means to the overall mean, when the data were approximately normally distributed. In cases where the data were clearly non-Gaussian and could not be transformed, we used the nonparametric version of the ANOM analysis, i.e., **ANOM with Transformed Ranks**. It compares each group's mean transformed rank to the overall mean transformed rank. In both, we used the methods implemented in *JMP*-13. (ii) **Pairwise Wilcoxon Rank Sum Tests** were computed using the *pairwise.wilcox.test* function from the *stats* v.3.5.0 R package. We used the Bonferroni correction during multiple comparison testing. (iii) **Fisher's exact tests** were computed using the *fisher.test* function from *stats* v.3.5.0 R package. They were performed for testing the null of independence of rows (phylogroups) and columns (sources) in a 2x2 contingency table. (iv) **Correlation coefficients.** Pearson's and Spearman's rank correlation rho were computed using the *cor* function from *stats* v.3.5.0 R package. The correlation matrices were represented using the *corrplot* v.0.84 R package (https://cran.r-project.org/web/packages/corrplot/index.html). (v) **Smooth regression**: We used the generalized additive model (*gam)* smoothing method from the *mgcv* v.1.8.23 R package (https://cran.r-project.org/web/packages/mgcv/index.html). (vi) **Stepwise multiple regressions** were computed with *JMP*-13. This standard statistical method consists in a stepwise integration of the different variables in the regression by decreasing order of contribution to the explanation of the variance of the data [124]. We used the forward algorithm and the BIC criterion for model choice in the multiple stepwise regressions. The P-values associated with each variable were assessed using an F-test.

Identification of Mobile Genetic Elements (MGEs)

Prophages: Prophages were predicted using VirSorter v.1.0.3 [53] with the RefSeqABVir database in all genomes from Australian and RefSeq datasets, as a control. The least confident predictions, *i.e.*, categories 3 and 6, were excluded from the analyses in both datasets. The prophage-associated regions in drafts are more numerous and shorter than in complete RefSeq genomes (S11 Fig). These results reveal that such regions are sometimes split in assemblies. In complete genomes, the cumulative size of the prophage-associated regions (X) is highly correlated with the number of prophages (Y) present in the genomes ($Y = 1.2923362 + 1.6767.10^{-5}$ X, $R^2 = 0.91$, $P<10^{-4}$, S11 Fig). Hence, we used this linear equation to estimate the number of

prophages in drafts using the cumulated size of prophage regions in the draft genomes. **Plasmids**: In the RefSeq dataset, all the extrachromosomal replicons were considered as plasmids. In the Australian dataset, plasmid sequences were identified using PlaScope v.1.3 [54] with the database dedicated to *E. coli*. PlaScope provides a method for plasmid and chromosome classification of *E. coli* contigs. It has the specificity to select a unique assignment to each contig of a draft genome to plasmid, chromosome or unclassified. The number (~16, max: 124) and size (~9 kb, max: 166 kb) of contigs predicted as plasmid were highly variable (S12 Fig) in the Australian dataset. Their size is much smaller than that of the average plasmid in complete genomes (~80 kb), reflecting the split of plasmids across different contigs because of the presence of repeated sequences, *e.g.* IS elements. Hence, we have not attempted to estimate the exact number of plasmids per genome and focus our analysis on the number of genes predicted to be in plasmid contigs. **MGEs (Plasmids + Prophages)**: We found 11,864 gene families specifically related to plasmid elements, 14,188 to prophage elements, and 2,599 shared by both (9% of the MGEs gene families). In complete genomes, prophage and plasmids elements account for half of the pan-genome, of which 1 third were singletons. The large fraction of singletons from MGEs confirms that these elements are extremely diverse and evolved very rapidly, which underlines the difficulty of accurately detecting them and probably leads to their under-estimation in draft genomes. **Loci encoding conjugative or mobilizable elements** were detected with the CONJscan module of MacSyFinder [125], using protein profiles and definitions following a previous work [55, 126]. 87% of conjugative systems and 75% of putative mobilizable elements were located on contigs predicted as plasmids by Plascope. **Integrons** were identified using IntegronFinder v.1.5 with the–local_max option [58]. 186 integron-integrase (*intI*) were detected with one quarter located at the edges of contigs. We only found one copy per genome. They were often located on very short contigs (20 proteins on average), and five make all the contigs. Most (86%) were located on contigs predicted as plasmid by Plascope, the remaining were on unclassified contigs. Except for the latter, *intI* genes were always located next to ARGs. **IS elements** were identified using ISfinder [57]. Only hits with an e-value lower than 10^{-10}, a minimum alignment coverage of 50% and with at least 70% identity were selected, we extracted the IS name of the best hit. Therefore, we identified 47,592 genes encoded for IS elements, among them 43% were located at the edges of contigs (20,329/47,592). They represented 1,006 gene families (~1% of the pan-genome), of which 41% were singletons. Only 13% were multigenic protein families (*i.e.*, with more than one member in at least one genome). Among them, 9 protein families were found in more than 10 copies in at least one genome, i.e., ISEc1 (10 copies), IS1397 (11), ISSoEn2 (11), IS621 (11), IS2 (15), IS629 (17), IS200C (17) IS1203 (18), and the most extreme case IS1F (107). Very large numbers of ISs, usually a sign of recent proliferation, was restricted to a small number of genomes (S1 Dataset), but this may be an under-estimate caused by the loss of ISs in the assembling process. ISs were often fragmented, characterized by numerous singletons, and six times more frequently present at the edges of contigs than expected by chance. All the results are reported in S1 Dataset.

Capsule systems

We used CapsuleFinder as published in [127] to search for Group I (Wzy-dependent), Group II and III (ABC-dependent), Group IV (subtypes e, f and s), synthase-dependent (subtypes cps3-like and hyaluronic acid) and PGA (Poly-γ-d-glutamate) capsules in the genome database. This allowed the detection of 2,829 systems: 1,236 Group I, 123 Group II, 777 Group IV e and 693 Group IV s. All the results are reported in S1 Dataset.

Antibiotic resistance genes (ARG) were detected using 2 curated databases of antibiotic resistance protein: Resfinder v.3.1 [128] and ARG-ANNOT v.3 [129]. Therefore, we used

BlastP and selected the hits with an e-value lower than 10^{-5}, with at least 90% of identity and a minimum alignment coverage of 50%. We found a strong positive correlation between the number of ARGs per genome using each database (pearson's r = 0.97, $P<10^{-4}$). The main difference is the additional detection of three ARGs by ARG-ANNOT, i.e., AmpC2, AmpH, Mfd, which are persistent in Australian dataset and normally do not confer antibiotic resistance in *E. coli*. All the results are reported in S1 Dataset.

Virulence factors (VF) were identified using VFDB (downloaded in February 2018, [65]). The two databases, *i.e.*, VFDB_setA and VFDB_setB were used independently. We used BlastP and selected the hits with an e-value lower than 10^{-5}, at least 70% of identity and minimum alignment coverage of 50%. We found 1,332 (vs. 3481) gene families encoding virulence factors with the setA (vs. setB). In spite of these differences, we found qualitatively similar conclusion with the 2 sets because they are very correlated (pearson's r = 0.97, $P<10^{-4}$). All the results are reported in S1 Dataset.

Supporting information

S1 Text. Isolates description.
(DOCX)

S2 Text. Quality control of the genomic sequences.
(DOCX)

S3 Text. Effect of contig breaks on the estimates of pan-genomes.
(DOCX)

S1 Table. Overall diversity of the three datasets.
(PDF)

S2 Table. Genetic diversification across phylogroups of Australian dataset.
(PDF)

S3 Table. The effects of phylogroup and the strains' source on genome size: results of the stepwise multiple regression. Stepwise regression is an approach to selecting a subset of parameters (among the strains' source and phylogroup) for a regression model. In forward selection, terms are entered into the model and most significant terms are added until all of the terms are significant. We used the minimum Bayesian Information Criterion to choose the best model. The Stepwise regression report (1) shows the statistics of the best model. As each step is taken, the Step History report (2) records the effect of adding a term to the model, and shows the order in which the terms entered the model and the statistics for each model. The Current Estimates report (3) indicates whether a term is currently in the best model and shows the statistics of each term for this model.
(PDF)

S4 Table. The effects of phylogroup and the strains' source on genome size without MGE: results of the stepwise multiple regression. Same approach described in S3 Table.
(PDF)

S5 Table. Genetic diversification across sources of Australian dataset.
(PDF)

S6 Table. The effects of phylogroup and the strains' source on MGE content: results of the stepwise multiple regression. Same approach described in S3 Table.
(PDF)

S1 Dataset. The main characteristics of each genome of this study.
(XLSX)

S2 Dataset. Association of the pan-genome with phylogroups and isolation sources: results of the GWAS analyses.
(XLSX)

S1 Fig. General genomic characteristics of the 1,294 Australian *E. coli* genomes. A. Histogram and boxplot of genomic features, *i.e*., the genome size (Mb), the number (#) of genes encoding proteins, the GC content (GC%), the gene density, the number of essential genes, the number of contigs and the L90 (Methods). For each case, the dash line corresponds to the smoothed curve, the red arrow to the median and the blue arrow to the average of each distribution. B. Strong positive correlation between the genome size and the number of genes (spearman's rho = 0.98, $P<10^{-4}$). C. Weak positive correlation between the genome size and the number of contigs (spearman's rho = 0.23, $P<10^{-4}$). The genomes with the greatest number of contigs were not necessarily the largest. Linear regression (dash line) and statistics were reported.
(EPS)

S2 Fig. The large Australian *E. coli* pan-genome. A. Number of gene families according to their occurrence in genomes. Singletons (in green), *i.e*., genes present in a single genome, represent 44% of the pan-genome. Persistent gene families (in gold), *i.e*., present in at least 99% of genomes, represent only 3% of the pan-genome. B. Fraction of gene families (%) according to their frequency among the pan-genome and the average genome. Frequencies were represented by a color code ranging from light grey (present in less than 1% of genomes) to black (up to 99%), persistent genes (>99%) were represented in gold. 82% of the gene families are rare, *i.e*., present in less than 1% of genomes including the 33,705 singletons. Persistent gene families represent 53% of the average genome, while singletons less than 1%. C. Rarefaction curve of the full pan-genome and of the pan-genome after removing the 33,705 singletons (wo. S). In each case, we used 1,000 permutations (genomes orderings) and then averaged the results. The *alpha* (inferred using the heaps' law model) is lower than 1 in both, indicating that the pan-genome is open in both. D. Rarefaction curve of the persistent genome (in gold) and of the core genome (in red), *i.e*., the cumulative number of gene families shared by 100% of the genomes. The evolution of the average number of new genes per genome is also reported (in green). When considering 1,294 genomes, there is on average 2,486 persistent proteins and only 26 singletons per genome. E. Violin-plots of the average sequence identity [left, mean], and the minimal sequence identity [right, min] observed in each of the 2,486 persistent gene families. The observed average sequence identity is 98.3% across families of persistent genes. The average minimal value observed across persistent gene families is 95.5%.
(EPS)

S3 Fig. Construction of the concatenated alignments of persistent gene families. A. Graphical representation of the different steps of the phylogenetic trees build process from the persistent genome. Among persistent gene families, there are families that are core (present in 100% of the genomes, in red) and the remaining that have missing genes (not-core, in gold). B. Number of persistent gene families according to their number of missing genes in the Australian dataset. Only 12% of families are core, i.e., present in all genomes (in red). C. Violin-plot of the number of missing genes per genome in the Australian dataset. On average, the number of missing genes is around 8 per genome. It can reach up 93 in a single genome, but this represents less than 4% of persistent families.
(EPS)

S4 Fig. The genus and species phylogenetic trees. A. Distance tree of 1,294 Australian *E. coli* and 86 outgroups genomes performed from the matrix of mash distances computed between all pairs of genomes using bionj. The number of genomes in each species (or clade) was indicated. The different phylogroups of *E. coli* were displayed: A (in blue), B1 (in green), E (in purple), D (in yellow), F (in orange), G (in brown) and B2 (in red). B. Boxplot of the mash distances computed between all pairs of genomes belonging to the same species (or clades). In both cases, the maximal mash distance was lower than 0.05. For *E. coli* species, the median was around 0.027 and the maximal value was 0.04. C. Phylogenetic tree of 100 Australian *E. coli* genomes representative to the diversity of the dataset and 86 outgroups genomes performed from the persistent-genome of the genus with IQ-TREE under the GTR+F+I+G4 model. We made 1,000 ultra-fast boostrap to assess the robustness of the topology of the tree. We found that all boostrap supports were higher than 95%. D. We rooted the species phylogenetic tree from the genus phylogenetic tree. The resulting rooted species tree was reported, and for simplicity, the main phylogenetic groups were collapsed.
(EPS)

S5 Fig. Singleton characterization. A. Boxplots of gene size (bp) in the three categories of gene families, *i.e.*, persistent (in gold), accessory (in grey) and singleton (in green). The average was represented by a black dot. The pairwise Wilcoxon Rank Sum test with bonferroni correction was applied to all comparisons ($P<0.001$:***). B. Same analysis as in A, but distinguishing the genomic location of the gene of each set : inside of contigs (I, dark color) or at the edge of contigs (E, light color). The average gene size for each case was reported in the table. C. Percentage of genes located inside contigs (dark color) or at the edge of contigs (light color) in the 3 sets. The last column corresponds to the fraction of the 3 sets located at the edge of contigs. D. Heatmap of the observed/expected (O/E) ratios of genes located inside or at the edges of contigs in the 3 sets. The ratio (O/E) was reported for all comparisons with a color code ranging from blue (under-representation) to red (over-representation). The level of significance of each Fisher's exact test was also indicated ($P<0.001$:***). It was performed on each $2^{*}2$ contingency table. E. Fraction of singletons with no hit (in light gey), with a small domain (in grey) or fully included (black) in larger accessory or persistent gene families (S3 Text). F. Violin plots of the number of singletons (in green) or persistent (in gold) observed in the rarefied Australian and RefSeq datasets. In each case, 1,000 permutations of 50 randomly selected genomes were performed (i.e., we used rarefied datasets). The boxplot is in white and the mean is represented by a black dot. While the average number of singletons is significantly higher (30% more) in the rarefied Australian dataset (Wilcoxon test, $P<10^{-4}$), the average number of persistent is also significantly higher (5% more, $P<10^{-4}$) than the rarefied RefSeq dataset. Singletons represent 43%, and 35% of the rarefied Australian and RefSeq pangenomes, resp.
(EPS)

S6 Fig. Association between GRR (%, Gene Repertoire Relatedness) and the patristic distance of each pair of genomes. Here, the GRR were computed excluding singletons in all genomes. Due to the large amount of comparisons (points), we divided the plot area in regular hexagons. Color intensity is proportional to the number of cases (count) in each hexagon. The linear fit (full line, linear model (lm)) and the spline fit (dash line, generalized additive model (gam)) were reported for the whole (in black, all the species) or the intra-ST (in blue) comparisons. There was a significant negative correlation between GRR and the patristic distance (spearman's rho = -0.69, $P<10^{-4}$). The summary of the linear fit was: Y = 90.722391–

76.2919X, $R^2 = 0.50$,$P<10^{-4}$. Hence, with or without singletons, the results were similar.
(EPS)

S7 Fig. Comparison of Australian, ECOR and RefSeq datasets. A. Violin plots of the nucleotide diversity per site in the 3 datasets computed from the multiple alignments of 112 core gene families (see Methods). The pairwise Wilcoxon Rank Sum test with bonferroni correction was applied to all comparisons (P>0.05:ns). B. Rarefaction curve of the full pan-genomes of the 3 datasets. In each case, we used 1,000 permutations (genomes orderings) and then averaged the results. C. Violin plots of the size of the pan-genomes computed from the three rarefied datasets: In each case, 1,000 permutations of 50 randomly selected genomes were performed to calculated the rarefied pan-genomes. The pairwise Wilcoxon Rank Sum test with bonferroni correction was applied to all comparisons ($P<10^{-3}$:***). D. Average number of persistent (in gold), accessory (in grey) and singleton (in green) in the rarefied pan-genomes of each dataset.
(EPS)

S8 Fig. Intra- and Inter-phylogroup genetic diversity. A. Violin plots of the nucleotide diversity per site (left), the MASH (center) and the patristic distances (right) computed with/between genomes belonging to the same phylogroup (intra-phylogroup, in seagreen), to different phylogroups (inter-phylogroup, in purple), or all together (ALL, in darkgrey). In all cases, intra- and inter-phylogroup distributions were significantly different (Wilcoxon tests, $P<10^{-4}$). B. Boxplots of the nucleotide diversity (left), the MASH (center) and the patristic distances (right) computed with/between genomes in each phylogroup. The pairwise Wilcoxon Rank Sum test with bonferroni correction was applied to all comparisons. Here, only the non-significant (ns : P> = 0.05) comparisons were indicated, all other were higly significant $P<10^{-4}$. C. Density of the patristic distances between all pairs of genomes of the same phylogroup (*intra-phylogroup*). The dash vertical line corresponds to the median of each distribution. (A-B-C) In all cases, similar results were obtained with rarefied datasets (i.e., comparing 50 randomly selected genomes in each groups, thus ignoring the small G phylogroup).
(EPS)

S9 Fig. Pan-genomes, Pan-MGE, and rarefied Pan-genomes of each phylogroup and isolation source. A. Size of the pan-genome in each phylogroup and in each isolation source. The pan-genome sizes were correlated to the number of genomes in each group, even after excluding the singletons from the analysis (both, adjusted $R^2>0.88$, $P<10^{-4}$). The Rarefaction curve of the pan-genomes of the full dataset was also reported (All, in black). B. Rarefaction curves of the pan-genomes of each phylogroup and of the full dataset (All). C. Rarefaction curves of the gene-families associated to MGE in each phylogroup and in the full dataset (All). D. Rarefaction curves of the pan-genomes of each isolation sources. In each case, (i) we used 1,000 permutations (genomes orderings) and then averaged the results (full line = mean, dash line = s.d), (ii) the pan-genomes remained open (with an *alpha* lower than one, see methods) that we considered them as a whole or without singletons, (iii) the boxplots of the rarefied pan-genomes (using a number of genomes = 50) were reported. The color code used was displayed in the insert (top right).
(EPS)

S10 Fig. Gene repertoire relatedness (GRR) within and between phylogroups. A. Average GRR (%) computed between pairs of genomes belonging to the same phylogroup (intra-phylogroup) and to different phylogroups (inter-phylogroup). The color code used was displayed in the insert (top right). B. Correlation between the different distances and indexes, *i.e.*, GRR, Manhattan, Jaccard, MASH and patristic, computed between pairs of genomes belonging to

the same phylogroup (intra-phylogroup) with the whole dataset or excluding singletons (woS). Spearman's rank correlation rho matrix. Positive correlations were displayed in red and negative correlations in blue color. Color intensity and the size of the circle were proportional to the correlation coefficients. The p-value of each correlation was highly significant ($P<10^{-4}$). We found similar results with rarefied datasets, i.e., considering only 50 randomly selected genomes in each phylogroup. We also found higher correlation coefficients using all the comparisons (intra- and inter-phylogroup).
(EPS)

S11 Fig. Detection and Estimation of the number of prophages. A. Boxplot of the number of regions detected as prophage-related by Virsorter in the 370 complete RefSeq GenBank genomes and in the 1,294 draft Australian genomes. These distributions were significantly different, on average the number of regions detected was significantly higher in draft than in complete genomes (Wilcoxon test, $P<10^{-4}$). B. Boxplot and histogram of the size of the detected regions in complete and draft genomes. These distributions were significantly different (Wilcoxon test, $P<10^{-4}$). On average the regions were almost 4 times larger in the complete genomes than in draft genomes and few regions (644) in draft genomes had a typical size of known dsDNA phages (around 44kb). (A-B) showed that prophage elements were less assembled and were probably divided into several small contigs. The large regions (>60 kb) in complete genomes corresponded to tandem elements (consecutive on the genomic sequence). Thus, the number of detected regions did not correspond to the number of prophages either in the complete genomes (due to tandem elements) or in the drafts genomes (the elements being fragmented). C. Strong association between the cumulative size of the detected regions (X) with the number of detected regions (Y). Linear regression (dash red line) and statistics were reported. D. Boxplot of the predicted number of prophage elements in both the complete and the draft genomes using the linear equation showed in (C) from the cumulative size of the regions detected by VirSorter. These distributions were significantly different (Wilcoxon test, $P<10^{-4}$). On average, there was 6.0 prophages in complete genomes, and 4.25 in draft genomes. The medians of the two data sets were closer reflecting probably the assembly problem related to the presence of prophages in tandem combined with the fact that they are often genetically close (most of them are lambdoids). In each panel, the red arrow corresponds to the median and the blue arrow to the average of each distribution.
(EPS)

S12 Fig. Detection of plasmid elements. A. Boxplot of the number of contigs classified as plasmid by PlaScope in the 370 complete RefSeq GenBank (Complete) genomes and in the 1,294 draft Australian genomes (Draft). All the extrachromosomal replicons of the complete genomes were perfectly identified as plasmid elements by PlaScope. Hence, results based on the extrachromosomal replicons or on the contigs detected as plasmid by PlaScope were identical (Complete*). The average number of contigs was eight times larger in draft genomes than in complete genomes (15.4 vs 1.9) and reached up to 124 contigs. B. Boxplot and histogram of the size of the contigs detected as plasmid in complete and draft genomes. These distributions were significantly different (Wilcoxon test, $P<10^{-4}$). On average the contigs were almost 10 times larger in the complete genomes than in draft genomes (81 kb vs. 8.9 kb). We identified 2347, 562 and 53 contigs larger than 20, 50 and 100 kb, resp. (A-B) showed that plasmid elements were poorly assembled and probably divided into several small contigs. C. Boxplot of the fraction of the proteome encoding plasmid elements per genome (*i.e.*, the cumulative number of proteins located on contigs classified as plasmid divided by the total number of proteins of the genome) in complete and draft genomes. These distributions were similar (Wilcoxon

test, P>0.1) with an average of 3.2% in both.
(EPS)

S13 Fig. General genomic characteristics of the mobilome of Australian *E. coli*. Three types of MGEs were detected, *i.e*., prophage (left column), plasmid (middle columns) and IS elements (right column). A. Histogram and boxplot of genomic features of each type of MGEs, *i.e*, the cumulative size of the elements per genome (Kb), the total number (#) of genes encoded by the elements per genome, the fraction of the genome encoding these elements per genome. For each case, the dash line corresponds to the smoothed curve, the red arrow to the median and the blue arrow to the average of each distribution. B. Histogram and boxplot of the number of conjugation systems per genome. C. Number of conjugative systems: (MPF) and isolated relaxases (MOB) detected in our dataset. The different MPF types were indicated and also their genomic location, *i.e*., located on a contig classified as plasmid or as chromosome by PlaScope.
(EPS)

S14 Fig. Contribution of MGEs to genome size variation. A. Association between the genome size (*i.e*., # of genes per genome) and the total number of genes associated to the MGE elements. B. Histogram and boxplot of the genome size (in grey), and of the genome size without MGE (in red), i.e., after removing all the genes encoding MGE elements (in red). These distributions were significantly different (Wilcoxon test, $P<10^{-4}$). C. Same representation as in (a), but distinguishing the different types of MGEs, i.e., prophage, plasmid and IS elements. (A-C) We found a strong correlation in each case. Linear regression (dash red line) and statistics were reported. Similar results were obtained with the genome size (Mb). D. Number of singletons (in green) and accessory gene families encoding MGEs. The fraction of the pangenome encoding such elements was reported in each case (%).
(EPS)

S15 Fig. Distribution of gene families related to MGEs across phylogroups and sources. Number of accessory gene families associated to prophage and plasmid present in one (i.e., phylogroup-specific) to seven phylogroups (A), or in one (i.e., source specific) to seven sources (B). The Z-score obtained for the observed number with respect to the expected distribution (as in Fig 5A, we randomized 1,000 times, only the phylogroup (A) or the source (B) assignment of genomes) was reported for each case with a color code ranging from blue (under-representation) to red (over-representation). The frequency of these families (average number of genomes) was also indicated in (C) for phylogroups, and in (D) for sources.
(EPS)

S16 Fig. Network of recent co-occurence of gains (co-gains) of MGE genes within and between phylogroups. Nodes are phylogroups and edges the O/E ratio of the number of pairs of MGE genes (from the same gene family) acquired in the terminal branches of the tree. Only significant O/E values (and edges) are plottted (|Z-score|>1.96). Under-represented values are in dash blue and over-represented in red (see Methods).
(EPS)

S17 Fig. Genome size and MGE content according to sources within each phylogroup. A. Heatmap of the average genome size of strains from different sources in each phylogroup. The deviation to the overall intra-phylogroup mean (i.e., the average genome size of all strains belonging to a given phylogroup) was reported for all comparisons with a color code ranging from blue (below average) to red (above average). The level of significance of each ANOM test was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). It was performed within

each phylogroup (each line). (B-C-D) Same representation as in (A), but in relation with the average number of genes associated to MGEs (B), to prophage (C), or plasmid elements (D). (EPS)

S18 Fig. Association of integrons and ARGs with human (or domesticated animals). A. Violin plots of the number of ARGs in genomes encoding integron-integrase (*int1*+) or not (*int1*-). The level of significance of the Wilcoxon test was indicated ($P<10^{-3}$). B. Heatmap of the proportion of genomes *int1*+ in each phylogroup and source. A cross marks the absence of data. C. Same as in (B) but we merged sources related to human activity (with), or not directly associated to human (without). The level of significance of each ANOM for proportions test was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). Here, we compared response proportions for the X levels to the overall response proportion from the contingency table. This method uses the normal approximation to the binomial. Therefore, in some cases sample sizes were too small to be tested. D. Heatmap of the average number of ARGs per genome in each phylogroup and source. E. Heatmap of the average number of ARGs when we merged sources related (with) or not (without) to human activity. The level of significance of each non-parametric ANOM test (ANOM with Transformed Ranks) was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). The deviation to the overall mean (i.e., in all genomes) was reported for all comparisons with a color code ranging from blue (below average) to red (above average). The color code used was displayed in the top of each panel. (EPS)

S19 Fig. Distribution of VFs and Colicins MGEs across phylogroups and sources. (A-B). Heatmap of the average number of VFs per strain from different sources in each phylogroup. The deviation to the overall mean (i.e., whole dataset, in A) or to the intra-phylogroup mean (i.e., the average number of all strains belonging to a given phylogroup, in B) was reported for all comparisons with a color code ranging from blue (below average) to red (above average). The level of significance of each ANOM test was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). It was performed within each phylogroup (each line, in B). C. Heatmap of the average number of Colicins per genome in each phylogroup and source. D. Same representation as in (B), but in relation with the average number of Colicins per genome. (EPS)

S20 Fig. Distribution of capsule systems across phylogroups and sources. A. Heatmap of the average number of capsule systems per genome in each phylogroup and source. B. The deviation to the intra-phylogroup mean (i.e., the average number of all strains belonging to a given phylogroup) was reported for all comparisons with a color code ranging from blue (below average) to red (above average). The level of significance of each ANOM test was indicated (P> = 0.05 : ns; P<0.05 : *; P<0.01 : **; P<0.001 :***). It was performed within each phylogroup (each line). C. Prevalence (%) of each capsule groups across phylogroups and sources. (EPS)

Acknowledgments

This work used the computational and storage services (TARS cluster) provided by the IT department at Pasteur Institute, Paris.

Author Contributions

Conceptualization: Marie Touchon, Erick Denamur, David Gordon, Eduardo PC Rocha.

Data curation: Marie Touchon, Amandine Perrin, David Gordon.

Formal analysis: Marie Touchon, Eduardo PC Rocha.

Funding acquisition: Marie Touchon, Erick Denamur, David Gordon, Eduardo PC Rocha.

Investigation: Marie Touchon, Eduardo PC Rocha.

Methodology: Marie Touchon, Eduardo PC Rocha.

Project administration: Marie Touchon, Eduardo PC Rocha.

Resources: Marie Touchon, Amandine Perrin, Jorge André Moura de Sousa, Belinda Vangchhia, Samantha Burn, Claire L. O'Brien, Erick Denamur, David Gordon, Eduardo PC Rocha.

Software: Marie Touchon, Amandine Perrin, Jorge André Moura de Sousa, Eduardo PC Rocha.

Supervision: Marie Touchon, Eduardo PC Rocha.

Validation: Marie Touchon, Eduardo PC Rocha.

Visualization: Marie Touchon.

Writing – original draft: Marie Touchon, Eduardo PC Rocha.

Writing – review & editing: Marie Touchon, Amandine Perrin, Jorge André Moura de Sousa, Belinda Vangchhia, Samantha Burn, Claire L. O'Brien, Erick Denamur, David Gordon, Eduardo PC Rocha.

References

1. Berg RD. The indigenous gastrointestinal microflora. Trends Microbiol. 1996; 4(11):430–5. https://doi.org/10.1016/0966-842x(96)10057-3 PMID: 8950812.
2. Gordon DM, Cowling A. The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects. Microbiology. 2003; 149(Pt 12):3575–86. https://doi.org/10.1099/mic.0.26486-0 PMID: 14663089.
3. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat Rev Microbiol. 2010; 8(3):207–17. https://doi.org/10.1038/nrmicro2298 PMID: 20157339.
4. Ishii S, Ksoll WB, Hicks RE, Sadowsky MJ. Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. Appl Environ Microbiol. 2006; 72(1):612–21. https://doi.org/10.1128/AEM.72.1.612-621.2006 PMID: 16391098; PubMed Central PMCID: PMC1352292.
5. Ishii S, Sadowsky MJ. Escherichia coli in the Environment: Implications for Water Quality and Human Health. Microbes Environ. 2008; 23(2):101–8. https://doi.org/10.1264/jsme2.23.101 PMID: 21558695.
6. van Elsas JD, Semenov AV, Costa R, Trevors JT. Survival of Escherichia coli in the environment: fundamental and public health aspects. ISME J. 2011; 5(2):173–83. https://doi.org/10.1038/ismej.2010.80 PMID: 20574458; PubMed Central PMCID: PMC3105702.
7. Berthe T, Ratajczak M, Clermont O, Denamur E, Petit F. Evidence for coexistence of distinct Escherichia coli populations in various aquatic environments and their survival in estuary water. Appl Environ Microbiol. 2013; 79(15):4684–93. https://doi.org/10.1128/AEM.00698-13 PMID: 23728810; PubMed Central PMCID: PMC3719502.
8. Donnenberg MS. Escherichia coli: virulence mechanisms of a versatile pathogen. New York: Academic Press, New York; 2002.
9. Kaper JB, Nataro JP, Mobley HL. Pathogenic Escherichia coli. Nat Rev Microbiol. 2004; 2(2):123–40. https://doi.org/10.1038/nrmicro818 PMID: 15040260.
10. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol. 2006; 60(5):1136–51. https://doi.org/10.1111/j.1365-2958.2006.05172.x PMID: 16689791; PubMed Central PMCID: PMC1557465.
11. Croxen MA, Finlay BB. Molecular mechanisms of Escherichia coli pathogenicity. Nat Rev Microbiol. 2010; 8(1):26–38. https://doi.org/10.1038/nrmicro2265 PMID: 19966814.

12. Leimbach A, Hacker J, Dobrindt U. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. Curr Top Microbiol Immunol. 2013; 358:3–32. https://doi.org/10.1007/82_2012_303 PMID: 23340801.

13. Gomes TA, Elias WP, Scaletsky IC, Guth BE, Rodrigues JF, Piazza RM, et al. Diarrheagenic Escherichia coli. Braz J Microbiol. 2016; 47 Suppl 1:3–30. https://doi.org/10.1016/j.bjm.2016.10.015 PMID: 27866935; PubMed Central PMCID: PMC5156508.

14. Vila J, Saez-Lopez E, Johnson JR, Romling U, Dobrindt U, Canton R, et al. Escherichia coli: an old friend with new tidings. FEMS Microbiol Rev. 2016; 40(4):437–63. https://doi.org/10.1093/femsre/fuw005 PMID: 28201713.

15. Cassini A, Hogberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. Lancet Infect Dis. 2019; 19(1):56–66. https://doi.org/10.1016/S1473-3099(18)30605-4 PMID: 30409683; PubMed Central PMCID: PMC6300481.

16. Chaudhuri RR, Henderson IR. The evolution of the Escherichia coli phylogeny. Infect Genet Evol. 2012; 12(2):214–26. https://doi.org/10.1016/j.meegid.2012.01.005 PMID: 22266241.

17. Ochman H, Selander RK. Standard reference strains of Escherichia coli from natural populations. J Bacteriol. 1984; 157(2):690–3. PMID: 6363394; PubMed Central PMCID: PMC215307.

18. Didelot X, Meric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012; 13:256. https://doi.org/10.1186/1471-2164-13-256 PMID: 22712577; PubMed Central PMCID: PMC3505186.

19. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci U S A. 2015; 112(29):9070–5. https://doi.org/10.1073/pnas.1510839112 PMID: 26153419; PubMed Central PMCID: PMC4517234.

20. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microb Genom. 2018; 4(7). https://doi.org/10.1099/mgen.0.000192 PMID: 29916797; PubMed Central PMCID: PMC6113867.

21. Lu S, Jin D, Wu S, Yang J, Lan R, Bai X, et al. 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. Emerg Microbes Infect. 2016; 5(12):e122. https://doi.org/10.1038/emi.2016.122 PMID: 27924811; PubMed Central PMCID: PMC5180367.

22. Clermont O, Dixit OVA, Vangchhia B, Condamine B, Bridier-Nahmias A, Denamur E, et al. Characterisation and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. Environ Microbiol. 2019. https://doi.org/10.1111/1462-2920.14713 PMID: 31188527.

23. Bergthorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol. 1998; 15(1):6–16. https://doi.org/10.1093/oxfordjournals.molbev.a025847 PMID: 9491600.

24. Escobar-Paramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, et al. Identification of forces shaping the commensal Escherichia coli genetic structure by comparing animal and human isolates. Environ Microbiol. 2006; 8(11):1975–84. https://doi.org/10.1111/j.1462-2920.2006.01077.x PMID: 17014496.

25. Vollmerhausen TL, Ramos NL, Gundogdu A, Robinson W, Brauner A, Katouli M. Population structure and uropathogenic virulence-associated genes of faecal Escherichia coli from healthy young and elderly adults. J Med Microbiol. 2011; 60(Pt 5):574–81. https://doi.org/10.1099/jmm.0.027037-0 PMID: 21292854.

26. Smati M, Clermont O, Bleibtreu A, Fourreau F, David A, Daubie AS, et al. Quantitative analysis of commensal Escherichia coli populations reveals host-specific enterotypes at the intra-species level. Microbiologyopen. 2015; 4(4):604–15. https://doi.org/10.1002/mbo3.266 PMID: 26033772; PubMed Central PMCID: PMC4554456.

27. Bok E, Mazurek J, Myc A, Stosik M, Wojciech M, Baldy-Chudzik K. Comparison of Commensal Escherichia coli Isolates from Adults and Young Children in Lubuskie Province, Poland: Virulence Potential, Phylogeny and Antimicrobial Resistance. Int J Environ Res Public Health. 2018;15(4). https://doi.org/10.3390/ijerph15040617 PMID: 29597292; PubMed Central PMCID: PMC5923659.

28. Gordon DM, Stern SE, Collignon PJ. Influence of the age and sex of human hosts on the distribution of Escherichia coli ECOR groups and virulence traits. Microbiology. 2005; 151(Pt 1):15–23. https://doi.org/10.1099/mic.0.27425-0 PMID: 15632421.

29. Escobar-Paramo P, Grenet K, Le Menac'h A, Rode L, Salgado E, Amorin C, et al. Large-scale population structure of human commensal Escherichia coli isolates. Appl Environ Microbiol. 2004; 70

(9):5698–700. https://doi.org/10.1128/AEM.70.9.5698-5700.2004 PMID: 15345464; PubMed Central PMCID: PMC520916.

30. Skurnik D, Bonnet D, Bernede-Bauduin C, Michel R, Guette C, Becker JM, et al. Characteristics of human intestinal Escherichia coli with changing environments. Environ Microbiol. 2008; 10(8):2132–7. https://doi.org/10.1111/j.1462-2920.2008.01636.x PMID: 18459976.

31. Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventre A, Elion J, et al. Commensal Escherichia coli isolates are phylogenetically distributed among geographically distinct human populations. Microbiology. 2001; 147(Pt 6):1671–6. https://doi.org/10.1099/00221287-147-6-1671 PMID: 11390698.

32. Power ML, Littlefield-Wyer J, Gordon DM, Veal DA, Slade MB. Phenotypic and genotypic characterization of encapsulated Escherichia coli isolated from blooms in two Australian lakes. Environ Microbiol. 2005; 7(5):631–40. https://doi.org/10.1111/j.1462-2920.2005.00729.x PMID: 15819845.

33. Walk ST, Alm EW, Calhoun LM, Mladonicky JM, Whittam TS. Genetic diversity and population structure of Escherichia coli isolated from freshwater beaches. Environ Microbiol. 2007; 9(9):2274–88. https://doi.org/10.1111/j.1462-2920.2007.01341.x PMID: 17686024.

34. Ratajczak M, Laroche E, Berthe T, Clermont O, Pawlak B, Denamur E, et al. Influence of hydrological conditions on the Escherichia coli population structure in the water of a creek on a rural watershed. BMC Microbiol. 2010; 10:222. https://doi.org/10.1186/1471-2180-10-222 PMID: 20723241; PubMed Central PMCID: PMC2933670.

35. Anastasi EM, Matthews B, Stratton HM, Katouli M. Pathogenic Escherichia coli found in sewage treatment plants and environmental waters. Appl Environ Microbiol. 2012; 78(16):5536–41. https://doi.org/10.1128/AEM.00657-12 PMID: 22660714; PubMed Central PMCID: PMC3406122.

36. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, et al. The link between phylogeny and virulence in Escherichia coli extraintestinal infection. Infect Immun. 1999; 67(2):546–53. PMID: 9916057; PubMed Central PMCID: PMC96353.

37. Johnson JR, Delavari P, Kuskowski M, Stell AL. Phylogenetic distribution of extraintestinal virulence-associated traits in Escherichia coli. J Infect Dis. 2001; 183(1):78–88. https://doi.org/10.1086/317656 PMID: 11106538.

38. Moulin-Schouleur M, Reperant M, Laurent S, Bree A, Mignon-Grasteau S, Germon P, et al. Extraintestinal pathogenic Escherichia coli strains of avian and human origin: link between phylogenetic relationships and common virulence patterns. J Clin Microbiol. 2007; 45(10):3366–76. https://doi.org/10.1128/JCM.00037-07 PMID: 17652485; PubMed Central PMCID: PMC2045314.

39. Riley LW. Pandemic lineages of extraintestinal pathogenic Escherichia coli. Clin Microbiol Infect. 2014; 20(5):380–90. https://doi.org/10.1111/1469-0691.12646 PMID: 24766445.

40. Stoppe NC, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, et al. Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates. Front Microbiol. 2017; 8:2512. https://doi.org/10.3389/fmicb.2017.02512 PMID: 29312213; PubMed Central PMCID: PMC5742620.

41. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. 2008; 190(20):6881–93. https://doi.org/10.1128/JB.00619-08 PMID: 18676672; PubMed Central PMCID: PMC2566221.

42. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009; 5(1):e1000344. https://doi.org/10.1371/journal.pgen.1000344 PMID: 19165319; PubMed Central PMCID: PMC2617782.

43. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 2010; 60(4):708–20. https://doi.org/10.1007/s00248-010-9717-3 PMID: 20623278; PubMed Central PMCID: PMC2974192.

44. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015; 15(2):141–61. https://doi.org/10.1007/s10142-015-0433-4 PMID: 25722247; PubMed Central PMCID: PMC4361730.

45. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, et al. Global dissemination of a multidrug resistant Escherichia coli clone. Proc Natl Acad Sci U S A. 2014; 111(15):5694–9. https://doi.org/10.1073/pnas.1322678111 PMID: 24706808; PubMed Central PMCID: PMC3992628.

46. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008; 11(5):472–7. https://doi.org/10.1016/j.mib.2008.09.006 PMID: 19086349.

47. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral

sequences. Nucleic Acids Res. 2016; 44(D1):D286–93. https://doi.org/10.1093/nar/gkv1248 PMID: 26582926; PubMed Central PMCID: PMC4702882.

48. Patel IR, Gangiredla J, Mammel MK, Lampel KA, Elkins CA, Lacher DW. Draft Genome Sequences of the Escherichia coli Reference (ECOR) Collection. Microbiol Resour Announc. 2018; 7(14). https://doi.org/10.1128/MRA.01133-18 PMID: 30533715; PubMed Central PMCID: PMC6256646.
49. Gonzalez-Alba JM, Baquero F, Canton R, Galan JC. Stratified reconstruction of ancestral Escherichia coli diversification. BMC Genomics. 2019; 20(1):936. https://doi.org/10.1186/s12864-019-6346-1 PMID: 31805853; PubMed Central PMCID: PMC6896753.
50. Wagner A, Lewis C, Bichsel M. A survey of bacterial insertion sequences using IScan. Nucleic Acids Res. 2007; 35(16):5284–93. https://doi.org/10.1093/nar/gkm597 PMID: 17686783; PubMed Central PMCID: PMC2018620.
51. Touchon M, Rocha EP. Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol. 2007; 24(4):969–81. https://doi.org/10.1093/molbev/msm014 PMID: 17251179.
52. Bobay LM, Touchon M, Rocha EP. Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci U S A. 2014; 111(33):12127–32. https://doi.org/10.1073/pnas.1405336111 PMID: 25092302; PubMed Central PMCID: PMC4143005.
53. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015; 3:e985. https://doi.org/10.7717/peerj.985 PMID: 26038737; PubMed Central PMCID: PMC4451026.
54. Royer G, Decousser JW, Branger C, Dubois M, Medigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. Microb Genom. 2018; 4(9). https://doi.org/10.1099/mgen.0.000211 PMID: 30265232.
55. Guglielmini J, Neron B, Abby SS, Garcillan-Barcia MP, de la Cruz F, Rocha EP. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 2014; 42(9):5715–27. https://doi.org/10.1093/nar/gku194 PMID: 24623814; PubMed Central PMCID: PMC4027160.
56. Cury J, Oliveira PH, de la Cruz F, Rocha EPC. Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. Mol Biol Evol. 2018; 35 (11):2850. https://doi.org/10.1093/molbev/msy182 PMID: 30418640; PubMed Central PMCID: PMC6231490.
57. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006; 34(Database issue):D32–6. https://doi.org/10.1093/nar/gkj014 PMID: 16381877; PubMed Central PMCID: PMC1347377.
58. Cury J, Jove T, Touchon M, Neron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. Nucleic Acids Res. 2016; 44(10):4539–50. https://doi.org/10.1093/nar/gkw319 PMID: 27130947; PubMed Central PMCID: PMC4889954.
59. Domingues S, da Silva GJ, Nielsen KM. Integrons: Vehicles and pathways for horizontal dissemination in bacteria. Mob Genet Elements. 2012; 2(5):211–23. https://doi.org/10.4161/mge.22967 PMID: 23550063; PubMed Central PMCID: PMC3575428.
60. Vieira G, Sabarly V, Bourguignon PY, Durot M, Le Fevre F, Mornico D, et al. Core and panmetabolism in Escherichia coli. J Bacteriol. 2011; 193(6):1461–72. https://doi.org/10.1128/JB.01192-10 PMID: 21239590; PubMed Central PMCID: PMC3067614.
61. Sabarly V, Aubron C, Glodt J, Balliau T, Langella O, Chevret D, et al. Interactions between genotype and environment drive the metabolic phenotype within Escherichia coli isolates. Environ Microbiol. 2016; 18(1):100–17. https://doi.org/10.1111/1462-2920.12855 PMID: 25808978.
62. Diaz E, Ferrandez A, Prieto MA, Garcia JL. Biodegradation of aromatic compounds by Escherichia coli. Microbiol Mol Biol Rev. 2001; 65(4):523–69, table of contents. https://doi.org/10.1128/MMBR.65.4.523-569.2001 PMID: 11729263; PubMed Central PMCID: PMC99040.
63. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, et al. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci U S A. 2013; 110(50):20338–43. https://doi.org/10.1073/pnas.1307797110 PMID: 24277855; PubMed Central PMCID: PMC3864276.
64. Meric G, Kemsley EK, Falush D, Saggers EJ, Lucchini S. Phylogenetic distribution of traits associated with plant colonization in Escherichia coli. Environ Microbiol. 2013; 15(2):487–501. https://doi.org/10.1111/j.1462-2920.2012.02852.x PMID: 22934605.
65. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res. 2016; 44(D1):D694–7. https://doi.org/10.1093/nar/gkv1239 PMID: 26578559; PubMed Central PMCID: PMC4702877.

66. Cascales E, Buchanan SK, Duche D, Kleanthous C, Lloubes R, Postle K, et al. Colicin biology. Microbiol Mol Biol Rev. 2007; 71(1):158–229. https://doi.org/10.1128/MMBR.00036-06 PMID: 17347522; PubMed Central PMCID: PMC1847374.

67. van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res. 2013; 41(Web Server issue):W448–53. https://doi.org/10.1093/nar/gkt391 PMID: 23677608; PubMed Central PMCID: PMC3692055.

68. Jang J, Hur HG, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. Environmental Escherichia coli: ecology and public health implications-a review. J Appl Microbiol. 2017; 123(3):570–81. https://doi.org/10.1111/jam.13468 PMID: 28383815.

69. Hazen TH, Leonard SR, Lampel KA, Lacher DW, Maurelli AT, Rasko DA. Investigating the Relatedness of Enteroinvasive Escherichia coli to Other E. coli and Shigella Isolates by Using Comparative Genomics. Infect Immun. 2016; 84(8):2362–71. https://doi.org/10.1128/IAI.00350-16 PMID: 27271741; PubMed Central PMCID: PMC4962626.

70. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, et al. Evolutionary History of the Global Emergence of the Escherichia coli Epidemic Clone ST131. MBio. 2016; 7(2):e02162. https://doi.org/10.1128/mBio.02162-15 PMID: 27006459; PubMed Central PMCID: PMC4807372.

71. Shaik S, Ranjan A, Tiwari SK, Hussain A, Nandanwar N, Kumar N, et al. Comparative Genomic Analysis of Globally Dominant ST131 Clone with Other Epidemiologically Successful Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. MBio. 2017; 8(5). https://doi.org/10.1128/mBio.01596-17 PMID: 29066550; PubMed Central PMCID: PMC5654935.

72. Gordon DM, Geyik S, Clermont O, O'Brien CL, Huang S, Abayasekara C, et al. Fine-Scale Structure Analysis Shows Epidemic Patterns of Clonal Complex 95, a Cosmopolitan Escherichia coli Lineage Responsible for Extraintestinal Infection. mSphere. 2017; 2(3). https://doi.org/10.1128/mSphere.00168-17 PMID: 28593194; PubMed Central PMCID: PMC5451516.

73. Johnson TJ, Elnekave E, Miller EA, Munoz-Aguayo J, Flores Figueroa C, Johnston B, et al. Phylogenomic Analysis of Extraintestinal Pathogenic Escherichia coli Sequence Type 1193, an Emerging Multidrug-Resistant Clonal Group. Antimicrob Agents Chemother. 2019;63(1). https://doi.org/10.1128/AAC.01913-18 PMID: 30348668; PubMed Central PMCID: PMC6325179.

74. Jorgensen SL, Stegger M, Kudirkiene E, Lilje B, Poulsen LL, Ronco T, et al. Diversity and Population Overlap between Avian and Human Escherichia coli Belonging to Sequence Type 95. mSphere. 2019; 4(1). https://doi.org/10.1128/mSphere.00333-18 PMID: 30651401; PubMed Central PMCID: PMC6336079.

75. Dobrindt U, Chowdary MG, Krumbholz G, Hacker J. Genome dynamics and its impact on evolution of Escherichia coli. Med Microbiol Immunol. 2010; 199(3):145–54. https://doi.org/10.1007/s00430-010-0161-2 PMID: 20445988.

76. Juhas M. Horizontal gene transfer in human pathogens. Crit Rev Microbiol. 2015; 41(1):101–8. https://doi.org/10.3109/1040841X.2013.804031 PMID: 23862575.

77. Stokes HW, Gillings MR. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. FEMS Microbiol Rev. 2011; 35(5):790–819. https://doi.org/10.1111/j.1574-6976.2011.00273.x PMID: 21517914.

78. von Wintersdorff CJ, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front Microbiol. 2016; 7:173. https://doi.org/10.3389/fmicb.2016.00173 PMID: 26925045; PubMed Central PMCID: PMC4759269.

79. Goldstone RJ, Smith DGE. A population genomics approach to exploiting the accessory 'resistome' of Escherichia coli. Microb Genom. 2017; 3(4):e000108. https://doi.org/10.1099/mgen.0.000108 PMID: 28785420; PubMed Central PMCID: PMC5506381.

80. Frazao N, Sousa A, Lassig M, Gordo I. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. Proc Natl Acad Sci U S A. 2019; 116(36):17906–15. https://doi.org/10.1073/pnas.1906958116 PMID: 31431529; PubMed Central PMCID: PMC6731689.

81. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics. 2012; 13:577. https://doi.org/10.1186/1471-2164-13-577 PMID: 23114024; PubMed Central PMCID: PMC3575317.

82. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, Pitout JDD. Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. Clin Microbiol Rev. 2019;32(3). https://doi.org/10.1128/CMR.00135-18 PMID: 31189557; PubMed Central PMCID: PMC6589867.

83. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. 2012; 29(11):3413–25. https://doi.org/10.1093/molbev/mss163 PMID: 22752048.

84. Wolf YI, Makarova KS, Lobkovsky AE, Koonin EV. Two fundamentally different classes of microbial genes. Nat Microbiol. 2016; 2:16208. https://doi.org/10.1038/nmicrobiol.2016.208 PMID: 27819663.

85. Paul JH. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME J. 2008; 2(6):579–89. https://doi.org/10.1038/ismej.2008.35 PMID: 18521076.

86. Bichsel M, Barbour AD, Wagner A. Estimating the fitness effect of an insertion sequence. J Math Biol. 2013; 66(1–2):95–114. https://doi.org/10.1007/s00285-012-0504-2 PMID: 22252506.

87. San Millan A, MacLean RC. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. Microbiol Spectr. 2017;5(5). https://doi.org/10.1128/microbiolspec.MTBP-0016-2017 PMID: 28944751.

88. Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends Genet. 2001; 17(10):589–96. https://doi.org/10.1016/s0168-9525(01)02447-7 PMID: 11585665.

89. Lawrence JG, Hendrix RW, Casjens S. Where are the pseudogenes in bacterial genomes? Trends Microbiol. 2001; 9(11):535–40. https://doi.org/10.1016/s0966-842x(01)02198-9 PMID: 11825713.

90. van Houte S, Buckling A, Westra ER. Evolutionary Ecology of Prokaryotic Immune Mechanisms. Microbiol Mol Biol Rev. 2016; 80(3):745–63. https://doi.org/10.1128/MMBR.00011-16 PMID: 27412881; PubMed Central PMCID: PMC4981670.

91. Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ, Turner P, et al. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. Mol Biol Evol. 2017; 34(10):2537–54. https://doi.org/10.1093/molbev/msx173 PMID: 28595308; PubMed Central PMCID: PMC5850285.

92. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. ISME J. 2016; 10(11):2744–54. https://doi.org/10.1038/ismej.2016.47 PMID: 27015004; PubMed Central PMCID: PMC5113838.

93. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol. 1997; 23(6):1089–97. https://doi.org/10.1046/j.1365-2958.1997.3101672.x PMID: 9106201.

94. Penades JR, Chen J, Quiles-Puchalt N, Carpena N, Novick RP. Bacteriophage-mediated spread of bacterial virulence genes. Curr Opin Microbiol. 2015; 23:171–8. https://doi.org/10.1016/j.mib.2014.11.019 PMID: 25528295.

95. Touchon M, Bobay LM, Rocha EP. The chromosomal accommodation and domestication of mobile genetic elements. Curr Opin Microbiol. 2014; 22:22–9. https://doi.org/10.1016/j.mib.2014.09.010 PMID: 25305534.

96. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. Ecology drives a global network of gene exchange connecting the human microbiome. Nature. 2011; 480(7376):241–4. https://doi.org/10.1038/nature10571 PMID: 22037308.

97. Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, Jenkins AP, et al. Mobile genes in the human microbiome are structured from global to individual scales. Nature. 2016; 535(7612):435–9. https://doi.org/10.1038/nature18927 PMID: 27409808; PubMed Central PMCID: PMC4983458.

98. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol. 2014; 12(12):841–50. https://doi.org/10.1038/nrmicro3331 PMID: 25220308.

99. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. Nat Microbiol. 2016; 2:16198. https://doi.org/10.1038/nmicrobiol.2016.198 PMID: 27798560.

100. O'Brien CL, Bringer MA, Holt KE, Gordon DM, Dubois AL, Barnich N, et al. Comparative genomics of Crohn's disease-associated adherent-invasive Escherichia coli. Gut. 2017; 66(8):1382–9. https://doi.org/10.1136/gutjnl-2015-311059 PMID: 27196580.

101. Blyton MD, Gordon DM. Genetic Attributes of E. coli Isolates from Chlorinated Drinking Water. PLoS One. 2017; 12(1):e0169445. https://doi.org/10.1371/journal.pone.0169445 PMID: 28107364.

102. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14):2068–9. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063.

103. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, et al. In silico serotyping of E. coli from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microb Genom. 2016; 2(7):e000064. https://doi.org/10.1099/mgen.0.000064 PMID: 28348859; PubMed Central PMCID: PMC5343136.

104. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 2014; 31(7):1929–36. https://doi.org/10.1093/molbev/msu136 PMID: 24739305; PubMed Central PMCID: PMC4069620.

105. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009; 106(45):19126–31. https://doi.org/10.1073/pnas.0906412106 PMID: 19855009; PubMed Central PMCID: PMC2776425.

106. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17(1):132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842; PubMed Central PMCID: PMC4915045.

107. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017; 35(11):1026–8. https://doi.org/10.1038/nbt.3988 PMID: 29035372.

108. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018; 9 (1):2542. https://doi.org/10.1038/s41467-018-04964-5 PMID: 29959318; PubMed Central PMCID: PMC6026198.

109. Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. BMC Bioinformatics. 2015; 16:79. https://doi.org/10.1186/s12859-015-0517-0 PMID: 25888166; PubMed Central PMCID: PMC4375852.

110. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018; 34(14):2490–2. https://doi.org/10.1093/bioinformatics/bty121 PMID: 29506019; PubMed Central PMCID: PMC6041967.

111. Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2008; 4(5):e1000069. https://doi.org/10.1371/journal.pcbi.1000069 PMID: 18516236; PubMed Central PMCID: PMC2396288.

112. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361; PubMed Central PMCID: PMC3197634.

113. Filipski A, Murillo O, Freydenzon A, Tamura K, Kumar S. Prospects for building large timetrees using molecular data with incomplete gene coverage among species. Mol Biol Evol. 2014; 31(9):2542–50. https://doi.org/10.1093/molbev/msu200 PMID: 24974376; PubMed Central PMCID: PMC4137717.

114. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. mBio. 2014; 5(6):e02158. https://doi.org/10.1128/mBio.02158-14 PMID: 25425237; PubMed Central PMCID: PMC4251999.

115. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EP. The Impact of Selection, Gene Conversion, and Biased Sampling on the Assessment of Microbial Demography. Mol Biol Evol. 2016; 33(7):1711–25. https://doi.org/10.1093/molbev/msw048 PMID: 26931140; PubMed Central PMCID: PMC4915353.

116. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430; PubMed Central PMCID: PMC4271533.

117. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018; 35(2):518–22. https://doi.org/10.1093/molbev/msx281 PMID: 29077904; PubMed Central PMCID: PMC5850222.

118. Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. Genome sequencing of environmental Escherichia coli expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A. 2011; 108(17):7200–5. https://doi.org/10.1073/pnas.1015622108 PMID: 21482770; PubMed Central PMCID: PMC3084108.

119. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty633 PMID: 30016406.

120. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet. 1999; 21(1):108–10. https://doi.org/10.1038/5052 PMID: 9916801.

121. Csuros M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics. 2010; 26(15):1910–2. https://doi.org/10.1093/bioinformatics/btq315 PMID: 20551134.

122. Oliveira PH, Touchon M, Rocha EP. Regulation of genetic flux between bacteria by restriction-modification systems. Proc Natl Acad Sci U S A. 2016; 113(20):5658–63. https://doi.org/10.1073/pnas.1603257113 PMID: 27140615; PubMed Central PMCID: PMC4878467.

123. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016; 17(1):238. https://doi.org/10.1186/s13059-016-1108-8 PMID: 27887642; PubMed Central PMCID: PMC5124306.

124. Draper NR SH. Applied Regression Analysis. York JWSN, editor1998.

125. Abby SS, Neron B, Menager H, Touchon M, Rocha EP. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One. 2014; 9(10):e110726. https://doi.org/10.1371/journal.pone.0110726 PMID: 25330359; PubMed Central PMCID: PMC4201578.

126. Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet. 2011; 7(8): e1002222. https://doi.org/10.1371/journal.pgen.1002222 PMID: 21876676; PubMed Central PMCID: PMC3158045.

127. Rendueles O, Garcia-Garcera M, Neron B, Touchon M, Rocha EPC. Abundance and co-occurrence of extracellular capsules increase environmental breadth: Implications for the emergence of pathogens. PLoS Pathog. 2017; 13(7):e1006525. https://doi.org/10.1371/journal.ppat.1006525 PMID: 28742161; PubMed Central PMCID: PMC5542703.

128. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012; 67(11):2640–4. https://doi.org/10.1093/jac/dks261 PMID: 22782487; PubMed Central PMCID: PMC3468078.

129. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother. 2014; 58(1):212–20. https://doi.org/10.1128/AAC.01310-13 PMID: 24145532; PubMed Central PMCID: PMC3910750.

8 Population structure of carbapenemase-producing *Morganella* species

I am currently participating to the analysis of the population of *Morganella morganii* species. Very little studied so far, there is an increasing interest for this opportunistic pathogen, as it is naturally resistant to many antibiotics. This study is in collaboration with the French National Reference Center for Antibiotic Resistance dedicated to Carbapenemase-Producing Enterobacteriaceae. My contribution to the study is on the bioinformatics side.

In order to analyse the whole species, I started by downloading all *Morganella* sequences available in Refseq. As downloading reference genomes is quite recurrent while doing comparative genomics, we decided to add this functionality to PanACoTA. With the `prepare` module, it is now possible to automatically download all available sequences in refseq or genbank of a given genera, species or strain.

The `prepare` module provides, in addition to this downloading step, tools to filter the dataset, mainly based on the alignment-free comparison tool Mash [135]. This was very important here, as very little is known on the *Morganella morganii* species. Mash analysis helped to organize its taxonomy, and highlighted the existence of two subspecies.

In order to better understand these subspecies, my work then consisted in computing the species pangenome, as well as computing and comparing the two subspecies core genomes.

The following document is a draft for a paper which has just been submitted to Nature Microbiology.

To be submitted as original article in Nature Microbiology

Population structure of carbapenemase-producing *Morganella spp.*

Rémy A. Bonnin[1,2]*, Elodie Creton[1,2$], Amandine Perrin[3,$], Delphine Girlich[1], Agnès B. Jousset[1,2,4], Cecile Emeraud[1,2,4], Katie Hopkins[5], Pierre Bogaerts,[6] Youri Glupczynski,[6] Niels Pfennigwerth,[7] Marek Gniadkowski,[8] Antoni Hendrickx,[9] Kim van der Zwaluw,[9] Petra Apfalter,[10] Rainer Hartl, [10] Vendula Heringova,[11] Jaroslav Hrabak,[11] Gerald Larrouy-Maumus,[12] Eduardo Rocha,[3] Thierry Naas[1,2,4], Laurent Dortet[1,2,4]

[1] Team "Resist" UMR1184 "Immunology of Viral, Auto-Immune, Hematological and Bacterial diseases (IMVA-HB)," INSERM, Université Paris-Saclay, CEA, LabEx LERMIT, Faculty of Medicine, Le Kremlin-Bicêtre, France.
[2] Associated French National Reference Center for Antibiotic Resistance: Carbapenemase-Producing Enterobacteriaceae, Le Kremlin-Bicêtre, France.
[3] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris 75015, France
[4] Bacteriology-Hygiene Unit, Assistance Publique-Hôpitaux de Paris, AP-HP Paris Saclay, Bicêtre Hospital Le Kremlin-Bicêtre, France.
[5] National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Imperial College London, Hammersmith Hospital, Du Cane Road, London, Antimicrobial Resistance and Healthcare Associated Infections (AMRHAI) Reference Unit, National Infection Service, Public Health England, London, NW9 5EQ, UK
[6] National Reference Laboratory for Monitoring of Antimicrobial Resistance in Gram-Negative Bacteria, CHU Dinant-Godinne, UCL Namur, B 5530 Yvoir, Belgium.
[7] German National Reference Centre for Multidrug-Resistant Gram-Negative Bacteria, Department of Medical Microbiology, Ruhr-University Bochum, Bochum, Germany
[8] Department of Molecular Microbiology, National Medicines Institute, Warsaw, Poland
[9] Laboratory for Infectious Diseases & Screening, National Institute for Public Health & the Environment (RIVM), Bilthoven, The Netherlands
[10] National Reference Center for Antimicrobial Resistance and Nosocomial Infections, Institute for Hygiene, Microbiology and Tropical Medicine, Ordensklinikum Linz Elisabethinen, Fadingerstrasse 1, 4020 Linz, Austria.
[11] Biomedical Center, Faculty of Medicine in Pilsen, Charles University, 30100 Pilsen, Czech Republic
[12] MRC Centre for Molecular Bacteriology and Infection, Department of Life Sciences, Faculty of Natural Sciences, Imperial College London, London, United Kingdom
[$] Equal contribution

Keywords: CHDL, carbapenem, transposon, insertion sequence, antibiotic resistance, detection
Word count: 4135 words
Abstract: 150 words
Tables: 1
Figures : 4
Extended data: 10 Supplementary Figures, 3 Supplementary Tables

*** Corresponding author's mailing address:**
Dr. Rémy Bonnin

Service de Bactériologie-Hygiène, Hôpital de Bicêtre, 78 rue du Général Leclerc, 94275 Le Kremlin-Bicêtre Cedex, France.
Fax: + 33 1 45 21 63 40.
E-mail: remy.bonnin@u-psud.fr

ABSTRACT

Morganella are opportunistic pathogens involved in various infections. In *Morganella*, intrinsic resistance to multiple antibiotics including polymyxin combined with the emergence of carbapenemase-producers (CP) strongly limits the antimicrobial armamentarium.

We deeply characterized an international collection of 172 highly drug-resistant *Morganella* spp. including 145 CP. Whole genome sequencing combined with cloning, biochemical experiments, antimicrobial susceptibility testing and epidemiological data of several outbreaks allowed to decipher several mechanisms responsible for intrinsic characters of *Morganella*, such as trehalose assimilation of *Morganella sibonii* and the addition of L-Ara4N on the lipid A leading to intrinsic polymyxin resistance. We highlighted the need to refine the *Morganella* taxonomy. The CP were distributed among five "high-risk" clones inside the *Morganella morganii* subsp. *morganii*. Epidemiological data allowed to decipher that a single nucleotide polymorphism cut-off of 100 was accurate to identify outbreaks. Finally, cefepime-zidebactam and ceftazidime-avibactam were the most potent last resort antimicrobials towards CP except for metallo-β-lactamase-producers.

INTRODUCTION

Morganella morganii is a facultative anaerobic Gram-negative rod belonging to Enterobacterales, firstly reported in 1907.[1] Initially reported as *Proteus morganii,* it has been reclassified as *Morganella morganii* gen nov in 1943.[2] Currently, *Morganella* genus is composed of two species: *M. morganii* and *Morganella psychrotolerans.*[3] Whereas *M. morganii* is frequently encountered in clinical specimen, the psychrotolerant *M. psychrotolerans* is associated to seafood poisoning by production of histamine.[4] In 1992, using biochemical analysis, *M. morganii* has been divided into two subspecies based on the differential utilization of trehalose.[5] These two subspecies are the trehalose-fermentating *M. morganii* subsp. *sibonii* and the *M. morganii* subsp. *morganii* unable to metabolize trehalose.

M. morganii is an opportunistic pathogen responsible for a wide variety of infections such as urinary tract infections, septic shock, surgical site infections, osteomyelitis, and pneumonia.[6,7] In addition, except for a cluster of *M. morganii* infections that has been reported in the late 1970's,[8] no other outbreak has been studied at the microbial and genomic levels. Accordingly, the molecular epidemiology of *Morganella* spp. recovered from clinical samples has never been explored.

Morganella spp. isolates are intrinsically resistant to colistin, macrolides, fosfomycin, amoxicillin, first- and second-generation cephalosporins (due to the production of a class C β-lactamase) and possess decreased susceptibility to imipenem due to a low affinity of its penicillin binding protein PBP-2.[7] The treatment of infections caused by Enterobacterales (including *Morganella*) often involve β-lactams. However, since 1980s the dissemination of extended-spectrum β-lactamases (ESBLs) and acquired- or overproduced-cephalosporinase (AmpC) has limited the therapeutic options with carbapenems remaining the only alternative among β-lactams. Unfortunately, since 2000's, the carbapenem resistance emerged in Enterobacterales. Carbapenem resistance is due to (i) the production of ESBLs or overproduced-AmpC associated with decreased outer-membrane permeability or (ii) to the production of an enzyme with significant hydrolytic activity towards carbapenems and named carbapenemases.[9] In Enterobacterales, the main carbapenemases are Ambler class A enzymes with mainly KPC-like enzymes, metallo β-lactamases (MBLs) (Ambler class B) of NDM-, VIM- and IMP-type and Ambler class D carbapeneme-hydrolyzing β-lactamases of OXA-48-type.[10] Carbapenemase-producing *M. morganii* are rarely reported. However, the most prevalent carbapenemases produced by *M. morganii* are NDM-like enzymes.[11–14] More sporadically, OXA-48- and KPC-like enzymes have also been reported in *M. morganii*.[15–18] To our knowledge, only one report of GES-5 carbapenemase in *M. morganii* has been

published.[19] In *Morganella* spp., combining intrinsic resistance (particularly to colistin) with carbapenemase production might lead to nearly "untreatable" infections.[20]

Here, we report a deep characterization of an international collection of carbapenemase-producing *Morganella* spp. Whole genome sequencing (WGS) combined with biochemical experiments, antimicrobial susceptibility testing and epidemiological data of several outbreaks allowed not only to decipher several mechanisms responsible for intrinsic characters of *Morganella* species, such as trehalose assimilation of *M. sibonii* and natural resistance to colistin, but also to identify "high-risk" clones inside the *Morganella* genus.

RESULTS

Carbapenemase-producing *Morganella* spp. in France

From January 1st 2013 to March 1st 2021, a total of 68 non-duplicate carbapenemase-producing *M. morganii* were collected at the French NRC among 14,672 carbapenemase-producing Enterobacterales (CPE) isolates representing 0·46% of the French CPE. These isolates were recovered from 8 different french regions (**Supplementary Figure S1B**). Of note, 63·2% (43/68) of the *M. morganii* isolates were recovered from the same area in South-West of France (**Supplementary Figure S1B**). Phylogenetic analysis performed on all the 68 *M. morganii* revealed that among the 51 NDM-1-producing *M. morganii*, 80·.4% (41/51) belonged to the same clone recovered from 2013 to 2021 (**Supplementary Figure S1A and S1C**). This clone included 39 isolates collected in three different cities located 22 km and 72 km away in South-West of France (suggesting patient-to-patient cross-contamination), and 2 isolates collected from a distant area in North-East of France (**Supplementary Figure S1A**). The 41 isolates that composed this clone (clone I hereafter) were genetically very close with a median of 25 SNPs (1st quartile = 16, 3rd quartile = 37) along their whole genome.

Resistome analysis demonstrated that the main clone identified in South-West of France carried the $bla_{\text{NDM-1}}$ carbapenemase encoding gene, the $bla_{\text{CTX-M-15}}$ ESBL gene and two copies of the bla_{DHA}-like cephalosporinase gene. The first copy corresponded to the chromosome encoded $bla_{\text{DHA-4}}$ gene and the second copy was a truncated $bla_{\text{DHA-1}}$ gene deleted in its 3' end. The association of these β-lactamases encoding genes was responsible for full resistance to all β-lactams including new β-lactam-inhibitors associations (ceftazidime-avibactam, ceftolozane-tazobactam, imipenem-relebactam and meropenem-vaborbactam) (**Figure 1**). In addition, this clone I produced two 16S RNA methylase, ArmA and RmtC, and two aminoglycoside-modifying enzymes (AAC(6')-Ib and AAD-1) conferring resistance to all aminoglycosides. Resistance to quinolones was mediated by mutations in GyrA (S83I), ParC

(S84I and D313E), ParE (N84K and S459Y),[21,22] associated with the production of QnrA1. Combined with the intrinsic resistance to polymyxins and tigecycline, acquired resistance determinants were responsible to full resistance to all commonly tested molecules. Attempts to transfer the bla_{NDM-1} gene by conjugation or electrotransformation failed suggesting a chromosome location of the bla_{NDM-1} gene. Long-read sequencing of the first isolate identified in France confirmed the chromosomal location of bla_{NDM-1}, inside a novel transposon Tn*7340* described in **Supplementary Figure S2**.

Out of this clone I, the 27 remaining isolates sent to the NRC were polyclonal and carried diverse carbapenemases (10 OXA-48, one OXA-162, twelve NDM-1, one NDM-7, one VIM-4 and one NDM-1 + VIM-1 producers) (**Supplementary Figure S1A** and **S1C**).

Carbapenemase-producing *Morganella* spp. in Europe

To determine if the main clone (clone I) observed in France has already spread abroad, we analyzed additional 104 *Morganella* spp. isolates with decrease susceptibility to ertapenem or meropenem, collected from 7 reference centers across Europe from 2013 to 2021. Resistome analysis allowed the identification of a wide variety of carbapenemases produced by these isolates (**Table 1**). Of note, a wide diversity of carbapenemases was identified in Germany including NDM-1, NDM-5, VIM-1, OXA-48, OXA-181 and OXA-641 (a variant of OXA-372 reported only once in *Citrobacter freundii*).[23] In contrast, in Czech Republic KPC-2 carbapenemase was highly prevalent (16/21) followed by OXA-48 (n=1). The resistomes of all isolates are summarized in **Supplementary Table S1**.

On top of carbapenemases, several acquired β-lactamases were also acquired. As previously reported for other Enterobacterales,[24] CTX-M-15 was the most prevalent ESBL identified in 34·9% (60/172) of the *Morganella* spp. isolates, followed by CTX-M-14, CTX-M-1, SHV-12 and VEB-6-like enzymes that were identified in four, one, six and one isolates, respectively. Twelve isolates produced an acquired cephalosporinase of the CMY family (**Supplementary Table S1**).

Regarding quinolones resistance, mutations in gyrase and topoisomerase IV were observed as well as production of plasmid-mediated quinolone resistance determinants (43 *qnrA1*-like, 27 *qnrD*-like, 13 *qnrB*-like and 6 *qnrS*-like) (**Table 1**). The QepA quinolone efflux pump was identified in two strains.

Regarding acquired aminoglycosides resistance genes, the prevalence of 16S rRNA methylases encoding genes was high with 42 and 39 isolates carrying *armA*-like and *rmtC*-like genes, respectively. In addition, several other genes encoding aminoglycosides modifying

enzymes (*aac(6')-Ib*, *aac(3')-IIa*, *aadA*, *aadB*, *aph(3')-I*, *aph(3')-VI* and *aph(4')-I*) were also identified in class 1 integrons. Accordingly, they were identified in isolates co-carrying other integron-born resistance genes such as trimethoprim (*dfrA*-like), chloramphenicol (*catA*-like, *catB*-like, *cmlA*-like) or sulfamide (*sul1*) resistance determinants (**Table 1** and **Supplementary Table S1**).

Antimicrobial alternatives for the treatment of highly drug resistant *Morganella* spp.

The treatment of infections caused by carbapenem-resistant *Morganella* spp. is of great concern. Accordingly, we tested several last-resort antibiotics such as three carbapenems (imipenem, ertapenem, meropenem), new β-lactams-β-lactamase inhibitor associations (ceftazidime-avibactam, meropenem-vaborbactam, imipenem-relebactam, cefepime-zidebactam and ceftolozane-tazobactam), temocillin and two last-resort cyclins (eravacycline and tigecycline).

For each antimicrobial, MICs distributions are presented in **Figure 1** and interpreted according to EUCAST guidelines. Carbapenemase producers (n=145) were separated according to their carbapenemase class content and compared to isolates that do not produce any carbapenemase (n=32). As expected, non-carbapenemase producers showed only a moderate susceptibility to imipenem. The combination with relebactam did not significantly increase the efficiency of imipenem. Surprisingly, MIC of ertapenem remained in the susceptible range for 100% and 81·3% of Ambler class A and class D carbapenemase producers and 93·8% of non-carbapenemase producers. As expected vaborbactam helped to restore meropenem susceptibility for all KPC producers. But this association showed a moderate effect on class D and no effect on MBL-producing isolates. The distribution of MICs of ceftazidime was heterogeneous except for MBL producers that remained highly resistant. The ceftazidime-avibactam was an accurate option for the treatment of infection caused by Ambler class A or class D carbapenemase- and non-carbapenemase-producers with 100%, 95·5% and 96·9% of susceptibility respectively. As expected, all MBL-producers were fully resistant to ceftazidime-avibactam. Temocillin showed a bi-modal distribution with 79·3% and 100% of MBL and class D carbapenemase producers being highly resistant, respectively. Oppositely, 78·9% of the Ambler class A carbapenemase producers remained susceptible. Cefepime/zidebactam demonstrated the highest efficacy with most of isolates remaining below the resistance threshold (≤ 4 mg/L). Zidebactam (formerly WCK 5222) is a β-lactamase inhibitor of the diazabicyclooctane (DBO) family and is used in combination with cefepime to potent β-lactam efficiency.[25] In addition to its inhibition properties towards

Ambler class A and class D β-lactamases, zidebactam possesses intrinsic antimicrobial activity through its binding to PBP-2 and PBP-3.[26] However, this molecule does not show any efficacy by itself against Proteae.[25] Accordingly, 100%, 95% and 100% of Ambler class A carbapenemase-producers, Ambler class D carbapenemase-producers and non-carbapenemase-producers were susceptible to cefepime/zidebactam, respectively. Only 49% of MBL-producing isolates were susceptible to cefepime/zidebactam. Of note, cefepime-zidebactam susceptible isolates also exhibited cefepime MIC ≤ 4 mg/L, confirming the absence of intrinsic action of zidebactam towards the PBP of *Morganella* spp.

Finally, MICs of last generation cyclines (tigecycline and eravacycline) demonstrated only very moderate susceptibility to these molecules (2·3% to 31·8%).

Deciphering polymyxin resistance mechanism in *Morganella* spp.

Morganella spp. are known to be resistant to polymyxin at high-level. However, the underlying mechanism remained unknown. Usually, resistance to polymyxins is mediated by modifications of the lipid A, the membrane anchor of the lipopolysaccharide (LPS), through covalent addition of phosphoethanolamine (pEtN) or 4-deoxyaminoarabinose (L-Ara4N). In our collection two isolates (BEL-5 and BEL-6) exhibited pellicular susceptibility to polymyxin with MICs to colistin at 2 and 0·5 mg/L respectively. To decipher the structure of the lipid A, a MALDIxin test (mass-spectrometry assay dedicated to lipid A analysis) was performed on 10 colistin-resistant isolates (MICs > 256 mg/L) (nine *M. morganii subsp. morganii* and one *M. sibonii*) and BEL-5 and BEL-6. It clearly identified a strong decrease in L-Ara4N-modified lipid A in susceptible isolates (**Supplementary Figure S3**). It demonstrated that resistance to polymyxin is caused by addition of L-Ara4N in *Morganella* genus and that some genetic events could occurred, leading to the decrease in L-Ara4N modifications and acquired susceptibility to polymyxins. In *K. pneumoniae*, addition of L-Ara4N is mediated by the up-regulation of the operon *arnBCADTEF* under the control of different two-component systems (TCS) PhoP/Q and PmrA/B.[27] By homology, nine similar TCS have been identified in this study in *Morganella* spp including PhoP/Q and QseB/C. Additionally, Guckes *et al.* demonstrated that QseB/C, involved in *quorum sensing*, could also interfere on the PmrA/B regulon.[28] Comparative genomics were performed on BEL-5 and BEL-6 isolates. In BEL-5 isolate colistin resistance was likely due to the insertion of IS*10R* immediately upstream the *arn* operon leading to the truncation of the PmrA/Qse binding site that likely modified the expression of the *arn* operon (**Supplementary Figure S4**). In BEL-6, the PmrA/Qse binding site is intact but *arn operon* exhibited 10 times more

SNPs in *arnA*, *arnB* & *arnC* than the whole bracketing region in comparison to the polymyxin-resistant *M. morganii* subsp. *morganii* isolate 177A6 (**Supplementary Figure S5**), suggesting an impact of these mutations in the function of the *arn* operon. However, the role of each mutation remains to be elucidated.

Identification of "high-risk" clones among carbapenemase-producing *Morganella* spp. from worldwide

First, phylogenetic analysis was conducted by creating a MASH distance similarity matrix including the whole genome of 270 *M. morganii* (68 carbapenemase-producing and 2 susceptible isolates from France, 104 from Europe, 2 from Pasteur's Institute collection, 1 from Canada and 93 from NCBI) and 5 genomes of *M. psychrotolerans* from NCBI (**Supplementary Figure S6**). Since, *M. psychrotolerans* was found to be very distant from *M. morganii* (less than 86% average nucleotide identity (ANI)) (**Supplementary Figure S6**), it was discarded from the further whole-genome comparison and phylogenetic tree construction (**Figure 2**).

Morganella isolates can be roughly separated into five subpopulations (**Supplementary Figure S6** and **Figure 2**). The first subpopulation corresponds to *M. sibonii* (formerly *M. morganii* subsp. *sibonii*) including 23 isolates from our study and two reference strains (CIP103648 and CIP 103649). Since, these 23 isolates possessed less than 92% ANI with the *M. morganii* subsp. *morganii* and *M. psychrotolerans*, they were reclassified as an independent species named *M. sibonii* instead of a subspecies of *M. morganii*. In agreement with this new taxonomy, differential biochemical and phenotypic characteristics were observed between *M. morganii* and *M. sibonii* (cf. below). The second and main subpopulation (n=215) corresponds to *M. morganii* subsp. *morganii*. A 3rd subpopulation included 31 isolates, that phylogenetically formed an independent "undefined" group with MASH distances of 94-95% with *M. morganii* subsp. *morganii* (**Figure 2** and **Supplementary Table S2**). This subpopulation was thus reclassified as a new subspecies of *M. morganii*. Since some isolates of this subspecies possessed biochemical characteristics of both *M. morganii* subsp. *morganii* and *M. sibonii* (cf. below), it was named *M. morganii* subsp. *intermedius*. Finally, a unique isolate (Genbank accession number NRQY0000000) is separated from the four other populations (less than 94% ANI). This peculiar isolate was recovered from a grass grub *Costelytra sp.* in New Zealand and might be further recognized as a novel *Morganella* species if several other isolates will be reported.

As revealed by the phylogenetic tree, *M. morganii* subsp. *morganii*, which includes the majority of clinical isolates, might be divided in a wide diversity of subclones (**Figure 2**). Among them, five subclones (clone I to V) were more prevalent and disseminated worldwide. The clone I, which includes the NDM-1-producing isolates of the French outbreak (n=39) also includes unrelated carbapenemase-producing isolates from Germany (n=4), United-Kingdom (n= 2), France (n=2) and Belgium (n=1). It confirms that this "high-risk" clone I has already disseminated in Europe. A sub-tree and a SNP matrix constructed with isolates of this clone I revealed that this clone can be divided into three independent clusters (**Figure 3**). The cluster A included the unique Belgian isolate, the cluster B included the French cluster (n=41) and the cluster C included 7 isolates from Germany (n=4), France (n=1) and United-Kingdom (n=2). Isolates of the cluster B, which are all epidemiologically related, had less than 50 SNPs except for isolates 126H7 and 105F9 (ca. 70 to 100 SNPs). Accordingly, we decided to use a reasonable cut-off at 100 SNPs to separate clusters based on genomic results and epidemiological investigations. One French isolate 81B3 did not belong to the main French cluster (cluster B). As expected, epidemiological data revealed that the patient travelled in India and was not related to the outbreak. This strain was part of the cluster C that includes German (n=3) and English (n=2) isolates. Despite a close genetic relationship, resistomes of the isolates of this cluster C were different (**Supplementary Table S2**).

To dive deeper into the evolution of this clone I, a molecular clock was calculated using the strains of the French outbreak (cluster B) by dividing the number of SNPs by number of days that separate the collection dates from the 1st isolate of the outbreak considered to be collected at Day 0 (24A3) (**Supplementary Figure S7**). A mean of ca. 3·9 SNPs per year was observed (50% of isolates have a molecular clock comprise between 1·0 and 6·8 SNPS per year). This result is in agreement with what we previously observed for *Klebsiella pneumoniae* (7·5 SNP/year) or *Pseudomonas aeruginosa* (7·0 SNP/year).[29,30]

Apart from the clone I which was overrepresented due to the French outbreak, four additional prevalent clones were evidenced in our collection (**Figure 2**). We analyzed the subtree of each clones and compared the results with epidemiological data (**Supplementary Figure S8**). The clone III is mostly composed by KPC-producers from Czech Republic that have been reported to be part of the same outbreak. Our genomic analysis confirmed epidemiological data with SNPs ranging between of 21 to 70, which confirmed the robustness of our SNP cut-off of 100. We identified several cross-country disseminations for each main clones. As example, inside the clone IV we identified five isolates in Czech Republic carrying either OXA-48 or KPC-2 and close relationship between three isolates from France and Belgium (156C10, 249E6 and

BEL-22, respectively). Out of the five main clones, we identified three isolates from United-Kingdom (ANG-7, ANG-9 and ANG-12) with clear epidemiological link (44 to 81 SNPs) (**Supplementary Figure S9**) that were previously confirmed to be part of the same outbreak using PFGE (K. Hopkins personal data).

To better understand the spread of the five main clones their core genomes were compared to the global core genome of *M. morganii* subsp. *morganii*. We identified 13, 13 and 1 unique genes for the clone I, II and V respectively. No specific gene was identified for the clone III and IV (**Table 2**). Most of these genes corresponded to hypothetical proteins. However, some interesting genes might be involved in the international dissemination of these clones including a putative colicin and a ferrochrome-iron receptor. However, their roles in fitness, virulence and dissemination remain to be elucidated.

Core genome analysis of *Morganella* spp.

Phylogeny and MASH analysis demonstrated that *Morganella* genus could be separate into 4 species named *M. psychrotolerans*, *M. sibonii*, *M. morganii* and a new species represented by a unique strain isolate from *Costelytra* sp. In addition, *Morganella morganii* can be split into two subspecies named *M. morganii* subsp. *morganii* and a novel subspecies named *M. morganii* subsp. *intermedius* (**Figure 2**). The core genomes of *M. sibonii* and *M. morganii* subsp. *morganii* were compared (**Figure 5A**). Almost 2,700 genes were present in 95% of both species whereas 47 genes were specific to *M. morganii* subsp. *morganii* and 63 were predicted to be specific to *M. sibonii* (**Figure 4A** and **Supplementary Table S3**). As demonstrated below, biochemical characterization, as well as intrinsic resistance profile to tetracycline and conservation of metabolic pathways, secretion systems, etc… correlate this refined taxonomy.

As previously described, *M. sibonii* (previously *M. morganii* subsp. *sibonii*) differs from *M. morganii* subsp. *morganii* by assimilation of trehalose.[31] The biochemical characterization using Api20E and Api50CH galleries of all isolates of our collection confirmed that all *M. sibonii* isolates were able to use trehalose as unique carbone source as opposed to *M. morganii* subsp. *morganii* strains. Among *M. sibonii* specific genes we identified the presence of an operon involved in sugar transport (**Figure 4B**). The expression of this entire operon in a *M. morganii* subsp. *morganii* strain restore the ability of this strain to utilize trehalose as sole carbon source, demonstrating the functionality of this operon. The **supplementary**

Figure S10 summarizes the putative role of each partner of this operon in trehalose utilization in *M. sibonii*.

In *M. sibonii*, immediately downstream of the trehalose operon, a putative type VI secretion system (T6SS) has been systematically identified (**Figure 4B**). We also identified a putative type III secretion system (T3SS) specific from *M. siboni* and absent from *M. morganii* subsp. *morganii* genomes (**Supplementary Table S2**).

Sometimes wrongly considered as tetracycline resistant, *M. morganii* subsp. *morganii* is intrinsically susceptible to tetracycline. Oppositely, *M. sibonii* possess in its core genome a *tetD*-like resistance gene as well as its *tetR* regulatory gene that might lead to intrinsic tetracyclin resistance (**Figure 4C**). Expressed in *E. coli*, we confirmed that *tetD*-like gene was able to confer resistance to tetracycline (MIC > 128 mg/L) but not to tigecyline (MIC <0·25 mg/L) or eravacycline (MIC <0·25 mg/L). In *M. sibonii,* genetic environment surrounding *tetD*-like gene did not show any mobile element indicating a recent acquisition. Accordingly, partial loss of this *tetD*-like locus in *M. morganii* subsp. *morganii* might be the result of a deletion/contraction at this locus (**Figure 4C**).

Persistent genome comparison identified several other genes specific of *M. morganii* subsp. *morganii* (**Table S2**) including the whole locus *mglB/A/C* encoding the ABC transporter of galactose/methyl galactoside and its transcriptional regulator encoded by *galS* (**Figure 4E**).

Of note, *M. morganii* subsp. *intermedius* is at the halfway between *M. morganii* subsp. *morganii* and of *M. sibonii*. As example, all *M. morganii* subsp. *intermedius* isolates possess the *mglB/A/C* locus specific to *M. morganii* subsp. *morganii* but few strains also possess the trehalose operon, the T6SS and the resistance gene *tetD* associated to *M. sibonii* (**Figure 2, Supplementary Table S2**). Of note, the trehalose operon was identified in the isolate 131E1 with a similar synteny but with only 88·2% nucleotide identity compared to the *M. sibonii* reference strain CIP103648, indicating that this operon was not recently acquired but rather has evolved in parallel within each species.

DISCUSSION

The starting point of this study was to compare and decipher an outbreak of NDM-1-producing *M. morganii* subsp. *morganii* in France over a ten-year period. This analysis revealed a longitudinal outbreak with the same genetic background. Surprisingly, the clone was well conserved according to SNP analyses with, for some isolates, less than 20 SNPs over a ten year-period (**Figure 3**). Using epidemiological data and SNP analyses, a cut-off value of 100 SNPs along the whole genome was considered reasonable to discriminate

outbreak-related strains from non-clonally related isolates. Recently, a cut-off value of 20 SNPs along the core-genome was advocated to decipher if the *K. pneumoniae* isolates were clonally-related or not.[32] This lower SNPs number might be explained by differences in genome comparison processes used in both studies. Indeed, in our study whole genome comparison was performed instead of a core genome comparison used for *K. pneumoniae*. Besides, we determined that the molecular clock of *M. morganii* was around 3·9 SNPs/year, slightly lower than *K. pneumoniae* and *P. aeruginosa* that possess evolution rates of 7·5 and 7·0 SNPs/years, respectively, calculated with the same approach.[29,30]

Comparison of carbapenemase-producing French *Morganella* spp. isolates (n=68) with an international (mainly Europe) collection of multidrug resistant *Morganella* spp. (n=104) and with genomes from the Genbank database (n=104), the population structure of *Morganella* spp. was deciphered. Phylogeny and MASH analysis demonstrated that *Morganella* genus could be separate into 4 species named *M. psychrotolerans*, *M. sibonii*, *M. morganii* and a new species represented by a unique strain isolate from a grass grub *Costelytra* sp. In addition, *Morganella morganii* could be split into two subspecies named *M. morganii* subsp. *morganii* and *M. morganii* subsp. *intermedius*. Intrinsic resistance profile to tetracycline, conservation of metabolic pathways, secretion systems, etc… correlated this refined taxonomy. Carbapenemase-producing *Morganella* spp. isolates were mostly identified among five "high-risk" clones of *M. morganii* subsp. *morganii* that have already disseminated worldwide.

Morganella spp. was well-known to be intrinsically resistant to polymyxins. However, the molecular and biochemical mechanism remained unknown. In this study, the analysis of two *Morganella* isolates susceptible to colistin allowed us to demonstrate that intrinsic addition of L-Ara-4N on the lipid A *via* expression of *arnBCADTEF* leads to polymyxin resistance.

Finally, antimicrobial susceptibility testing allowed us to identify the best therapeutic options for the treatment of infections caused by MDR *Morganella* spp. As expected, we identified that relebactam do not restore imipenem susceptibility since *Morganella* spp. possess intrinsic decreased susceptibility to imipenem through PBP with low affinity to this molecule. We also demonstrated that the ceftazidime-avibactam, meropenem-vaborbactam and cefepime-zidebactam are suitable options for the treatment of Ambler class A and D carbapenemase-producing isolates as well as for non-carbapeneamse producers. Of note, as observed with Enterobacterales, the novel β-lactamase inhibitors (avibactam, relebactam, vaborbactam and zidebactam) are inefficient to restore the activity of carbapenems (imipenem or meropenem) or broad-spectrum cephalosporins (ceftazidime, cefepime) when a MBL (Ambler class B) was

411 produced. Finally, despite the presence of *tetD* gene responsible for tetracycline resistance
412 (acquired or intrinsic in *M. sibonii*) but not to tigecycline and eravacycline, most of
413 *Morganella* spp. isolates were resistant to both molecules (**Figure 1**).

414 To conclude, this work deeply analyzed the largest collection of *Morganella* spp. ever
415 published leading to a reorganization in *Morganella's* taxonomy using whole genome
416 sequencing data validated by key phenotypes (trehalose assimilation, tetracycline
417 resistance…). Our results identified new components or virulence factors of some *Morganella*
418 species (e.g. T6SS and T3SS in *M. sibonii*) that might be implicated the bacterial lifestyle.
419 Regarding the antimicrobial resistance potential of *Morganella* spp. (intrinsic resistance to
420 colistin, chromosome-encoded cephalosporinase associated, acquired carbapenemase
421 encoding genes), this genus might become a threatening issue in a next future.[20] Accordingly,
422 *Morganella* deserve more comprehensive studies to understand its lifestyle and its ability to
423 acquire resistance. It includes a better knowledge of the dissemination of "high-risk clones"
424 such as the one which was responsible for a large outbreak in France and that had already
425 spread at least in Europe.
426
427 **ONLINE METHODS**
428
429 **Strains collections**
430 All *Morganella morganii* isolates sent to the French National Reference Center (NRC) for
431 Antimicrobial Resistance from January 1st 2013 to March 1st 2021 were included (n=68). The
432 bacterial isolates referred to NRC were recovered from clinical and screening human
433 specimens collected in french microbiology laboratories. Additionally, 104 *M. morganii*
434 isolates referred to European antimicrobial resistance reference centers were added to the
435 collection: Germany (n=32), Belgium (n=26), England (n=17), Austria (n=3), The
436 Netherlands (n=4), Poland (n=1), Czech Republic (n=21). One additional carbapenemase-
437 producing *M. morganii* isolate was from Canada, two *M. morganii* subsp. *sibonii* reference
438 strains (CIP 103648 and CIP 103649) from the Pasteur Institute collection and two
439 susceptible isolates from Bicêtre Hospital. Genomes of all *Morganella* spp. isolates were
440 totally sequenced as described in supplementary methods.
441
442 **Bacterial identification**
443 The bacterial identification of all 172 *Morganella* spp. collection isolates was verified by
444 MALDI-TOF mass spectrometry (Biotyper, Bruker Daltonics).

Biochemical characterization

Biochemical characterization of the collection was performed using Api20E and Api50CH systems according to the manufacturer's recommendations (BioMérieux, France).

Antimicrobial susceptibility testing

Antimicrobial susceptibility testing was performed by the disc diffusion method on Mueller-Hinton (MH) agar (Bio-Rad, Marnes-La-Coquette, France) and interpreted according to EUCAST guidelines as updated in 2021 (http://www.eucast.org). MICs of temocillin, cefepime/zidebactam, ceftazidime, ceftazidime/avibactam, ceftolozane/tazobactam, ertapenem, imipenem, imipenem/relabactam, meropenem, meropenem/vaborbactam, colistin, eravacycline and tigecycline were determined by broth microdilution (Sensititre™ Thermofisher, France).

Carbapenemase detection.

Carbapenemase detection was performed using Carba NP-test as previously described, followed by an immunochromatographic detection of the carbapenemase enzyme using NG-Carba5 test (NG Biotech, Guipry, France). [33]

Whole genome sequencing

All *Morganella* spp. isolates were sequenced using Illumina's technology as previously described.[34] *De novo* assembly and read mappings were performed using CLC Genomics Workbench v12.0 (Qiagen, Les Ulis, France). Long read sequencing was performed on *M. morganii* 97F5 using PacBio's technologies as described previously. [29]

Bioinformatic analysis

The acquired antimicrobial resistance genes were identified using Resfinder server v3.1 (https://cge.cbs.dtu.dk/services/ResFinder/).[35]

The genomes were annotated using RAST server.[36,37] Phylogeny was performed using CSIphylogeny v1.4 server (www.cge.cbs.dtu.dk/services/CSIPhylogeny/) and visualised using iTOL software v4.[38] Sequences alignments were performed using ClustalW (https://www.genome.jp/tools-bin/clustalw). SNPs analysis was performed on whole genome using CSIphylogeny V1.4 with parameters as follow select min depth at SNP position at 10X, minimum distance between SNPs at 10 bp, minimum SNP quality at 30. Plasmid contents of

clinical isolates were analyzed by searching replicase gene using PlasmidFinder v2.1 and manual search for genes showing homology with a replicase gene.

Regarding Pangenome and persistent genomes determination, the 270 genomes of *Morganella morganii* were annotated using PanACoTA v1.2.0.[39] With default parameters (max L90 = 100 and a maximum of 999 contigs), 42 genomes did not pass the quality control. This gave our final dataset of 209 *Morganella morganii* genomes and 19 *Morganella sibonii* genomes. All 228 genomes were annotated with Prokka.[40] The pangenome of all 228 *Morganella* genomes was built, using the 'pangenome' module of PanACoTA (based on mmseqs2 v.13-45111) with default parameters.[41] Homologous families were determined by keeping only hits with at least 80% identity and an alignment covering at least 80% of both proteins. These proteins were then clustered by single linkage. This resulted into 22,952 families of homologous proteins. Persistent genomes were inferred from the pangenome previously calculated, using 'corepers' module of PanACoTA. We built a mixed persistent genome at 95% of the whole *Morganella morganii* species. This means that a gene is considered as persistent if it is present in a single copy in at least 95% of the genomes. A total of 2,697 persistent genes have been identified. Then, the persistent genome of the 2 subspecies was computed, from the *M. morganii* species pangenome. A total of 2,781 persistent genes in the 209 *M. morganii* genomes and 2,498 genes in the 19 *M. morganii subsp. siboni* genomes were identified.

Dataset of *Morganella* spp. genomes

A dataset of 275 *Morganella* spp. genomes including 103 genomes from Genbank and 172 genomes from our collection were used for bioinformatics analysis. Genomes similarity was estimated by calculating pairwise genetic distances with Mash v2.1.[42] For very similar genomes, the Mash distance D strongly correlates with alignment-based measures such as the Average Nucleotide Identity (ANI) based on whole-genome sequence, with $D \approx 1 - (ANI/100)$. Based on current taxonomy, 3 groups were identified: *Morganella psychotolerans* (5 genomes), *Morganella morganii* subsp. *sibonii* (23 genomes) and *Morganella morganii* subsp. *morganii* (247 genomes). Only genomes from *Morganella morganii* species were kept for further analysis corresponding to 270 genomes. Bioinformatic analysis to assess acquired antimicrobial resistance genes, phylogeny, single nucleotide polymorphisms (SNPs), plasmid content, pangenome and persistent genome are described in supplementary methods.

PCR and cloning experiments to assess trehalose assimilation and intrinsic tetracyline resistance in *Morganella siboni*

Whole cell-DNA was extracted as previously described.[43] DNA from *M. sibonii* GER-11 was used as template for the amplification of the trehalose operon using primers treF: 5'-ATTTGCGGTCAACACTCTCC-3' and treR 5'-CGGCATCTGTTCTGATAACC-3' and tetracyline resistance gene *tetD*-likeF : 5'-CGGGCAAAAACGAAAAGTCGC -3' and *tetD*-likeR : 5'-ACGGTTCCTCTGTGTCTGAG -3'. PCR were performed using Phusion polymerase (Thermo fischer scientific, Les Ulis, France) with an annealing temperature of 55°C respectively and an extended elongation time (5 minutes) to amplify the whole trehalose operon (7,517 bp in size). These amplicons were cloned into the Zero Blunt pTOPO (KanR) cloning vector (Thermo Fisher Scientific) and then transferred into electrocomptent *E. coli* TOP10 as previously described.[34] Plasmids were extracted from the recombinant *E. coli* using the GeneJet Plasmid miniprep kit according to the manufacturer (Thermo fischer scientific) and transferred into electrocompetent *M. morganii* O86D10. Electrocompetent *M. morganii* were prepared using the standard procedure for *E. coli* as previously described. [43]

MALDIxin test.

The protocol has been performed as previously described.[44] Briefly, a 10 µL inoculation loop of bacteria, grown on Mueller-Hinton agar for 18-24 hours, was resuspended in 200 µL of water. Mild-acid hydrolysis was performed on 100 µL of this suspension, by adding 100 µL of acetic acid 2 % v/v and incubating the mixture at 98°C for 30 min. Hydrolyzed cells were centrifuged at 17,000 *x g* for 2 min, the supernatant was discarded, and the pellet was washed 3 times with 300 µL of ultrapure water and resuspended to a density of McFarland 20 as measured using a McFarland Tube Densitometer. A volume of 0.4 µL of this suspension was loaded onto the MALDI target plate and immediately overlaid with 1.2 µL of a matrix Norharmane (Sigma-Aldrich) solubilized in chloroform/methanol 90:10 v/v to a final concentration of 10 mg/mL. For external calibration, 0.5 µL of calibration peptide was loaded along with 0.5 µL of the given calibration matrix (peptide calibration standard II, Bruker Daltonik, Germany). The samples were loaded onto a disposable MSP 96 target polished steel BC (Bruker Part-No. 8280800).

The bacterial suspension and matrix were mixed directly on the target by pipetting and the mix dried gently under a stream of air. The spectra were recorded in the linear negative-ion mode (laser intensity 95%, ion source 1 = 10.00 kV, ion source 2 = 8.98 kV, lens = 3.00 kV, detector voltage = 2652 V, pulsed ion extraction = 150 ns). Each spectrum corresponded to

546 ion accumulation of 5,000 laser shots randomly distributed on the spot. The spectra obtained
547 were processed with default parameters using FlexAnalysis v.3.4 software (Bruker Daltonik,
548 Germany).
549
550 **ACKNOWLEDGMENTS**
551 Transposon was named according the transposon Registry database. [45]
552
553 **DECLARATION OF INTERESTS**
554 None to declare
555

REFERENCES

1. Morgan, H. R. Upon the bacteriology of the summer diarrhoea of infants. *Br Med J* **2**, 908–912 (1907).
2. O'Hara, C. M., Brenner, F. W. & Miller, J. M. Classification, identification, and clinical significance of Proteus, Providencia, and Morganella. *Clin. Microbiol. Rev.* **13**, 534–546 (2000).
3. Emborg, J., Dalgaard, P. & Ahrens, P. Morganella psychrotolerans sp. nov., a histamine-producing bacterium isolated from various seafoods. *Int J Syst Evol Microbiol* **56**, 2473–2479 (2006).
4. Wang, D., Yamaki, S., Kawai, Y. & Yamazaki, K. Histamine Production Behaviors of a Psychrotolerant Histamine-Producer, Morganella psychrotolerans, in Various Environmental Conditions. *Curr Microbiol* **77**, 460–467 (2020).
5. Jensen, K. T. *et al.* Recognition of Morganella subspecies, with proposal of Morganella morganii subsp. morganii subsp. nov. and Morganella morganii subsp. sibonii subsp. nov. *Int J Syst Bacteriol* **42**, 613–620 (1992).
6. Bandy, A. Ringing bells: Morganella morganii fights for recognition. *Public Health* **182**, 45–50 (2020).
7. Liu, H., Zhu, J., Hu, Q. & Rao, X. Morganella morganii, a non-negligent opportunistic pathogen. *Int J Infect Dis* **50**, 10–17 (2016).
8. Tucci, V. & Isenberg, H. D. Hospital cluster epidemic with Morganella morganii. *J Clin Microbiol* **14**, 563–566 (1981).
9. Bonnin, R. A. *et al.* Genetic Diversity, Biochemical Properties, and Detection Methods of Minor Carbapenemases in Enterobacterales. *Front Med (Lausanne)* **7**, 616490 (2020).
10. Nordmann, P., Dortet, L. & Poirel, L. Carbapenem resistance in Enterobacteriaceae: here is the storm! *Trends Mol Med* **18**, 263–272 (2012).
11. Schultz, E. *et al.* Multidrug Resistance Salmonella Genomic Island 1 in a Morganella morganii subsp. morganii Human Clinical Isolate from France. *mSphere* **2**, (2017).
12. Olaitan, A. O. *et al.* Genome analysis of NDM-1 producing Morganella morganii clinical isolate. *Expert Rev Anti Infect Ther* **12**, 1297–1305 (2014).
13. Guo, X. *et al.* Detection and Genomic Characterization of a Morganella morganii Isolate From China That Produces NDM-5. *Front Microbiol* **10**, 1156 (2019).

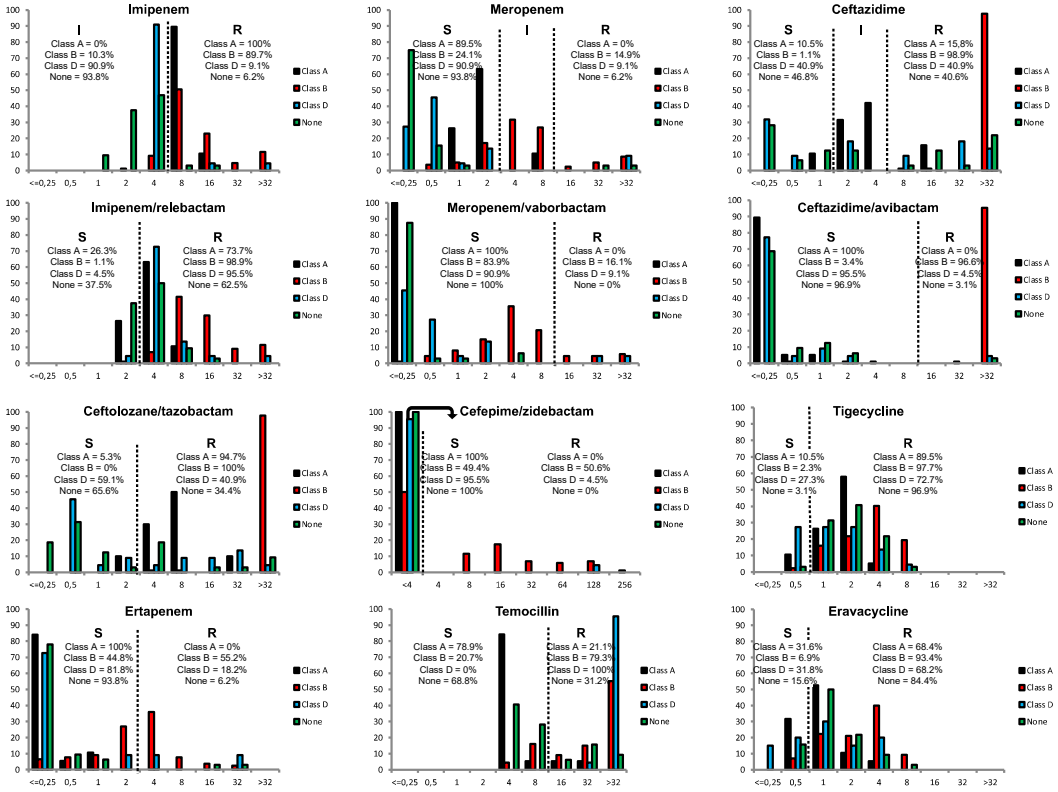
14. Aires-de-Sousa, M. *et al.* Occurrence of NDM-1-producing Morganella morganii and Proteus mirabilis in a single patient in Portugal: probable in vivo transfer by conjugation. *J. Antimicrob. Chemother.* **75**, 903–906 (2020).
15. Jamal, W. Y., Albert, M. J., Khodakhast, F., Poirel, L. & Rotimi, V. O. Emergence of New Sequence Type OXA-48 Carbapenemase-Producing Enterobacteriaceae in Kuwait. *Microb Drug Resist* **21**, 329–334 (2015).
16. Shi, D.-S., Wang, W.-P., Kuai, S.-G., Shao, H.-F. & Huang, M. Identification of bla KPC-2 on different plasmids of three Morganella morganii isolates. *Eur J Clin Microbiol Infect Dis* **31**, 797–803 (2012).
17. Cai, J. C. *et al.* Detection of KPC-2 and qnrS1 in clinical isolates of Morganella morganii from China. *Diagn Microbiol Infect Dis* **73**, 207–209 (2012).
18. Kukla, R. *et al.* Characterization of KPC-Encoding Plasmids from Enterobacteriaceae Isolated in a Czech Hospital. *Antimicrob Agents Chemother* **62**, (2018).
19. Moura, Q., Cerdeira, L., Fernandes, M. R., Vianello, M. A. & Lincopan, N. Novel class 1 integron (In1390) harboring blaGES-5 in a Morganella morganii strain recovered from a remote community. *Diagn. Microbiol. Infect. Dis.* **91**, 345–347 (2018).
20. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* S0140-6736(21)02724–0 (2022) doi:10.1016/S0140-6736(21)02724-0.
21. Li, B. *et al.* Prevalence and characteristics of ST131 clone among unselected clinical Escherichia coli in a Chinese university hospital. *Antimicrob Resist Infect Control* **6**, 118 (2017).
22. Sorlozano, A., Gutierrez, J., Jimenez, A., de Dios Luna, J. & Martínez, J. L. Contribution of a new mutation in parE to quinolone resistance in extended-spectrum-beta-lactamase-producing Escherichia coli isolates. *J Clin Microbiol* **45**, 2740–2742 (2007).
23. Antonelli, A., D'Andrea, M. M., Vaggelli, G., Docquier, J.-D. & Rossolini, G. M. OXA-372, a novel carbapenem-hydrolysing class D β-lactamase from a Citrobacter freundii isolated from a hospital wastewater plant. *J Antimicrob Chemother* **70**, 2749–2756 (2015).
24. Woerther, P.-L., Burdet, C., Chachaty, E. & Andremont, A. Trends in human fecal carriage of extended-spectrum β-lactamases in the community: toward the globalization of CTX-M. *Clin. Microbiol. Rev.* **26**, 744–758 (2013).

25. Livermore, D. M., Mushtaq, S., Warner, M., Vickers, A. & Woodford, N. In vitro activity of cefepime/zidebactam (WCK 5222) against Gram-negative bacteria. *J. Antimicrob. Chemother.* **72**, 1373–1385 (2017).
26. Rajavel, M. *et al.* Structural Characterization of Diazabicyclooctane β-Lactam 'Enhancers' in Complex with Penicillin-Binding Proteins PBP2 and PBP3 of Pseudomonas aeruginosa. *mBio* **12**, (2021).
27. Jeannot, K., Bolard, A. & Plésiat, P. Resistance to polymyxins in Gram-negative organisms. *Int J Antimicrob Agents* **49**, 526–535 (2017).
28. Guckes, K. R. *et al.* Strong cross-system interactions drive the activation of the QseB response regulator in the absence of its cognate sensor. *Proc Natl Acad Sci U S A* **110**, 16592–16597 (2013).
29. Boulant, T. *et al.* A 2.5-years within-patient evolution of a Pseudomonas aeruginosa with in vivo acquisition of ceftolozane-tazobactam and ceftazidime-avibactam resistance upon treatment. *Antimicrob Agents Chemother* (2019) doi:10.1128/AAC.01637-19.
30. Jousset, A. B. *et al.* A 4.5-Year Within-Patient Evolution of a Colistin-Resistant Klebsiella pneumoniae Carbapenemase-Producing K. pneumoniae Sequence Type 258. *Clin. Infect. Dis.* **67**, 1388–1394 (2018).
31. Janda, J. M., Abbott, S. L., Khashe, S. & Robin, T. Biochemical investigations of biogroups and subspecies of Morganella morganii. *J Clin Microbiol* **34**, 108–113 (1996).
32. David, S. *et al.* Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. *Nat Microbiol* (2019) doi:10.1038/s41564-019-0492-8.
33. Dortet, L., Bréchard, L., Poirel, L. & Nordmann, P. Impact of the isolation medium for detection of carbapenemase-producing Enterobacteriaceae using an updated version of the Carba NP test. *J. Med. Microbiol.* **63**, 772–776 (2014).
34. Girlich, D. *et al.* Chromosomal amplification of the blaOXA-58 carbapenemase gene in a Proteus mirabilis clinical isolate. *Antimicrob. Agents Chemother.* (2016) doi:10.1128/AAC.01697-16.
35. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).
36. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

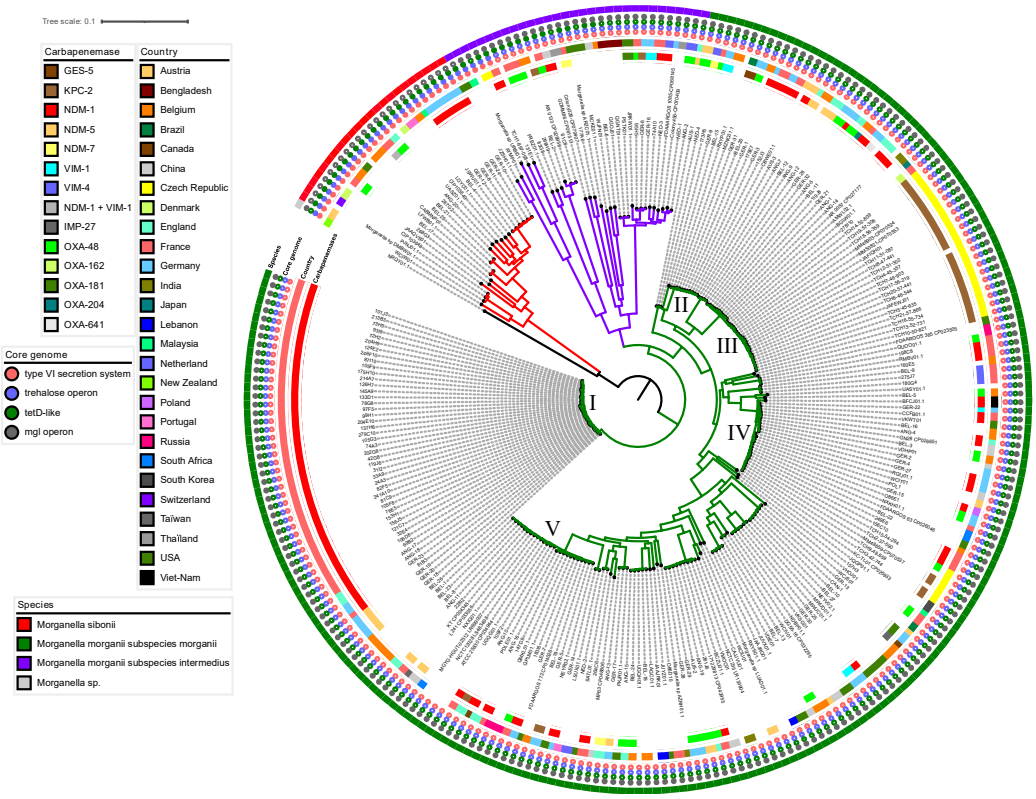
37. Aziz, R. K. *et al.* SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* **7**, e48053 (2012).
38. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256–W259 (2019).
39. Perrin, A. & Rocha, E. P. C. PanACoTA: a modular tool for massive microbial comparative genomics. *NAR Genom Bioinform* **3**, lqaa106 (2021).
40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
41. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
42. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
43. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular cloning: a laboratory manual*. (2001).
44. Dortet, L. *et al.* Rapid detection and discrimination of chromosome- and MCR-plasmid-mediated resistance to polymyxins by MALDI-TOF MS in Escherichia coli: the MALDIxin test. *J Antimicrob Chemother* **73**, 3359–3367 (2018).
45. Tansirichaiya, S., Rahman, M. A. & Roberts, A. P. The Transposon Registry. *Mob DNA* **10**, 40 (2019).

FIGURES'S LEGENDS

Figure 1. MICs distribution of twelve last resort antimicrobials. MICs were obtained using microbroth dilution. MIC distributions were separated according carbapenemase content and is indicated on the histograms. Clinical breakpoints correspond to EUCAST guidelines. S: Susceptible; R: Resistant.



673 **Figure 2. Phylogenetic analysis of the 270 isolates of *Morganella* spp.** Phylogenetic tree
674 was constructed using CSIphylogeny (cge.cbs.dtu.dk/services/CSIPhylogen) and visualized
675 using iTOL (itol.embl.de/). Produced carbapenemases are indicated by colored squares. The
676 number indicated the five main clones of *M. morganii* subsp. *morganii*.
677



678
679

Figure 3. Phylogenetic analysis of *M. morganii* subsp. *morganii* clone I.
This phylogenetic tree was obtained by comparing all isolates from *M. morganii* subsp. *morganii* clone I. A SNP-based matrix was visualized. Year of isolation, Country and carbapenemase of each isolate is indicated at the vicinity of the isolate name.

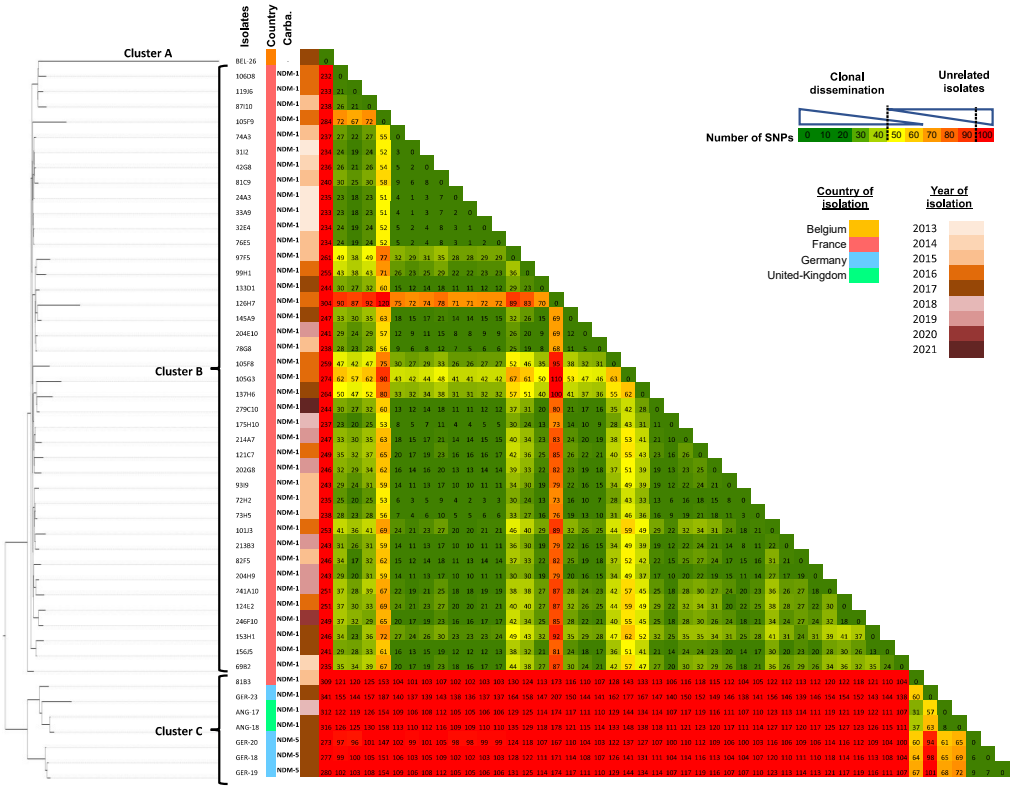
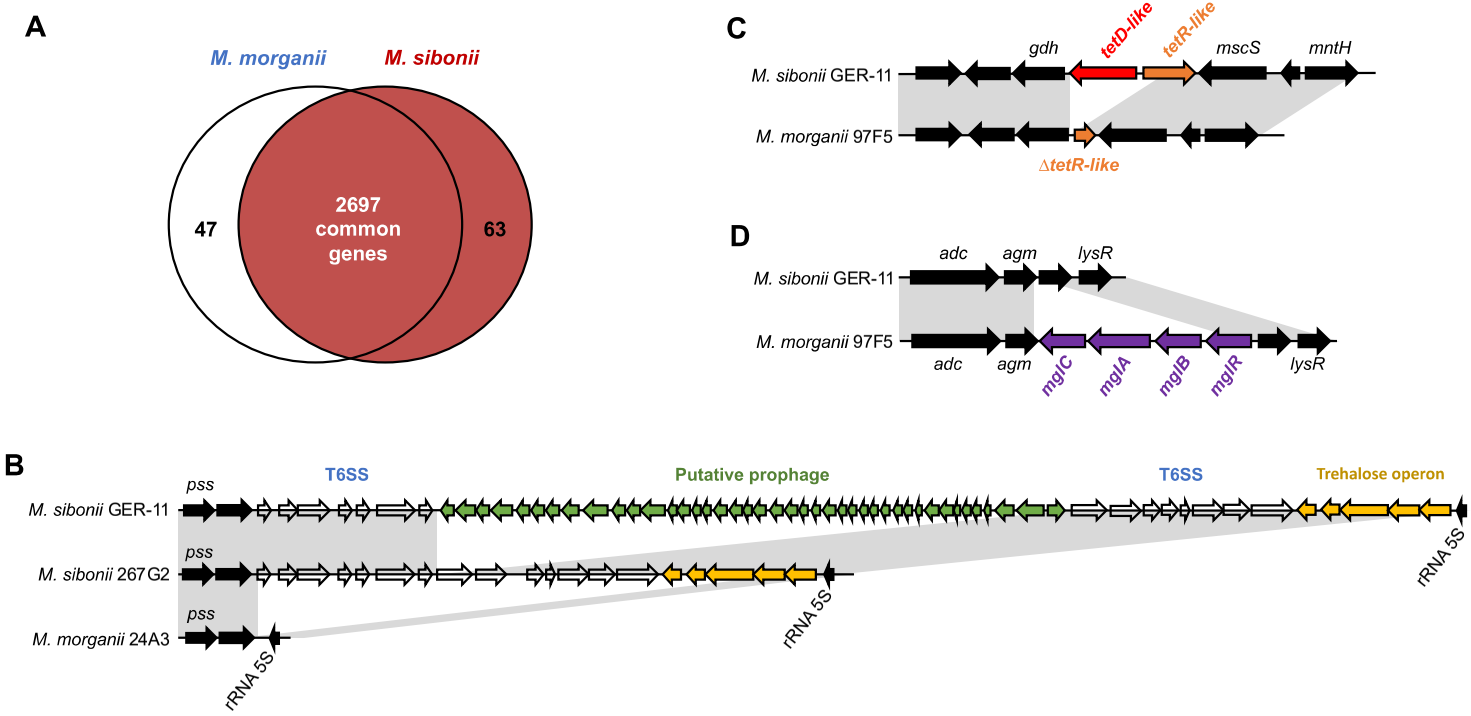


Figure 4. Core genome analysis of *Morganella spp.* A. Schematic representation of genes identified only in *M. morganii* subsp. *morganii* and *M. morganii* subsp. *sibonii*. **B.** Schematic representation and comparison of the trehalose operon locus in *M. sibonii* GER-11, *M. sibonii* 267G2 and *M. morganii* subsp. *morganii* 24A3. CDS are represented by arrows. Putative prophage is indicated in green, trehalose operon in yellow, type VI secretion system in blue and other genes in black. **C.** Schematic representation of the *tetD* locus in *M. morganii* subsp. *morganii* and *M. sibonii*. The *tetD*-like gene is represented in red and its regulator (*tetR*-like) in orange. (D) Schematic representation of the galatactose/methylgalactoside ABC transporter *mgl* locus in *M. morganii* subsp. *morganii* and *M. sibonii*. In all panels, nucleotide sequences with ≥99% identity are highlighted in grey.

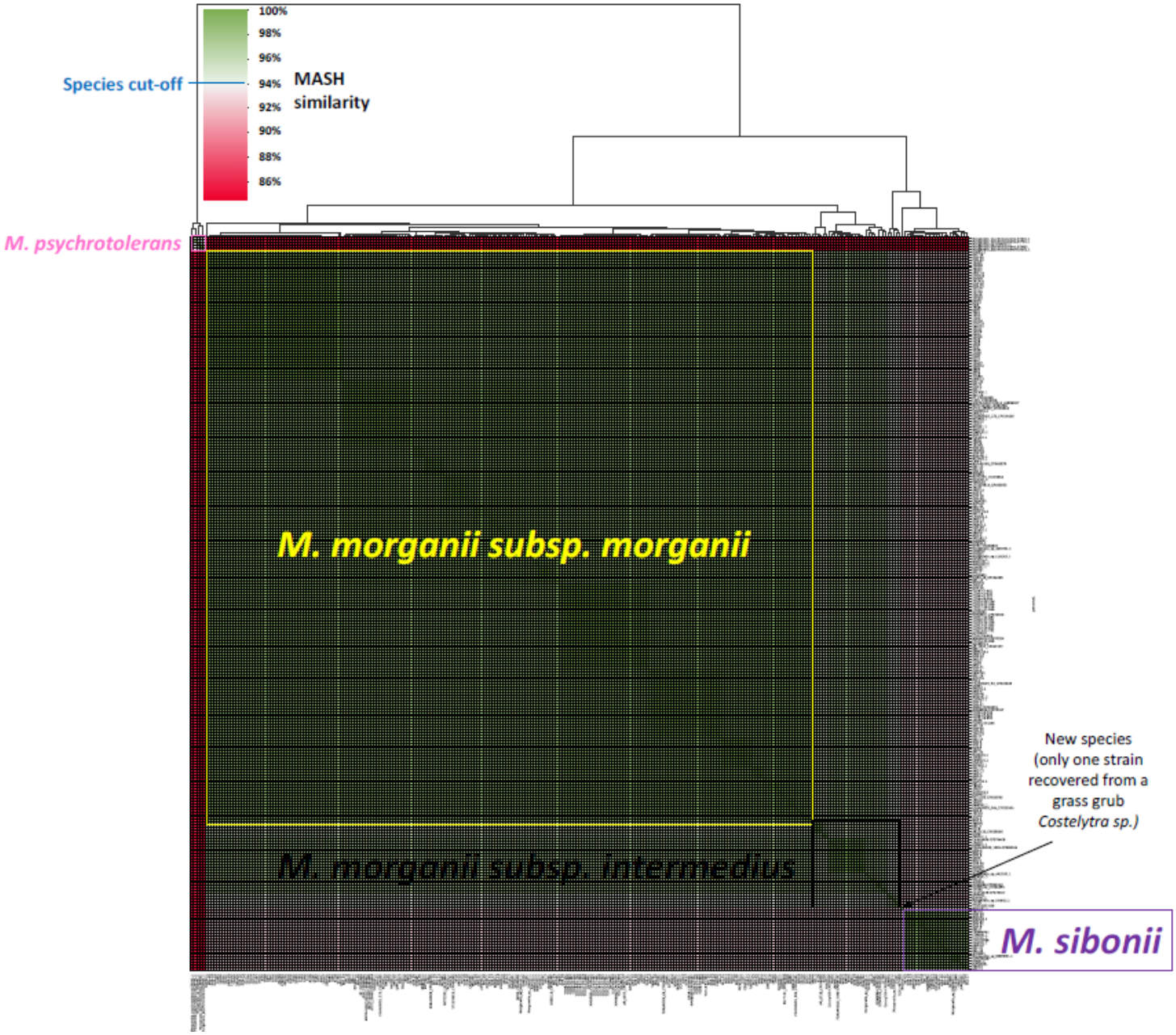


696 **Table 1:** Genetic features and diversity of acquired resistance genes of *Morganella* spp.
697 isolates
698

	M. morganii	***M. sibonii***	***M. psychrotolerans***
Number of genomes	247	23	5
Mean genome size (Mbp)	3.97	4.14	4.21
Mean contig number	165.5	154.95	40.4
N50 (kbp)	597.97	308.89	386.40
β -lactam resistance			
Class A carbapenemases	*bla*$_{KPC-2}$ (n=26) *bla*$_{GES-5}$ (n=1)	None	None
Class B carbapenamases	*bla*$_{NDM-1}$ (n=75) *bla*$_{NDM-5}$ (n=10) *bla*$_{NDM-7}$ (n=3) *bla*$_{VIM-1}$ (n=3) *bla*$_{VIM-1-like}$ (n=1) *bla*$_{VIM-4}$ (n=2) *bla*$_{IMP-27}$ (n=1)	*bla*$_{NDM-1}$ (n=6) *bla*$_{VIM-1}$ (n=1)	None
Class D carbapenamases	*bla*$_{OXA-48}$ (n=22) *bla*$_{OXA-162}$ (n=1) *bla*$_{OXA-181}$ (n=2) *bla*$_{OXA-204}$ (n=1) *bla*$_{OXA-641}$ (n=1)	*bla*$_{OXA-48}$ (n=5)	None
ESBLs/Cepahlosporinases	*bla*$_{CTX-M-3}$ (n=1) *bla*$_{CTX-M-14}$ (n=4) *bla*$_{CTX-M-15}$ (n=1) *bla*$_{CMY-2}$ (n=2) *bla*$_{CMY-4}$-like (n=6) *bla*$_{CMY-16}$-like (n=2) *bla*$_{DHA/MOR}$-like (n=247) *bla*$_{SHV-12}$ (n=6) *bla*$_{TEM-2}$ (n=1) *bla*$_{VEB-6}$-like (n=1) *bla*$_{OXA-35}$ (n=1)	*bla*$_{CTX-M-1}$ (n=1) *bla*$_{CMY-16}$ (n=1) *bla*$_{DHA/MOR}$-like (n=23)	None
Penicillinases	*bla*$_{TEM-1}$-like (n=43) *bla*$_{TEM-110}$-like (n=2) *bla*$_{CARB-2}$-like (n=10) *bla*$_{OXA-1}$ (n=30) *bla*$_{OXA-9}$ (n=2) *bla*$_{OXA-10}$-like (n=5)	*bla*$_{OXA-10}$ (n=1)	
Non- β -lactam resistance			
Aminoglycosides	*aadA1*-like (n=109) *aadA2*-like (n=39) *aadA5* (n=19) aadA7 (n=1) aadA12-like (n=6) aadA13-like (n=2) aadA16-like (n=2) aadA17-like (n=1) *aadA24*-like (n=9) *aadB*-like (n=13) *strA*-like (n=37) *strB*-like (n=36) aac(3')-II-like (n=44) aac(3')-IVa-like (n=6) aac(6')Ib-like (n=116) aac(6')-aph(2")-like (n=1)	*aadA1*-like (n=6) *aadA2*-like (n=1) *aadA5* (n=1) *aadA24*-like (n=1) *aadB*-like (n=6) *strA*-like (n=6) *strB*-like (n=6) aac(3')-IVa-like (n=2) aac(6')Ib-like (n=1) aph(3')-Ic (n=2) aph(4')-Ia (n=1) aph(3')-VIa-like (n=1)	None

	aph(3')-I-like (n=36) aph(3')-VIa-like (n=16) aph(4)-Ia (n=6) armA-like (n=42) rmtB-like (n=4) rmtC-like (n=39)		
Fluoroquinolones	*qnrA1*-like (n=43) *qnrB1*-like (n=2) *qnrB2* (n=1) *qnrB6* (n=1) *qnrB19* (n=1) *qnrB32*-like (n=6) *qnrB58*-like (n=1) *qnrB66*-like (n=1) *qnrD*-like (n=17) *qnrS1*-like (n=5) *qnrS2*-like (n=1) *qepA*-like (n=2)	*qnrD* (n=10)	None
Chloramphenicol	catA1-like (n=63) catA2-like (n=111) *catB2*-like (n=8) *catB3*-like (n=71) *florR*-like (n=14) cmlA-like (n=7)	catA2-like (n=16) *catB2*-like (n=1) *catB3*-like (n=3) *florR*-like (n=4)	None
Tetracycline	*tetA*-like (n=23) *tetB*-like (n=124) *tetD*-like (n=3) *tetL*-like (n=1) *tetY*-like (n=1)	*tetA*-like (n=1) *tetB*-like (n=1) *tetD*-like (n=23) *tetY*-like (n=1)	None
Macrolides	*mph(A)*-like (n=41) *mph(E)*-like (n=52) *ere(A)*-like (n=1) *ere(B)*-like (n=10) *erm(42)*-like (n=4) *erm(B)*-like (n=9) *msr(E)*-like (n=52)	*mph(A)* (n=2)	None
Colistin	*mcr-1* (n=2)		None
Rifampin	*arr-2* (n=2) *arr-3* (n=37)		None
Trimethoprime	*dfrA1*-like (n=103) *dfrA7* (n=1) *dfrA12*-like (n=14) *dfrA14*-like (n=15) *dfrA15* (n=4) *dfrA17*-like (n=19) *dfrA18-like* (n=9) *dfrA24*-like (n=1) *dfrA27*-like (n=2) *dfrA30*-like (n=1)	*dfrA1* (n=7) *dfrA12* (n=1) *dfrA14*-like (n=6) *dfrA16*-like (n=3) *dfrA17* (n=1)	None
Sulfamides	*sul1*-like (n=144) *sul2*-like (n=46) *sul3* (n=3)	*sul1* (n=3) *sul2* (n=8)	None

Supplementary Figure S6. MASH distance analysis. Genomes similarity were estimated by calculating pairwise genetic distances with Mash v2.1.[42]



Part IV

Conclusion and perspectives

Conclusion and perspectives

The main goal of my PhD was to develop a tool to generate the basic datasets required for any large-scale bacterial comparative genomics study, namely, a set of quality controlled annotated genomes, a pan and a core and/or persistent genome, and the MSA of each core genome family.

More than 12000 lines of Python and 2000 commits later, the first version of `PanACoTA` was released to the community. This version fulfilled the different requirements stated when introducing PanACoTA's paper on page 81.

To minimize the annotation inconsistencies, it provides a quality control module which automatically removes the genomes not fulfilling the different criteria, and uniformly annotates the remaining sequences of the dataset. As each dataset is different, the quality control criteria can be adapted by the user: nothing is imposed.

Regarding the ability to handle large-scale datasets, I want to clarify something. When I started my PhD (and thus the development of PanACoTA), we were considering as "large-scale" datasets with several hundreds of genomes. For example, in February 2018, there were 437 complete *Klebsiella pneumonia* genomes in RefSeq. Today, there are more than 25000 strains of *E. coli* in Refseq, including more than 2000 complete genomes. And I can easily predict that the increasing growth will continue. Regarding the pangenome computation, the tool is even faster than our first expectations, as it can compute a reliable pangenome of almost 4000 strains in 30 minutes. An analysis of the gene families showed a good consistency with the functional annotations. Moreover, a comparison with a small subset of this huge dataset showed that the construction of the pangenome is robust to large variations in the number of input genomes.

The next point was the adaptation of the pangenome definition to large-scale datasets. While the initial definition of a core genome required genes in all genomes, some tools now use this term to refer to families with genes in most genomes but not necessarily all. For other tools, such gene families are part of a *soft core*, *relaxed core*, or else *persistent*. Moreover, there is no clear definition regarding the number of genes allowed per genome in these gene families. Most of the time, it is not even specified by the tool. With PanACoTA, we propose a standardization of these definitions. As the needs vary according to the underlying biological question, we propose three different types of persistent genomes, clearly defined: strict, mixed and multi persistent genome.

Finally, the tool is based on six different modules, which can be run separately, allowing

the user to start/end where he needed, and/or to rerun some steps with new parameters. Indeed, one may want to introduce different methods at one or more steps, while still using as input a dataset generated by PanACoTA. We took care to produce file formats compatible with the main existing tools.

As PanACoTA was developed alongside comparative genomics studies, we progressively adapted the tool to concrete situations, and provided new useful outputs that we had not thought about in the beginning. For example, for many studies (like *Elizabethkingia anophelis* outbreak in Wisconsin or Population structure of carbapenemase-producing *Morganella* species), we needed to download all sequences of a given species available on Refseq. We thus developed a new module of PanACoTA to automatically download the sequences of a given species.

PanACoTA has already demonstrated its usefulness in several comparative genomics studies (see chapter III).

For example, a few months before I started my PhD, an outbreak caused by *Elizabethkingia anophelis* arose in Wisconsin. Even if there were "only" 66 confirmed cases, this was an exceptional number for an outbreak caused by this bacterium. Indeed, *E. anophelis* is a very common environmental bacterium, most of the time harmless. Several cases of meningitis caused by this bacterium had already been reported before, but they were health-care associated, and sporadic. This time, the bacterium was spreading across the population, its source was unknown and, despite of the introduction of an antibiotic treatment, many patients lost their lives. To understand the genomic features of this outbreak strain, we used our developing method to analyse outbreak strains provided by the CDC, together with other available *E. anophelis* strains. The development of the tool at the moment of these analyses provided mutual benefits: the pangenome computed thanks to the embryo of PanACoTA allowed to discover the origin of the epidemic, and this study allowed us to improve the method, for instance by providing more adapted formats for the pangenome results.

In the same vein, we used PanACoTA to explore the *Morganella* genus, an opportunistic pathogen involved in various infections, and, above all, resistant to many (if not all) antibiotics. Thanks to the results generated by PanACoTA, we could identify two subspecies, which were confirmed by chemical tests. We could also identify an operon, only present in *M. siboni* (one of the subspecies), explaining its capacity to use trehalose as a carbon source.

Our lab was contacted by an Australian team to help with the analysis of a very important dataset of *E. coli* strains. Composed of thousands of strains with very diverse origins (healthy or sick host, plants, food, water, soil etc.), this dataset was a good opportunity to study the global genomic diversity of this species. This was possible thanks to the pan and persistent genomes computed by PanACoTA. Thus, the ability of PanACoTA to handle very large datasets can also allow to study the global genomic diversity of a species. To depict the overall genomic diversity of thousands of strains, other tools propose graph-based approaches. At the same time as we were developing PanACoTA, we collaborated with a team from the Genoscope which was developing a method to represent the pangenome using a graph structure (see Annexe on page 219). In PPanGGOLiN, PanACoTA's method is used to build a pangenome, which is then converted to a partitioned pangenome graph. They developed

statistical methods to classify the different gene families based on the graph.

Thus, PanACoTA can be used in many different contexts, from short-term studies (like an epidemic, to understand how the strain became pathogen) to long-term species evolution studies. It can be used both directly, by analysing the pan and core/persistent genome generated, or as an input for more specific methods.

We could imagine even more applications in which PanACoTA could be useful. For example, the core genome is a fundamental basis for any Genome-Wide Association Studies. GWAS aim at exploring the associations between genetic variations and observed traits and have much more statistical power when using very large genome datasets. It can be used, for example, to identify mutations causing antibiotic resistance. PanACoTA could be also be used for reverse vaccinology: core genes are most likely the most desirable targets for novel vaccine candidates. Actually, it is maybe already used for these kind of applications...and you might be more aware than me if you are actually using it!

PanACoTA has been developed in the aim of being useful for the community. By useful, I not only mean providing interesting scientific information, but above all providing them in a reliable and upgradable manner. In a way, PanACoTA is the pipette of the bioanalyst: it is used to prepare data, which will be used for further analyses. If the pipette is contaminated, or if it does not precisely extract the right amount of product, the culture will be biased, and the experiment will most likely fail. Similarly, if there is a mistake in the code of the software, the generated datasets will be wrong, and will bias the subsequent analyses. Thus, it is important to be thorough during the development to provide the most reliable tool possible.

First, I hosted PanACoTA on a Github to track changes in the code throughout. As science is constantly evolving, tools must be able to adapt: they must be easily maintainable and upgradable. For example, as stated above, we added new input or output formats to be compliant with other tools, or new features like the possibility to download genomes from Refseq. This was possible thanks to the structure of the code, which allows to easily add new features without the need to change everything. However, even if it is easy to add, it is also easy to introduce unwanted behaviors in other parts of the code as side effects. To minimize this risk, I set up a continuous integration process (CI), which is automatically run each time new code is pushed to the git repository. This process checks the installation step, and, if it is successful, runs tests to check each unit functionality.

Another important point is the accessibility of the software, which is essential for its usefulness. With PAnACoTA, I developed a full documentation, which provides details on each step, as well as a toy dataset to quickly get acquainted to the software. The latter is automatically regenerated by the CI when a new version is pushed. I also provide several means to install PanACoTA. The tool is available on conda and pip package-management platforms, and a singularity image including all needed dependencies hosted in Docker Hub provides the possibility to run PanACoTA without needing to install anything.

"Communication" with users is another important point for me. Providing a feedback to the user is important. Sometimes, nothing but a little print on the terminal is helpful to know at which step is the program, or even quite simply that it has not crashed. PanACoTA also

outputs log files, in order to keep a trace of the commands and parameters used, to guarantee reproductibility of the analysis. In the same vein, the issue tracking system of Github allows users to report bugs, ask questions, make suggestions or, for developers, participate to the software improvement. I have already handled several issues, some of which leading to the addition of a new feature to the software. For example, one can now extract the core genome of different subsets of genomes from a same pangenome. Since its first release, more than 8000 pip-package downloads of the tool have been identified.

In the short future, I would like to harness PanACoTA's results to develop a method to detect MGEs in bacterial genomes. Indeed, although MGEs are the main agents mediating Horizontal Gene Transfer, the main mechanism driving bacterial genome evolution, they are relatively little known. If we take the example of phages, in Refseq, only 200 genomes are annotated as "*Salmonella* phages", compared to the more than 10000 bacterial genomes of this gender (as of January 2022). Yet, understanding the evolution of bacterial genomes requires to understand the functioning and evolution of the elements shaping it. For example, the ICE discovered in the genomes of *Elizabethkingia anophelis* allowed to understand the origin of the outbreak. So, why are they so little known?

Detecting MGEs is a difficult task, as they are "hidden" among the other bacterial genes. MGEs identified so far are mostly plasmids and phages, partly because they are (or can be for phages) independent replicons, and can thus be cultured and sequenced separately. Several tools, like PHASTER and VirSorter2, propose the detection of prophages inside bacterial genomes based on homology search against known databases [7] [82]. But what about the unknown phages, absent from the databases? And the other MGEs, for which no database is available?

There is, yet, no automatic tool available to precisely detect MGEs *de novo*. By definition, MGEs are moving within and between genomes. Their short residence time implies that they cannot be part of the core genome. Moreover, as we saw in the first chapter, many of them need a more or less specific integration site to be inserted in the bacterial genome. Hence, close MGEs are likely to be inserted in the same loci in different genomes [20]. I wish to use the patterns of gene presence variation to identify the MGEs.

To start, we will try this method on the detection of prophages. We have a set of *Salmonella enterica* genomes for which we already know some prophages, and will use them to test the method. Having a method able to detect MGEs would open many new perspectives in the bacterial genomics world.

As I am a permanent engineer at Institut Pasteur, I will be able to explore this new horizon after my PhD defense. So, phages, take advantage of my holidays to find a good hiding place, and be ready for a hide and seek game!

References

Bibliography

[1] Carlos G. Acevedo-Rocha, Gang Fang, Markus Schmidt, et al. *From essential to persistent genes: A functional approach to constructing synthetic life*. 2013. DOI: 10.1016/j.tig.2012.11.001.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, et al. *Molecular Biology of the Cell*. Ed. by Garland Science. 2015, p. 1342. ISBN: 978-0-8153-4432-2.

[3] Stephen F. Altschul, Warren Gish, Webb Miller, et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2.

[4] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, et al. "Opportunities and challenges in long-read sequencing data analysis". In: *Genome Biology* 21.1 (2020), p. 30. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1935-5.

[5] Alexandre Ambrogelly, Sotiria Palioura, and Dieter Söll. "Natural expansion of the genetic code". In: *Nature Chemical Biology* 3.1 (2007), pp. 29–35. ISSN: 1552-4469. DOI: 10.1038/nchembio847.

[6] Masanori Arita, Ilene Karsch-Mizrachi, Guy Cochrane, and on behalf of the International Nucleotide Sequence Database Collaboration. "The international nucleotide sequence database collaboration". In: *Nucleic Acids Research* 49.D1 (2021), pp. D121–D124. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA967.

[7] David Arndt, Jason R. Grant, Ana Marcu, et al. "PHASTER: a better, faster version of the PHAST phage search tool". In: *Nucleic Acids Research* 44.W1 (2016), W16–W21. ISSN: 0305-1048. DOI: 10.1093/NAR/GKW387.

[8] Oswald T. Avery, Colin M. Macleod, and Maclyn McCarty. "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii". In: *Journal of Experimental Medicine* 79.2 (1944), pp. 137–158. ISSN: 15409538. DOI: 10.1084/jem.79.2.137.

[9] Sajad Babakhani and Mana Oloomi. "Transposons: the agents of antibiotic resistance in bacteria". In: *Journal of basic microbiology* 58.11 (2018), pp. 905–917. ISSN: 1521-4028. DOI: 10.1002/JOBM.201800204.

[10] Daniel N. Baker and Ben Langmead. "Dashing: fast and accurate genomic distances with HyperLogLog". In: *Genome Biology* 20.1 (2019), p. 265. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1875-0.

[11] Anton Bankevich, Sergey Nurk, Dmitry Antipov, et al. "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of Computational Biology* 19.5 (2012), pp. 455–477. ISSN: 1066-5277. DOI: 10.1089/cmb.2012.0021.

[12] Robert W. Bauman. "Microbial Genetics". In: *Microbiology with Diseases by Body System*. Ed. by Pearson. 5th. PRENTICE HALL, 2017. Chap. 7, pp. 192–236. ISBN: 9780134618449.

[13] Sion C Bayliss, Harry A Thorpe, Nicola M Coyle, et al. "PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria". In: *GigaScience* 8.10 (2019). ISSN: 2047-217X. DOI: 10.1093/gigascience/giz119.

[14] Frida Belinky, Igor B. Rogozin, and Eugene V. Koonin. "Selection on start codons in prokaryotes and potential compensatory nucleotide substitutions". In: *Scientific Reports* 7.1 (2017), pp. 1–10. ISSN: 2045-2322. DOI: 10.1038/s41598-017-12619-6.

[15] J. Manuel Bello-López, Omar A. Cabrero-Martínez, Gabriela Ibáñez-Cervantes, et al. *Horizontal gene transfer and its association with antibiotic resistance in the genus aeromonas spp.* 2019. DOI: 10.3390/microorganisms7090363.

[16] Pauline M Bennett. "Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria". In: *British Journal of Pharmacology* 153.Suppl 1 (2008), S357. DOI: 10.1038/SJ.BJP.0707607.

[17] G Bertani. "Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli". In: *Journal of bacteriology* 62.3 (1951), pp. 293–300. ISSN: 0021-9193. DOI: 10.1128/JB.62.3.293-300.1951.

[18] Jochen Blom, Stefan P. Albaum, Daniel Doppmeier, et al. "EDGAR: A software framework for the comparative analysis of prokaryotic genomes". In: *BMC Bioinformatics* 10.1 (2009), p. 154. ISSN: 14712105. DOI: 10.1186/1471-2105-10-154.

[19] Jochen Blom, Julian Kreis, Sebastian Spänig, et al. "EDGAR 2.0: an enhanced software platform for comparative gene content analyses". In: *Nucleic acids research* 44.W1 (2016), W22–W28. ISSN: 13624962. DOI: 10.1093/nar/gkw255.

[20] Louis Marie Bobay, Eduardo P.C. Rocha, and Marie Touchon. "The adaptation of temperate bacteriophages to their host genomes". In: *Molecular biology and evolution* 30.4 (2013), pp. 737–751. ISSN: 1537-1719. DOI: 10.1093/MOLBEV/MSS279.

[21] Vincenzo Bonnici, Rosalba Giugno, and Vincenzo Manca. "PanDelos: A dictionary-based method for pan-genome content discovery". In: *BMC Bioinformatics* 19.15 (2018), pp. 47–59. ISSN: 14712105. DOI: 10.1186/S12859-018-2417-6/FIGURES/3.

[22] Benjamin Buchfink, Klaus Reuter, and Hajk Georg Drost. "Sensitive protein alignments at tree-of-life scale using DIAMOND". In: *Nature Methods* 18.4 (2021), pp. 366–368. ISSN: 15487105. DOI: 10.1038/s41592-021-01101-x.

[23] Vincent Burrus, Guillaume Pavlovic, Bernard Decaris, and Gérard Guédon. "Conjugative transposons: the tip of the iceberg". In: *Molecular Microbiology* 46.3 (2002), pp. 601–610. ISSN: 1365-2958. DOI: 10.1046/J.1365-2958.2002.03191.X.

[24] Devon R. Byrd and Steven W. Matson. "Nicking by transesterification: The reaction catalysed by a relaxase". In: *Molecular Microbiology* 25.6 (1997), pp. 1011–1022. ISSN: 0950382X. DOI: 10.1046/j.1365-2958.1997.5241885.x.

[25] Allan Campbell. "The future of bacteriophage biology". In: *Nature Reviews Genetics* 4.6 (2003), pp. 471–477. ISSN: 1471-0064. DOI: 10.1038/nrg1089.

[26] Allan M Campbell. "Chromosomal insertion sites for phages and plasmids". In: *JOURNAL OF BACTERIOLOGY* 174.23 (1992), pp. 7495–7499.

[27] Carlos Canchaya, Ghislain Fournous, Sandra Chibani-Chennoufi, et al. "Phage as agents of lateral gene transfer". In: *Current Opinion in Microbiology* 6.4 (2003), pp. 417–424. ISSN: 1369-5274. DOI: 10.1016/S1369-5274(03)00086-9.

[28] Nicolas Carraro, Dominique Poulin, and Vincent Burrus. "Replication and active partition of Integrative and Conjugative Elements (ICEs) of the SXT/R391 family: The line between ICEs and conjugative plasmids is getting thinner". In: *PLoS Genetics* 11.6 (2015), pp. 98–102. ISSN: 15537404. DOI: 10.1371/journal.pgen.1005298.

[29] Eric Cascales and Peter J. Christie. "The versatile bacterial type IV secretion systems". In: *Nature Reviews Microbiology* 1.2 (2003), pp. 137–149. ISSN: 17401534. DOI: 10.1038/nrmicro753.

[30] Sherwood R. Casjens and Eddie B. Gilcrease. "Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions". In: *Methods in molecular biology* 502 (2009), pp. 91–111. DOI: 10.1007/978-1-60327-565-1{_}7.

[31] P. B. van Cauwenberge, A. M. Vander Mijnsbrugge, and K. J.A.O. Ingels. "The microbiology of acute and chronic sinusits and otitis media: a review". In: *European Archives of Oto-Rhino-Laryngology* 250.1 (1993), S3–S6. ISSN: 1434-4726. DOI: 10.1007/BF02540108.

[32] Sujoy Chatterjee and Eli Rothenberg. "Interaction of Bacteriophage λ with Its E. coli Receptor, LamB". In: *Viruses* 4.11 (2012), pp. 3162–3178. DOI: 10.3390/V4113162.

[33] Shouqiang Cheng, Yu Liu, Christopher S. Crowley, et al. "Bacterial microcompartments: their properties and paradoxes". In: *BioEssays* 30.11-12 (2008), pp. 1084–1095. ISSN: 1521-1878. DOI: 10.1002/BIES.20830.

[34] Alvin J. Clark, Michael Chamberlin, Richard P. Boyce, and Paul Howard-Flanders. "Abnormal metabolic response to ultraviolet light of a recombination deficient mutant of Escherichia coli K12". In: *Journal of Molecular Biology* 19.2 (1966), pp. 442–454. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(66)80015-3.

[35] Alvin J. Clark and Steven J. Sandler. "Homologous genetic recombination: The pieces begin to fall into place". In: *Critical Reviews in Microbiology* 20.2 (1994), pp. 125–142. DOI: 10.3109/10408419409113552.

[36] Jean Pierre Claverys and Leiv S. Håvarstein. "Cannibalism and fratricide: Mechanisms and raisons d'être". In: *Nature Reviews Microbiology* 5.3 (2007), pp. 219–229. DOI: 10.1038/NRMICRO1613.

[37] Stanley N Cohen, Annie C Y Chang, Herbert W Boyer, et al. "Construction of biologically functional bacterial plasmids in vitro". In: *Proceedings of the National Academy of Sciences of the United States of America* 70.11 (1973), pp. 3240–3244. ISSN: 0027-8424. DOI: 10.1073/PNAS.70.11.3240.

[38] Stanley N. Cohen. "Bacterial plasmids: their extraordinary contribution to molecular genetics". In: *Gene* 135.1–2 (1993), pp. 67–76. ISSN: 03781119. DOI: 10.1016/0378-1119(93)90050-D.

[39] Marta Colomer-Lluch, Juan Jofre, and Maite Muniesa. "Antibiotic Resistance Genes in the bacteriophage DNA fraction of environmental samples". In: *PLOS ONE* 6.3 (2011), e17549. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0017549.

[40] The Computational Pan-Genomics Consortium, Tobias Marschall, Manja Marz, et al. "Computational pan-genomics: status, promises and challenges". In: *Briefings in Bioinformatics* 19.1 (2018), pp. 118–135. ISSN: 1467-5463. DOI: 10.1093/BIB/BBW089.

[41] Bruno Contreras-Moreira and Pablo Vinuesa. "GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis". In: *Applied and Environmental Microbiology* 79.24 (2013), pp. 7696–7701. ISSN: 00992240. DOI: 10.1128/AEM.02411-13.

[42] Salvatore Cosentino and Wataru Iwasaki. "SonicParanoid: fast, accurate and easy orthology inference". In: *Bioinformatics* 35.1 (2019), pp. 149–151. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty631.

[43] J. Craig Venter, M. D. Adams, E. W. Myers, et al. "The sequence of the human genome". In: *Science* 291.5507 (2001), pp. 1304–1351. ISSN: 00368075. DOI: 10.1126/SCIENCE.1058040/SUPPL{_}FILE/1058040S3-20{_}THUMB.GIF.

[44] Jean Cury, Thomas Jové, Marie Touchon, et al. "Identification and analysis of integrons and cassette arrays in bacterial genomes". In: *Nucleic Acids Research* 44.10 (2016), p. 4550. DOI: 10.1093/NAR/GKW319.

[45] Félix D'Herelle. "Sur un microbe invisible antagoniste des bacilles dysentériques". In: *Comptes rendus Acad. Sci. Paris* 165.11 (1917), pp. 373–375.

[46] Félix D'Herelle. *The Bacteriophage, its role in immunity*. Baltimore: Williams \& Wilkins, 1922, pp. 443–444.

[47] Bernard D. Davis. "Nonfiltrability of the agents of genetic recombination in Escherichia coli". In: *Journal of bacteriology* 60.4 (1950), pp. 507–508. ISSN: 00219193. DOI: 10.1128/jb.60.4.507-508.1950.

[48] Fernando De La Cruz, Laura S. Frost, Richard J. Meyer, and Ellen L. Zechner. *Conjugative DNA metabolism in Gram-negative bacteria*. 2010. DOI: 10.1111/j.1574-6976.2009.00195.x.

[49] Erick Denamur, Olivier Clermont, Stéphane Bonacorsi, and David Gordon. "The population genetics of pathogenic Escherichia coli". In: *Nature Reviews Microbiology 2020 19:1* 19.1 (2020), pp. 37–54. ISSN: 1740-1534. DOI: 10.1038/s41579-020-0416-x.

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

[51] Erwin L. van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. "The Third Revolution in Sequencing Technology". In: *Trends in Genetics* 34.9 (2018), pp. 666–681. ISSN: 01689525. DOI: 10.1016/j.tig.2018.05.008.

[52] Wei Ding, Franz Baumdicker, and Richard A Neher. "panX: pan-genome analysis and exploration". In: *Nucleic Acids Research* 46.1 (2018), e5–e5. ISSN: 0305-1048. DOI: 10.1093/nar/gkx977.

[53] Sara Domingues, Gabriela J. da Silva, and Kaare M. Nielsen. "Integrons: Vehicles and pathways for horizontal dissemination in bacteria". In: *Mobile Genetic Elements* 2.5 (2012), pp. 211–223. DOI: 10.4161/MGE.22967.

[54] Sean R. Eddy. "Accelerated profile HMM searches". In: *PLoS Computational Biology* 7.10 (2011), e1002195. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002195.

[55] Robert C. Edgar. "MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping". In: *bioRxiv* (2021), p. 2021.06.20.449169. DOI: 10.1101/2021.06.20.449169.

[56] Robert C. Edgar. "MUSCLE: A multiple sequence alignment method with reduced time and space complexity". In: *BMC Bioinformatics* 5.1 (2004), pp. 1–19. ISSN: 14712105. DOI: 10.1186/1471-2105-5-113/FIGURES/16.

[57] A. J. Enright. "An efficient algorithm for large-scale detection of protein families". In: *Nucleic Acids Research* 30.7 (2002), pp. 1575–1584. ISSN: 1362-4962. DOI: 10.1093/nar/30.7.1575.

[58] J.P. Euzéby. "List of Bacterial Names with Standing in Nomenclature: a Folder Available on the Internet". In: *International Journal of Systematic and Evolutionary Microbiology* 47.2 (1997), pp. 590–592. ISSN: 1466-5026. DOI: 10.1099/00207713-47-2-590.

[59] Michael Feiss, R. A. Fisher, M. A. Crayton, and Carol Egner. "Packaging of the bacteriophage λ chromosome: Effect of chromosome length". In: *Virology* 77.1 (1977), pp. 281–293. ISSN: 0042-6822. DOI: 10.1016/0042-6822(77)90425-1.

[60] Walter Fiers, Roland Contreras, Fred Duerinck, et al. "Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene". In: *Nature* 260.5551 (1976), pp. 500–507. DOI: 10.1038/260500A0.

[61] Steven E. Finkel and Roberto Kolter. "DNA as a nutrient: Novel role for bacterial competence gene homologs". In: *Journal of Bacteriology* 183.21 (2001), pp. 6288–6293. ISSN: 00219193. DOI: 10.1128/JB.183.21.6288-6293.2001.

[62] R. D. Finn, J. Tate, J. Mistry, et al. "The Pfam protein families database". In: *Nucleic Acids Research* 36.Database (2007), pp. D281–D288. ISSN: 0305-1048. DOI: 10.1093/nar/gkm960.

[63] Robert D. Fleischmann, Mark D. Adams, Owen White, et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223 (1995), pp. 496–512. ISSN: 00368075. DOI: 10.1126/science.7542800.

[64] Derrick E. Fouts, Lauren Brinkac, Erin Beck, et al. "PanOCT: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species". In: *Nucleic Acids Research* 40.22 (2012), e172. ISSN: 03051048. DOI: 10.1093/nar/gks757.

[65] Victor J. Freeman. "Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphteriae". In: *Journal of Bacteriology* 61.6 (1951), pp. 675–688.

[66] Laura S. Frost, Raphael Leplae, Anne O. Summers, and Ariane Toussaint. "Mobile genetic elements: the agents of open source evolution". In: *Nature Reviews Microbiology* 3.9 (2005), pp. 722–732. ISSN: 1740-1534. DOI: 10.1038/nrmicro1235.

[67] Limin Fu, Beifang Niu, Zhengwei Zhu, et al. "CD-HIT: Accelerated for clustering the next-generation sequencing data". In: *Bioinformatics* 28.23 (2012), pp. 3150–3152. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts565.

[68] Asami Fukuda, Yuichi Kodama, Jun Mashima, et al. "DDBJ update: streamlining submission and access of human data". In: *Nucleic acids research* 49.D1 (2021), pp. D71–D75. ISSN: 1362-4962. DOI: 10.1093/NAR/GKAA982.

[69] Kenji Fukui. "DNA Mismatch Repair in Eukaryotes and Bacteria". In: *Journal of Nucleic Acids* 2010 (2010), p. 16. ISSN: 20900201. DOI: 10.4061/2010/260512.

[70] Carl W. Fuller, Lyle R. Middendorf, Steven A. Benner, et al. "The challenges of sequencing by synthesis". In: *Nature Biotechnology* 27.11 (2009), pp. 1013–1023. ISSN: 1546-1696. DOI: 10.1038/nbt.1585.

[71] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, et al. "PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph". In: *PLoS Computational Biology* 16.3 (2020). Ed. by Christos A. Ouzounis, e1007732. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1007732.

[72] Beth Gibson, Daniel J. Wilson, Edward Feil, and Adam Eyre-Walker. "The distribution of bacterial doubling times in the wild". In: *Proceedings of the Royal Society B: Biological Sciences* 285.1880 (2018). DOI: 10.1098/RSPB.2018.0789.

[73] M. Girvan and M. E.J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences* 99.12 (2002), pp. 7821–7826. ISSN: 0027-8424. DOI: 10.1073/PNAS.122653799.

[74] Agnieszka A. Golicz, Philipp E. Bayer, Prem L. Bhalla, et al. "Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications". In: *Trends in Genetics* 36.2 (2020), pp. 132–145. ISSN: 0168-9525. DOI: 10.1016/J.TIG.2019.11.006.

[75] Hans Gram and Carl Friedlaender. *Ueber die isolirte Färbung der Schizomyceten : in Schnitt-und Trockenpräparaten*. Vol. 2. Berlin: Theodor Fischer's medicinischer Buchhandlung, 1884, pp. 185–189.

[76] Fred Griffith. "The significance of Pneumococcal types". In: *Journal of Hygiene* 27.2 (1928), pp. 113–159. ISSN: 00221724. DOI: 10.1017/S0022172400031879.

[77] Anthony JF Griffiths, William M Gelbart, Jeffrey H Miller, and Richard C Lewontin. "Bacterial conjugation". In: *Modern Genetic Analysis*. W. H. Freeman, 1999.

[78] Nigel D.F. Grindley, Katrine L. Whiteson, and Phoebe A. Rice. "Mechanisms of site-specific recombination". In: *Annual review of biochemistry* 75 (2006), pp. 567–605. DOI: 10.1146/ANNUREV.BIOCHEM.73.011303.073908.

[79] Elisabeth Grohmann, Günther Muth, and Manuel Espinosa. "Conjugative plasmid transfer in Gram-positive bacteria". In: *Microbiology and Molecular Biology Reviews* 67.2 (2003), pp. 277–301. ISSN: 1092-2172. DOI: 10.1128/mmbr.67.2.277-301.2003.

[80] Gérard Guédon, Virginie Libante, Charles Coluzzi, et al. "The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems". In: *Genes* 8.11 (2017), p. 337. ISSN: 2073-4425. DOI: 10.3390/GENES8110337.

[81] Julien Guglielmini, Leonor Quintais, Maria Pilar Garcillán-Barcia, et al. "The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation". In: *PLoS Genetics* 7.8 (2011), p. 1002222. ISSN: 15537390. DOI: 10.1371/journal.pgen.1002222.

[82] Jiarong Guo, Ben Bolduc, Ahmed A. Zayed, et al. "VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses". In: *Microbiome* 9.1 (2021), pp. 1–13. ISSN: 20492618. DOI: 10.1186/S40168-020-00990-Y/FIGURES/5.

[83] Radhey S. Gupta. "Origin of diderm (Gram-negative) bacteria: Antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes". In: *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 100.2 (2011), pp. 171–182. ISSN: 00036072. DOI: 10.1007/S10482-011-9616-8/FIGURES/2.

[84] Bernard Hallet and David J. Sherratt. "Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements". In: *FEMS Microbiology Reviews* 21.2 (1997), pp. 157–178. ISSN: 0168-6445. DOI: 10.1111/J.1574-6976.1997.TB00349.X.

[85] Holly L. Hamilton and Joseph P. Dillard. "Natural transformation of Neisseria gonorrhoeae: From DNA donation to homologous recombination". In: *Molecular Microbiology* 59.2 (2006), pp. 376–385. ISSN: 0950382X. DOI: 10.1111/j.1365-2958.2005.04964.x.

[86] Rasika M Harshey. "Transposable phage Mu". In: *Microbiology spectrum* 2.5 (2014). ISSN: 2165-0497. DOI: 10.1128/MICROBIOLSPEC.MDNA3-0007-2014.

[87] Finbarr Hayes. *The function and organization of plasmids.* 2003. DOI: 10.1385/1-59259-409-3:1.

[88] Jack A. Heinemann and George F. Sprague. "Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast". In: *Nature* 340.6230 (1989), pp. 205–209. ISSN: 00280836. DOI: 10.1038/340205a0.

[89] Roger W Hendrix. "Bacteriophage genomics". In: *Current opinion in microbiology* 6.5 (2003), pp. 506–511. ISSN: 1369-5274. DOI: 10.1016/J.MIB.2003.09.004.

[90] S. Henikoff and J. G. Henikoff. "Amino acid substitution matrices from protein blocks." In: *Proceedings of the National Academy of Sciences* 89.22 (1992), pp. 10915–10919. ISSN: 0027-8424. DOI: 10.1073/pnas.89.22.10915.

[91] Karin Hollricher. "" Species Don't Really Mean Anything in the Bacterial World "". In: *Lab times* (2007), pp. 22–25.

[92] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, et al. "A new view of the tree of life". In: *Nature Microbiology* 1.5 (2016), pp. 1–6. ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2016.48.

[93] Doug Hyatt, Gwo Liang Chen, Philip F. LoCascio, et al. "Prodigal: Prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinformatics* 11 (2010), p. 119. ISSN: 14712105. DOI: 10.1186/1471-2105-11-119.

[94] Paul Hyman and Stephen T. Abedon. "Bacteriophage host range and bacterial resistance". In: *Advances in Applied Microbiology* 70 (2010), pp. 217–248. ISSN: 0065-2164. DOI: 10.1016/S0065-2164(10)70007-1.

[95] Ethel Noland Jackson, David A. Jackson, and Robert J. Deans. "EcoRI analysis of bacteriophage P22 DNA packaging". In: *Journal of Molecular Biology* 118.3 (1978), pp. 365–388. ISSN: 0022-2836. DOI: 10.1016/0022-2836(78)90234-6.

[96] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(61)80072-7.

[97] Miten Jain, Sergey Koren, Karen H Miga, et al. "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nature Biotechnology* 36.4 (2018), pp. 338–345. ISSN: 1087-0156. DOI: 10.1038/nbt.4060.

[98] Calum Johnston, Bernard Martin, Gwennaele Fichant, et al. "Bacterial transformation: Distribution, shared mechanisms and divergent control". In: *Nature Reviews Microbiology* 12.3 (2014), pp. 181–196. ISSN: 17401526. DOI: 10.1038/nrmicro3199.

[99] Carola Kanz, Philippe Aldebert, Nicola Althorpe, et al. "The EMBL Nucleotide Sequence Database". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D29–D33. ISSN: 0305-1048. DOI: 10.1093/NAR/GKI098.

[100] K. Katoh and D. M. Standley. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability". In: *Molecular Biology and Evolution* 30.4 (2013), pp. 772–780. ISSN: 0737-4038. DOI: 10.1093/molbev/mst010.

[101] W. James Kent. "BLAT—The BLAST-Like Alignment Tool". In: *Genome Research* 12.4 (2002), p. 656. ISSN: 1088-9051. DOI: 10.1101/GR.229202.

[102] Takehiko Kenzaka, Katsuji Tani, and Masao Nasu. "High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level". In: *The ISME Journal 2010 4:5* 4.5 (2010), pp. 648–659. ISSN: 1751-7370. DOI: 10.1038/ismej.2009.145.

[103] Konstantinos T. Konstantinidis and James M. Tiedje. "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (2005), pp. 2567–2572. ISSN: 00278424. DOI: 10.1073/pnas.0409727102.

[104] Sergey Koren, Brian P. Walenz, Konstantin Berlin, et al. "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". In: *Genome Research* 27.5 (2017), pp. 722–736. ISSN: 1088-9051. DOI: 10.1101/gr.215087.116.

[105] Stephen C Kowalczykowski, Dan A Dixon, Angela K Eggleston, et al. "Biochemistry of Homologous Recombination in Escherichia coli". In: *Microbiological Reviews* 58.3 (1994), pp. 401–465.

[106] Stefan Kurtz, Adam Phillippy, Arthur L. Delcher, et al. "Versatile and open software for comparing large genomes." In: *Genome biology* 5.2 (2004), p. 12. ISSN: 14656914. DOI: 10.1186/gb-2004-5-2-r12.

[107] Andrei Kuzminov. "Recombinational Repair of DNA Damage in Escherichia coli and Bacteriophage λ". In: *Microbiology and Molecular Biology Reviews* 63.4 (1999), pp. 751–813.

[108] Chad Laing, Cody Buchanan, Eduardo N. Taboada, et al. "Pan-genome sequence analysis using Panseq: An online tool for the rapid analysis of core and accessory genomic regions". In: *BMC Bioinformatics* 11.1 (2010), pp. 1–14. ISSN: 14712105. DOI: 10.1186/1471-2105-11-461.

[109] Miriam Land, Loren Hauser, Se-Ran Jun, et al. "Insights from 20 years of bacterial genome sequencing". In: *Functional & Integrative Genomics* 15.2 (2015), p. 141. DOI: 10.1007/S10142-015-0433-4.

[110] Pascal Lapierre and J. Peter Gogarten. *Estimating the size of the bacterial pan-genome*. 2009. DOI: 10.1016/j.tig.2008.12.004.

[111] D. Laslett. "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences". In: *Nucleic Acids Research* 32.1 (2004), pp. 11–16. ISSN: 1362-4962. DOI: 10.1093/nar/gkh152.

[112] Esther M. Lederberg and Joshua Lederberg. "Genetic studies of lysogenicity in Escherichia coli". In: *Genetics* 38.1 (1953), pp. 51–64.

[113] Joshua Lederberg and E. L. Tatum. "Gene recombination in Escherichia coli". In: *Nature* 158.4016 (1946), p. 558. ISSN: 00280836. DOI: 10.1038/158558a0.

[114] Heewook Lee, Ellen Popodi, Haixu Tang, and Patricia L. Foster. "Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing". In: *Proceedings of the National Academy of Sciences* 109.41 (2012), E2774–E2783. ISSN: 0027-8424. DOI: 10.1073/PNAS.1210309109.

[115] Anthony Leewenhoeck. "An abstract of a letter from Mr. Anthony Leevvenhoeck at Delft, dated Sep. 17. 1683. Containing some microscopical observations, about animals in the scurf of the teeth, the substance call'd worms in the nose, the cuticula consisting of scales". In: *Philosophical Transactions of the Royal Society of London* 14.159 (1684), pp. 568–574. ISSN: 0261-0523. DOI: 10.1098/rstl.1684.0030.

[116] Elliot J Lefkowitz, Donald M Dempsey, Robert Curtis Hendrickson, et al. "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)". In: *Nucleic Acids Research* 46.D1 (2018), pp. D708–D717. ISSN: 0305-1048. DOI: 10.1093/NAR/GKX932.

[117] Nicole A. Lerminiaux and Andrew D.S. Cameron. "Horizontal transfer of antibiotic resistance genes in clinical environments". In: *Canadian Journal of Microbiology* 65.1 (2019), pp. 34–44. ISSN: 14803275. DOI: 10.1139/cjm-2018-0275.

[118] Petra Anne Levin and Esther R. Angert. "Small but Mighty: Cell Size and Bacteria". In: *Cold Spring Harbor Perspectives in Biology* 7.7 (2015), a019216. ISSN: 19430264. DOI: 10.1101/CSHPERSPECT.A019216.

[119] Cynthia A. Liebert, Ruth M. Hall, and Anne O. Summers. "Transposon Tn21 , flagship of the floating genome". In: *Microbiology and Molecular Biology Reviews* 63.3 (1999), pp. 507–522. DOI: 10.1128/MMBR.63.3.507-522.1999.

[120] David J. Lipman and William R. Pearson. "Rapid and sensitive protein similarity searches". In: *Science* 227.4693 (1985), pp. 1435–1441. ISSN: 0036-8075. DOI: 10.1126/science.2983426.

[121] Jacques Mahillon and Michael Chandler. "Insertion Sequences". In: *Microbiology and Molecular Biology Reviews* 62.3 (1998), pp. 725–774. DOI: 10.1128/MMBR.62.3.725-774.1998.

[122] Shoshana Marcus, Hayan Lee, and Michael C. Schatz. "SplitMEM: A graphical algorithm for pan-genome analysis with suffix skips". In: *Bioinformatics* 30.24 (2014), pp. 3476–3483. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu756.

[123] A. J. Martinez-Murcia, S. Benlloch, and M. D. Collins. "Phylogenetic Interrelationships of Members of the Genera Aeromonas and Plesiomonas as Determined by 16S Ribosomal DNA Sequencing: Lack of Congruence with Results of DNA-DNA Hybridizations". In: *International Journal of Systematic Bacteriology* 42.3 (1992), pp. 412–421. ISSN: 0020-7713. DOI: 10.1099/00207713-42-3-412.

[124] Barbara McClintock. "The origin and behavior of mutable loci in maize". In: *Proceedings of the National Academy of Sciences* 36.6 (1950), pp. 344–355. ISSN: 0027-8424. DOI: 10.1073/PNAS.36.6.344.

[125] Rajiv C. McCoy, Ryan W. Taylor, Timothy A. Blauwkamp, et al. "Illumina TruSeq Synthetic Long-Reads Empower De Novo Assembly and Resolve Complex, Highly-Repetitive Transposable Elements". In: *PLoS ONE* 9.9 (2014), e106689. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0106689.

[126] James O. McInerney, Alan McNally, and Mary J. O'Connell. "Why prokaryotes have pangenomes". In: *Nature Microbiology* 2.4 (2017), pp. 1–5. ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2017.40.

[127] Duccio Medini, Claudio Donati, Hervé Tettelin, et al. "The microbial pan-genome". In: *Current Opinion in Genetics & Development* 15.6 (2005), pp. 589–594. ISSN: 0959437X. DOI: 10.1016/j.gde.2005.09.006.

[128] F. Meinhardt, R. Schaffrath, and M. Larsen. "Microbial linear plasmids". In: *Applied Microbiology and Biotechnology* 47.4 (1997), pp. 329–336. ISSN: 01757598. DOI: 10.1007/s002530050936.

[129] Eric S. Miller, Elizabeth Kutter, Gisela Mosig, et al. "Bacteriophage T4 Genome". In: *Microbiology and Molecular Biology Reviews* 67.1 (2003), pp. 86–156. DOI: 10.1128/MMBR.67.1.86-156.2003.

[130] J A Mongold. "DNA repair and the evolution of transformation in Haemophilus influenzae." In: *Genetics* 132.4 (1992), pp. 893–898.

[131] E. P. Nawrocki and S. R. Eddy. "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29.22 (2013), pp. 2933–2935. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt509.

[132] Saul B. Needleman and Christian D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 00222836. DOI: 10.1016/0022-2836(70)90057-4.

[133] Anders Norman, Lars Hestbjerg Hansen, Qunxin She, and Søren Johannes Sørensen. "Nucleotide sequence of pOLA52: a conjugative IncX1 plasmid from Escherichia coli which enables biofilm formation and multidrug efflux". In: *Plasmid* 60.1 (2008), pp. 59–74. ISSN: 0147-619X. DOI: 10.1016/J.PLASMID.2008.03.003.

[134] R. T. Okinaka, K. Cloud, O. Hampton, et al. "Sequence and organization of pXO1, the large Bacillus anthracis plasmid harboring the anthrax toxin genes". In: *Journal of Bacteriology* 181.20 (1999), pp. 6509–6515.

[135] Brian D Ondov, Todd J Treangen, Páll Melsted, et al. "Mash: fast genome and metagenome distance estimation using MinHash." In: *Genome biology* 17.1 (2016), p. 132. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0997-x.

[136] Aharon Oren, David R. Arahal, Ramon Rosselló-Móra, et al. "Emendation of rules 5b, 8, 15 and 22 of the international code of nomenclature of prokaryotes to include the rank of phylum". In: *International Journal of Systematic and Evolutionary Microbiology* 71.6 (2021), p. 004851. ISSN: 14665034. DOI: 10.1099/IJSEM.0.004851/CITE/REFWORKS.

[137] Andrew J. Page, Carla A. Cummins, Martin Hunt, et al. "Roary: Rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* 31.22 (2015), pp. 3691–3693. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv421.

[138] Charles T. Parker, Brian J. Tindall, and George M. Garrity. "International code of nomenclature of Prokaryotes". In: *International Journal of Systematic and Evolutionary Microbiology* 69.1A (2019), S1–S111. DOI: 10.1099/IJSEM.0.000778/SIDEBYSIDE.

[139] Julian Parkhill, Mohammed Sebaihia, Andrew Preston, et al. "Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica". In: *Nature Genetics 2003 35:1* 35.1 (2003), pp. 32–40. ISSN: 1546-1718. DOI: 10.1038/ng1227.

[140] Adam R. Parks, Zaoping Li, Qiaojuan Shi, et al. "Transposition into replicating DNA occurs through interaction with the processivity factor". In: *Cell* 138.4 (2009), pp. 685–695. DOI: 10.1016/J.CELL.2009.06.011.

[141] John Patterson, Thidathip Wongsurawat, and Analiz Rodriguez. "A Glioblastoma Genomics Primer for Clinicians". In: *Medical Research Archives* 8.2 (2020). ISSN: 23751916. DOI: 10.18103/mra.v8i2.2034.

[142] Alexander Payne, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. "BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files". In: *Bioinformatics* 35.13 (2019), pp. 2193–2198. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty841.

[143] José R. Penadés, John Chen, Nuria Quiles-Puchalt, et al. "Bacteriophage-mediated spread of bacterial virulence genes". In: *Current Opinion in Microbiology* 23 (2015), pp. 171–178. ISSN: 1369-5274. DOI: 10.1016/J.MIB.2014.11.019.

[144] Ye Peng, Shanmei Tang, Dan Wang, et al. "MetaPGN: A pipeline for construction and graphical visualization of annotated pangenome networks". In: *GigaScience* 7.11 (2018), pp. 1–11. ISSN: 2047217X. DOI: 10.1093/gigascience/giy121.

[145] Amandine Perrin, Elise Larsonneur, Ainsley C. Nicholson, et al. "Evolutionary dynamics and genomic features of the Elizabethkingia anophelis 2015 to 2016 Wisconsin outbreak strain". In: *Nature Communications* 8 (2017), p. 15483. ISSN: 2041-1723. DOI: 10.1038/ncomms15483.

[146] Amandine Perrin and Eduardo P C Rocha. "PanACoTA: a modular tool for massive microbial comparative genomics." In: *NAR genomics and bioinformatics* 3.1 (2021), lqaa106. ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa106.

[147] Andrea Pitzschke and Heribert Hirt. *New insights into an old story: Agrobacterium-induced tumour formation in plants by plant transformation*. 2010. DOI: 10.1038/emboj.2010.8.

[148] Leslie A. Pratt and Roberto Kolter. "Genetic analysis of Escherichia coli biofilm formation: roles of flagella, motility, chemotaxis and type I pili". In: *Molecular Microbiology* 30.2 (1998), pp. 285–293. ISSN: 1365-2958. DOI: 10.1046/J.1365-2958.1998.01061.X.

[149] Morgan N Price, Adam P Arkin, and Eric J Alm. "The life-cycle of operons." In: *PLoS genetics* 2.6 (2006), e96. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020096.

[150] Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic Acids Research* 33.suppl_1 (2005), pp. D501–D504. ISSN: 0305-1048. DOI: 10.1093/NAR/GKI025.

[151] Michael Quail, Miriam E Smith, Paul Coupland, et al. "A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers". In: *BMC Genomics* 13.1 (2012), p. 341. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-341.

[152] R. J. Redfield. "Genes for breakfast: The have-your-cake and-eat-lt-too of bacterial transformation". In: *Journal of Heredity* 84.5 (1993), pp. 400–404. ISSN: 14657333. DOI: 10.1093/oxfordjournals.jhered.a111361.

[153] Wanda C Reygaert. "An overview of the antimicrobial resistance mechanisms of bacteria". In: *AIMS Microbiology* 4.3 (2018), p. 501. DOI: 10.3934/MICROBIOL.2018.3.482.

[154] Eduardo P. C Rocha, Emmanuel Cornet, and Bénédicte Michel. "Comparative and evolutionary analysis of the bacterial homologous recombination systems". In: *PLOS Genetics* 1.2 (2005), e15. ISSN: 1553-7404. DOI: 10.1371/JOURNAL.PGEN.0010015.

[155] Eduardo P. C. Rocha. "The replication-related organization of bacterial genomes". In: *Microbiology* 150.6 (2004), pp. 1609–1627. ISSN: 1350-0872. DOI: 10.1099/mic.0.26974-0.

[156] Mostafa Ronaghi, Mathias Uhlén, and Pål Nyrén. "A sequencing method based on real-time pyrophosphate". In: *Science* 281.5375 (1998), pp. 363–365. ISSN: 0036-8075. DOI: 10.1126/SCIENCE.281.5375.363.

[157] J. A. Ruiz-Masó, C. MachóN, L. Bordanaba-Ruiseco, et al. "Plasmid Rolling-Circle replication". In: *Microbiology Spectrum* 3.1 (2015). ISSN: 2165-0497. DOI: 10.1128/MICROBIOLSPEC.PLAS-0035-2014.

[158] F. Sanger, G. M. Air, B. G. Barrell, et al. "Nucleotide sequence of bacteriophage phi X174 DNA". In: *Nature* 265.5596 (1977), pp. 687–695. ISSN: 0028-0836. DOI: 10.1038/265687A0.

[159] Eric W. Sayers, Mark Cavanaugh, Karen Clark, et al. "GenBank". In: *Nucleic Acids Research* 49.D1 (2021), pp. D92–D96. ISSN: 0305-1048. DOI: 10.1093/NAR/GKAA1023.

[160] Petra Schicklmaier, Elisabeth Moser, Thomas Wieland, et al. "A comparative study on the frequency of prophages among natural isolates of Salmonella and Escherichia coli with emphasis on generalized transducers". In: *Antonie van Leeuwenhoek* 73.1 (1998), pp. 49–54. ISSN: 1572-9699. DOI: 10.1023/A:1000748505550.

[161] Bettina E. Schirrmeister, Muriel Gugger, and Philip C.J. Donoghue. "Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils". In: *Palaeontology* 58.5 (2015), pp. 769–785. ISSN: 1475-4983. DOI: 10.1111/PALA.12178.

[162] Gunnar Schröder and Erich Lanka. *The mating pair formation system of conjugative plasmids - A versatile secretion machinery for transfer of proteins and DNA*. 2005. DOI: 10.1016/j.plasmid.2005.02.001.

[163] Torsten Seemann. "Prokka: rapid prokaryotic genome annotation." In: *Bioinformatics* 30.14 (2014), pp. 2068–9. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu153.

[164] Ron Sender, Shai Fuchs, and Ron Milo. "Revised estimates for the number of human and bacteria cells in the body". In: *PLOS Biology* 14.8 (2016), e1002533. ISSN: 1545-7885. DOI: 10.1371/JOURNAL.PBIO.1002533.

[165] Basem Al-Shayeb, Rohan Sachdeva, Lin-Xing Chen, et al. "Clades of huge phages from across Earth's ecosystems". In: *Nature* 578.7795 (2020), pp. 425–431. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2007-4.

[166] L J Shimkets. "Social and developmental biology of the myxobacteria". In: *Microbiological Reviews* 54.4 (1990), pp. 473–501. ISSN: 0146-0749. DOI: 10.1128/mr.54.4.473-501.1990.

[167] Patricia Siguier, Lionel Gagnevin, and Michael Chandler. "The new IS1595 family, its relation to IS1 and the frontier between insertion sequences and transposons". In: *Research in Microbiology* 160.3 (2009), pp. 232–241. ISSN: 09232508. DOI: 10.1016/j.resmic.2009.02.003.

[168] Patricia Siguier, Edith Gourbeyre, and Mick Chandler. "Bacterial insertion sequences: their genomic impact and diversity". In: *FEMS Microbiology Reviews* 38.5 (2014), pp. 865–891. DOI: 10.1111/1574-6976.12067.

[169] Erich W. Six and Carol A.Connelly Klug. "Bacteriophage P4: a satellite virus depending on a helper such as prophage P2". In: *Virology* 51.2 (1973), pp. 327–344. ISSN: 0042-6822. DOI: 10.1016/0042-6822(73)90432-7.

[170] Jeffrey M. Skerker and Michael T. Laub. "Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus". In: *Nature Reviews Microbiology* 2.4 (2004), pp. 325–337. ISSN: 1740-1526. DOI: 10.1038/nrmicro864.

[171] Lloyd M. Smith, Jane Z. Sanders, Robert J. Kaiser, et al. "Fluorescence detection in automated DNA sequence analysis". In: *Nature* 321.6071 (1986), pp. 674–679. ISSN: 0028-0836. DOI: 10.1038/321674A0.

[172] T.F. Smith and M.S. Waterman. "Identification of common molecular subsequences". In: *Journal of Molecular Biology* 147.1 (1981), pp. 195–197. ISSN: 00222836. DOI: 10.1016/0022-2836(81)90087-5.

[173] Jonathan M. Solomon and Alan D. Grossman. *Who's competent and when: Regulation of natural genetic competence in bacteria*. 1996. DOI: 10.1016/0168-9525(96)10014-7.

[174] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. "Horizontal gene transfer: building the web of life". In: *Natures Reviews genetics* 16 (2015), pp. 472–482. DOI: 10.1038/nrg3962.

[175] P. F. Sparling. "Genetic transformation of Neisseria gonorrhoeae to streptomycin resistance." In: *Journal of bacteriology* 92.5 (1966), pp. 1364–1371. ISSN: 00219193. DOI: 10.1128/jb.92.5.1364-1371.1966.

[176] Martin Steinegger and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature Biotechnology* 35.11 (2017), pp. 1026–1028. ISSN: 15461696. DOI: 10.1038/nbt.3988.

[177] Alexander Sulakvelidze, Zemphira Alavidze, and J Glenn Morris. "Bacteriophage therapy". In: *Antimicrobial Agents and Chemotherapy* 45.3 (2001), pp. 649–659. DOI: 10.1128/AAC.45.3.649-659.2001.

[178] Tatiana Tatusova, Stacy Ciufo, Scott Federhen, et al. "Update on RefSeq microbial genomes resources". In: *Nucleic Acids Research* 43.D1 (2015), pp. D599–D605. ISSN: 0305-1048. DOI: 10.1093/NAR/GKU1062.

[179] W. H. Taylor and E. Juni. "Pathways for biosynthesis of a bacterial capsular polysaccharide. I. Characterization of the organism and polysaccharide". In: *Journal of Bacteriology* 81.5 (1961), p. 693. ISSN: 00219193. DOI: 10.1128/jb.81.5.688-693.1961.

[180] Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, et al. "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.39 (2005), pp. 13950–13955. ISSN: 00278424. DOI: 10.1073/pnas.0506758102.

[181] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. "Comparative genomics: the bacterial pan-genome". In: *Current Opinion in Microbiology* 11.5 (2008), pp. 472–477. ISSN: 13695274. DOI: 10.1016/j.mib.2008.09.006.

[182] The UniProt Consortium. "UniProt: a hub for protein information". In: *Nucleic Acids Research* 43.D1 (2015), pp. D204–D212. ISSN: 1362-4962. DOI: 10.1093/nar/gku989.

[183] Christopher M Thomas and David Summers. "Bacterial Plasmids". In: *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, 2008. DOI: 10.1002/9780470015902.a0000468.pub2.

[184] Julie D. Thompson, Desmond G. Higgins, and Toby J. Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". In: *Nucleic acids research* 22.22 (1994), pp. 4673–4680. ISSN: 0305-1048. DOI: 10.1093/NAR/22.22.4673.

[185] Harry A Thorpe, Sion C Bayliss, Samuel K Sheppard, and Edward J Feil. "Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria". In: *GigaScience* 7.4 (2018). ISSN: 2047-217X. DOI: 10.1093/gigascience/giy015.

[186] C. Titus Brown and Luiz Irber. "sourmash: a library for MinHash sketching of DNA". In: *The Journal of Open Source Software* 1.5 (2016), p. 27. ISSN: 2475-9066. DOI: 10.21105/joss.00027.

[187] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, et al. "Producing polished prokaryotic pangenomes with the Panaroo pipeline". In: *Genome Biology* 21.1 (2020), pp. 1–21. ISSN: 1474760X. DOI: 10.1186/S13059-020-02090-4/FIGURES/7.

[188] Marie Touchon, Aude Bernheim, and Eduardo PC Rocha. "Genetic and life-history traits associated with the distribution of prophages in bacteria". In: *The ISME Journal* 10.11 (2016), pp. 2744–2754. ISSN: 1751-7370. DOI: 10.1038/ismej.2016.47.

[189] Marie Touchon, Jorge A. Moura de Sousa, and Eduardo PC Rocha. "Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer". In: *Current Opinion in Microbiology* 38 (2017), pp. 66–73. ISSN: 1369-5274. DOI: 10.1016/J.MIB.2017.04.010.

[190] Marie Touchon, Amandine Perrin, Jorge André Moura de Sousa, et al. "Phylogenetic background and habitat drive the genetic diversification of Escherichia coli". In: *PLOS Genetics* 16.6 (2020). Ed. by Xavier Didelot, e1008866. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008866.

[191] Todd J. Treangen and Eduardo P.C. Rocha. "Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes". In: *PLOS Genetics* 7.1 (2011), e1001284. ISSN: 1553-7404. DOI: 10.1371/JOURNAL.PGEN.1001284.

[192] Patrick Trieu-Cuot, Cécile Carlier, Patrice Martin, and Patrice Courvalin. "Plasmid transfer by conjugation from Escherichia coli to Gram-positive bacteria". In: *FEMS Microbiology Letters* 48.1-2 (1987), pp. 289–294. ISSN: 15746968. DOI: 10.1111/j.1574-6968.1987.tb02558.x.

[193] Frederic William Twort. "An investigation on the nature of ultra-microscopic viruses". In: *The Lancet* 186.4814 (1915), pp. 1241–1243. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(01)20383-3.

[194] George Vernikos, Duccio Medini, David R. Riley, and Hervé Tettelin. "Ten years of pan-genome analyses". In: *Current Opinion in Microbiology* 23 (2015), pp. 148–154. ISSN: 13695274. DOI: 10.1016/j.mib.2014.11.016.

[195] Christian J.H. Von Wintersdorff, John Penders, Julius M. Van Niekerk, et al. "Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer". In: *Frontiers in Microbiology* 7 (2016), p. 173. ISSN: 1664302X. DOI: 10.3389/fmicb.2016.00173.

[196] Matthew K. Waldor and John J. Mekalanos. "Lysogenic conversion by a filamentous phage encoding cholera toxin". In: *Science* 272.5270 (1996), pp. 1910–1913. ISSN: 00368075. DOI: 10.1126/science.272.5270.1910.

[197] James D. Watson and Francis H. Crick. "The structure of DNA". In: *Cold Spring Harbor Symposia on Quantitative Biology* 18 (1951), pp. 123–131. ISSN: 0091-7451. DOI: 10.1101/SQB.1953.018.01.020.

[198] L. G. Wayne, D. J. Brenner, R. R. Colwell, et al. "Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics". In: *International Journal of Systematic and Evolutionary Microbiology* 37.4 (1987), pp. 463–464. ISSN: 1466-5026. DOI: 10.1099/00207713-37-4-463.

[199] Whitman WB, Coleman DC, and Wiebe WJ. "Prokaryotes: the unseen majority". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.12 (1998), pp. 6578–6583. ISSN: 0027-8424. DOI: 10.1073/PNAS.95.12.6578.

[200] Linda M. Weigel, Don B. Clewell, Steven R. Gill, et al. "Genetic analysis of a high-level vancomycin-resistant isolate of Staphylococcus aureus". In: *Science* 302.5650 (2003), pp. 1569–1571. DOI: 10.1126/SCIENCE.1090956.

[201] Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. "Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads". In: *PLOS Computational Biology* 13.6 (2017), e1005595. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005595.

[202] Carl R Woese, Otto Kandlert, and Mark L Wheelis. "Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya (Euryarchaeota/Crenarchaeota/kingdom/evolution)". In: *Proc. Nati. Acad. Sci. USA* 87 (1990), pp. 4576–4579.

[203] Rosanna C. T. Wright. "Understanding the structure and dynamics of bacteria-phage infection networks". PhD thesis. University of York, 2018.

[204] Seok-Hwan Yoon, Sung-min Ha, Jeongmin Lim, et al. "A large-scale evaluation of algorithms to calculate average nucleotide identity". In: *Antonie van Leeuwenhoek* 110.10 (2017), pp. 1281–1286. ISSN: 0003-6072. DOI: 10.1007/s10482-017-0844-4.

[205] Yun William Yu and Griffin M. Weber. "HyperMinHash: MinHash in LogLog space". In: *IEEE Transactions on Knowledge and Data Engineering* (2020), pp. 1–1. ISSN: 1041-4347. DOI: 10.1109/TKDE.2020.2981311.

[206] Daniel R. Zerbino. "Using the Velvet de novo Assembler for Short [U+2010] Read Sequencing Technologies". In: *Current Protocols in Bioinformatics* 31.1 (2010). ISSN: 1934-3396. DOI: 10.1002/0471250953.bi1105s31.

[207] XiaoFei Zhao. "BinDash, software for fast genome distance estimation on a typical personal laptop". In: *Bioinformatics* 35.4 (2019), pp. 671–673. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty651.

[208] Yongbing Zhao, Jiayan Wu, Junhui Yang, et al. "PGAP: Pan-genomes analysis pipeline". In: *Bioinformatics* 28.3 (2012), pp. 416–418. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr655.

[209] Zhemin Zhou, Jane Charlesworth, and Mark Achtman. "Accurate reconstruction of bacterial pan- And core genomes with PEPPAN". In: *Genome Research* 30.11 (2020), pp. 1667–1679. ISSN: 15495469. DOI: 10.1101/GR.260828.120/-/DC1.

[210] Norton D. Zinder and Joshua Lederberg. "Genetic exchange in Salmonella". In: *Journal of Bacteriology* 64.5 (1952), pp. 679–699. ISSN: 0021-9193. DOI: 10.1128/jb.64.5.679-699.1952.

ANNEXES

TIMELINE

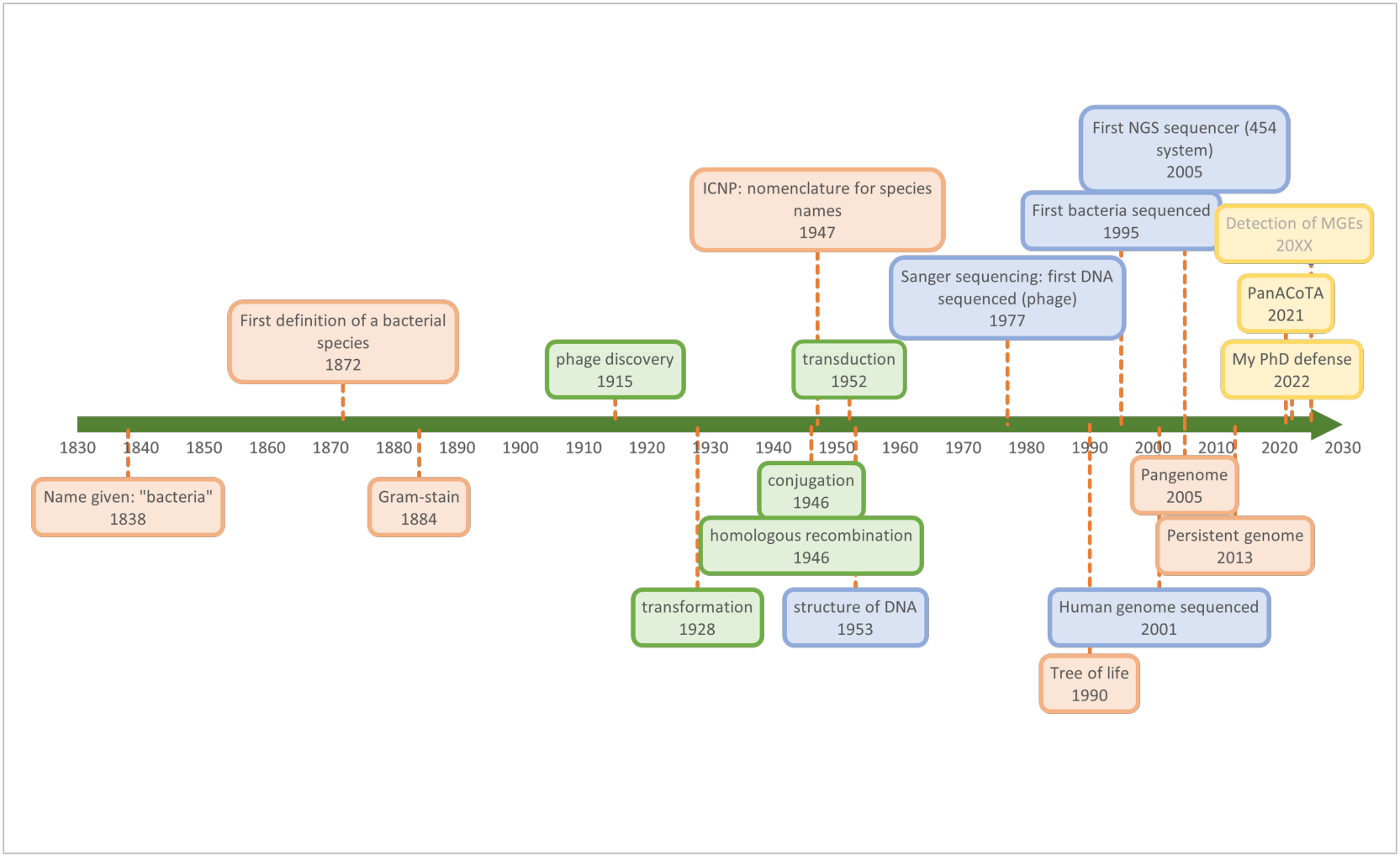


Figure 8.1: Timeline giving the main dates of the notions used in this PhD: notions related to species definition (orange), to HGT (green), to DNA (blue) and to my own work (yellow).

PPANGGOLIN: DEPICTING MICROBIAL DIVERSITY VIA A PARTITIONED PANGENOME GRAPH

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph

Guillaume Gautreau[1], Adelme Bazin[1], Mathieu Gachet[1], Rémi Planel[1¤a], Laura Burlot[1], Mathieu Dubois[1], Amandine Perrin[2,3], Claudine Médigue[1], Alexandra Calteau[1], Stéphane Cruveiller[1¤b], Catherine Matias[4], Christophe Ambroise[5], Eduardo P. C. Rocha[2], David Vallenet[1]*

1 LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, Evry, France, **2** Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, France, **3** Sorbonne Université, Collège doctoral, Paris, France, **4** Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Université de Paris, Centre National de la Recherche Scientifique, Paris, France, **5** Laboratoire de Mathématiques et Modélisation d'Evry, UMR CNRS 8071, Université d'Evry Val d'Essonne, Evry, France

¤a Current address: Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France
¤b Current address: PathoQuest SAS, BioPark – bâtiment B, 11 rue Watt, 75013 Paris, France
* vallenet@genoscope.cns.fr

Citation: Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. (2020) PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol 16(3): e1007732. https://doi.org/10.1371/journal.pcbi.1007732

Editor: Christos A. Ouzounis, CPERI, GREECE

Received: November 19, 2019

Accepted: February 12, 2020

Published: March 19, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007732

Data Availability Statement: Archaeal and bacterial genomes were downloaded from the NCBI FTP server (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank) 17 April 2019. Metagenome-Assembled Genomes were downloaded from

Abstract

The use of comparative genomics for functional, evolutionary, and epidemiological studies requires methods to classify gene families in terms of occurrence in a given species. These methods usually lack multivariate statistical models to infer the partitions and the optimal number of classes and don't account for genome organization. We introduce a graph structure to model pangenomes in which nodes represent gene families and edges represent genomic neighborhood. Our method, named PPanGGOLiN, partitions nodes using an Expectation-Maximization algorithm based on multivariate Bernoulli Mixture Model coupled with a Markov Random Field. This approach takes into account the topology of the graph and the presence/absence of genes in pangenomes to classify gene families into persistent, cloud, and one or several shell partitions. By analyzing the partitioned pangenome graphs of isolate genomes from 439 species and metagenome-assembled genomes from 78 species, we demonstrate that our method is effective in estimating the persistent genome. Interestingly, it shows that the shell genome is a key element to understand genome dynamics, presumably because it reflects how genes present at intermediate frequencies drive adaptation of species, and its proportion in genomes is independent of genome size. The graph-based approach proposed by PPanGGOLiN is useful to depict the overall genomic diversity of thousands of strains in a compact structure and provides an effective basis for very large scale comparative genomics. The software is freely available at https://github.com/labgem/PPanGGOLiN.

https://opendata.lifebit.ai/table/SGB. All analyses described here were run using PPanGGOLiN software (version 1.0). PPanGGOLiN source code is freely available from https://github.com/labgem/PPanGGOLiN under a CeCILL license. All relevant data are within the manuscript and its Supporting Information files.

Funding: This research was supported in part by the IRTELIS and Phare PhD programs of the French Alternative Energies and Atomic Energy Commission (CEA) for GG and AB respectively, the French Government "Investissements d'Avenir" programs (namely FRANCE GENOMIQUE [ANR-10-INBS-09-08], the INSTITUT FRANÇAIS DE BIOINFORMATIQUE [ANR-11-INBS-0013], and the Agence Nationale de la Recherche [Projet ANR-16-CE12-29 for EPCR]). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Microorganisms have the greatest biodiversity and evolutionary history on earth. At the genomic level, it is reflected by a highly variable gene content even among organisms from the same species which explains the ability of microbes to be pathogenic or to grow in specific environments. We developed a new method called PPanGGOLiN which accurately represents the genomic diversity of a species (i.e. its pangenome) using a compact graph structure. Based on this pangenome graph, we classify genes by a statistical method according to their occurrence in the genomes. This method allowed us to build pangenomes even for uncultivated species at an unprecedented scale. We applied our method on all available genomes in databanks in order to depict the overall diversity of hundreds of species. Overall, our work enables microbiologists to explore and visualize pangenomes alike a subway map.

Introduction

The analyses of the gene repertoire diversity of species—their pangenome—have many applications in functional, evolutionary, and epidemiological studies [1, 2]. The core genome is defined as the set of genes shared by all the genomes of a taxonomic unit (generally a species) whereas the accessory (or variable) genome contains genes that are only present in some genomes. The latter is crucial to understand bacterial adaptation as it contains a large repertoire of genes that may confer distinct traits and explain many of the phenotypic differences across species. Most of these genes are acquired by horizontal gene transfer (HGT) [3]. This usual dichotomy between core and accessory genomes does not consider the diverse ranges of gene frequencies in a pangenome. The main problem in using a strict definition of the core genome is that its size decreases as more genomes are added to the analysis [4] due to gene loss events and technical artifacts (i.e. sequencing, assembly or annotation issues). As a consequence, it was proposed in the field of synthetic biology to focus on persistent genes, i.e. those conserved in a large majority of genomes [5]. The persistent genome is also called the soft core [6], the extended core [7, 8] or the stabilome [9]. These definitions advocate for the use of a threshold frequency of a gene family within a species above which it is considered as *de facto* core gene. Persistent gene families are usually defined as those present in a range comprised between 90% [10] and 99% [11] of the strains in the species. This approach addresses some problems of the original definition of core genome but requires the setting of an appropriate threshold. The gene frequency distribution in pangenomes is extensively documented [7, 8, 12–16]. Due to the variation in the rates of gene loss and gain of genes, the gene frequencies tend to show an asymmetric U-shaped distribution regardless of the phylogenetic level and the clade considered (with the exception of few species having non-homogeneous distributions as described in [17]). Thereby, as proposed by Koonin and Wolf [12] and formally modeled by Collins and Higgs [14], the pangenome can be split into 3 classes: (1) persistent genome, for the gene families present in almost all genomes; (2) shell genome, for gene families present at intermediate frequencies in the species; (3) cloud genome, for gene families present at low frequency in the species.

The study of pangenomes in microbiology now relies on the comparison of hundreds to thousands of genomes of a single species. The analysis of this massive amount of data raises computational and algorithmic challenges that can be tackled because genomes within a species have many homologous genes and it is possible to design new compact ways of representing and manipulating this information. As suggested by Chan *et al.* [18], a consensus representation of multiple genomes would provide a better analytical framework than using individual

reference genomes. Among others, this proposition has led to a paradigm shift from the usual linear representation of reference genomes to a representation as variation graphs (also named "genome graphs" or "pangenome graphs") bringing together all the different known variations as multiple alternative paths. Methods [19–21] have been developed aiming at factorizing pangenomes at the genome sequence-level to capture all the nucleotide variations in a graph that enables variant calling and improves the sensitivity of the read mapping (summarized in [22]).

The method presented here, named PPanGGOLiN (Partitioned PanGenome Graph Of Linked Neighbors), introduces a new representation of the gene repertoire variation as a graph, where each node represents a family of homologous genes and each edge indicates a relation of genetic contiguity. PPanGGOLiN fills the gap between the standard pangenomic approach (that uses a set of independent and isolated gene families) and sequence-level pangenome graph (as reviewed in [23]). The interest of a gene-level graph compared to a sequence graph is that it provides a much more compact structure in clades where gene gains and losses are the major drivers of adaptation. This comes at the cost of disregarding polymorphism in genes and ignoring variation in intergenic regions and introns. However, the genomes of prokaryotes have very small intergenic regions and are almost devoid of introns justifying a focus on the variation of gene repertoires [12], which can be complemented by analysis of intergenic and intragenic polymorphism. PPanGGOLiN uses a new statistical model to classify gene families into persistent, cloud, and one or several shell partitions. To the best of our knowledge three statistical methods are available to partition a pangenome. Two of them use probabilistic models that partition dichotomously the pangenome only into core and accessory components [24, 25]. Conversely, the method proposed and implemented by Snipen *et al.* [26, 27] (micropan R package) classifies a pangenome in K partitions using a Binomial Mixture Model relying on gene family frequencies. Unlike these three methods, PPanGGOLiN is not based on frequencies but combines both the patterns of occurrence of gene families and the pangenome graph topology to perform the classification. In the following sections we present an overview of the method, an illustration of a pangenome graph and then the partitioning of a large set of prokaryotic species from GenBank. We evaluate the relevance of the persistent genome computed by PPanGGOLiN in comparison to the classical soft core genome. Next, we illustrate the importance of the shell structure and dynamics in the study of the evolution of microbial genomes. Finally, we compare GenBank results to the ones obtained with Metagenome-Assembled Genomes (MAGs) to validate the use of PPanGGOLiN for metagenomic applications.

Results and discussion

Overview of the PPanGGOLiN method

PPanGGOLiN builds pangenomes for large sets of prokaryotic genomes (i.e. several thousands) through a graphical model and a statistical method to classify gene families into three classes: persistent, cloud, and one or several shell partitions. It uses as input a set of annotated genomes with their coding regions classified in homologous gene families. As depicted in Fig 1, PPanGGOLiN integrates information on protein-coding genes and their genomic neighborhood to build a graph where each node is a gene family and each edge is a relation of genetic contiguity (two families are linked in the graph if they contain genes that are neighbors in the genomes). Thanks to this graphical model, the structure of the pangenome is resilient to fragmented assemblies: an assembly gap in one genome can be offset by information from other genomes, thus maintaining the link in the graph. To partition this graph, we established a statistical model taking into consideration that persistent genes share conserved genomic organizations along genomes (i.e. synteny conservation) [28] and that horizontally transferred genes

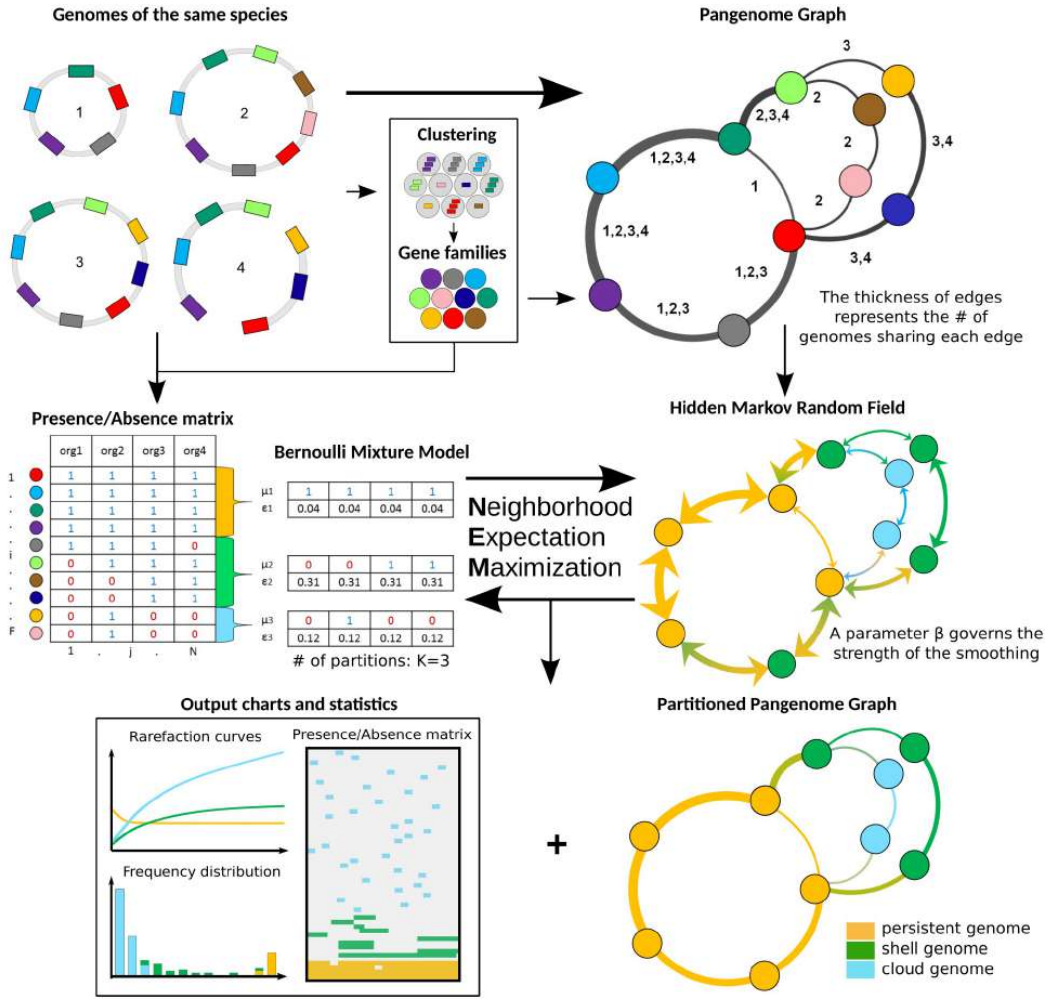


Fig 1. Flowchart of PPanGGOLiN on a toy example of 4 genomes. The method requires annotated genomes of the same species with their genes clustered into homologous gene families. Annotations and gene families can be predicted by PPanGGOLiN or directly provided by the user. Based on these inputs, a pangenome graph is built by merging homologous genes and their genomic links. Nodes represent gene families and edges represent genomic neighborhood. The edges are labeled by identifiers of genomes sharing the same gene neighborhood. In parallel, gene families are encoded as a presence/absence matrix that indicates for each family whether or not it is present in the genomes. The pangenome is then divided into *K* partitions (*K* = 3 in this example) by estimating the best partitioning parameters through an Expectation-Maximization algorithm. The method involves the maximization of the likelihood of a multivariate Bernoulli Mixture Model taking into account the constraint of a Markov Random Field (MRF). The MRF network is given by the pangenome graph and it favors two neighbors to be more likely classified in the same partition. At the end of this iterative process, PPanGGOLiN returns a partitioned pangenome graph where persistent, shell and cloud partitions are overlaid on the neighborhood graph. In addition, many tables, charts and statistics are provided by the software. The number of partitions (*K*) can either be provided by the user or determined by the algorithm.

https://doi.org/10.1371/journal.pcbi.1007732.g001

(i.e. shell and cloud genes) tend to insert preferentially in a few chromosomal regions (hotspots) [29]. Thereby, PPanGGOLiN favors two gene families that are consistent neighbors in the graph to be more likely classified in the same partition. This is achieved by a hidden Markov Random Field (MRF) whose network is given by the pangenome graph. In parallel, the

pangenome is also represented as a binary Presence/Absence (P/A) matrix where the rows correspond to gene families and the columns to genomes. Values are 1 for the presence of at least one member of the gene family and 0 otherwise. This P/A matrix is modeled by a multivariate Bernoulli Mixture Model (BMM). Its parameters are estimated via an Expectation-Maximization (EM) algorithm taking into account the constraints imposed by the MRF. Each gene family is then associated to its closest partition according to the BMM. This results in a partitioned pangenome graph made of nodes that are classified as either persistent, shell or cloud. The strength of the MRF constraints increases according to a parameter called β (if $\beta = 0$, the effect of the MRF is disabled and the partitioning only relies on the P/A matrix) and it depends on the weight of the edges of the pangenome graph which represents the number of gene pairs sharing the neighborhood. Another originality of our method is that, even if the number of partitions (K) is estimated to be equal to 3 (persistent, shell, cloud) in most cases (see 'Analyses of the most represented species in databanks' section), more partitions can be used if the pangenome matrix contains several contrasted patterns of P/A. These additional partitions are considered to belong to the shell genome and reflect a heterogeneous structure of the shell (see 'Shell structure and dynamics' section).

Illustration of a partitioned pangenome graph depicting the *Acinetobacter baumannii* species

We computed the pangenome of 3 117 *Acinetobacter baumannii* genomes from GenBank using PPanGGOLiN. For the persistent, shell and cloud genomes, we obtained 3 084, 1 529 and 64 833 gene families, respectively. If we compare our results with those of Chan *et al.* study [18], the size of the persistent genome predicted by PPanGGOLiN is included in their soft core estimation ranging from 2 833 (95% of presence) to 3 126 (75% of presence) gene families using 249 *A. baumannii* genomes. On the partitioned pangenome graph built with PPanGGOLiN (Fig 2), the gene families classified as persistent (orange nodes) correspond to the conserved paths that are interrupted by many islands composed of shell (green nodes) and cloud genomes (blue nodes). These islands appear to be frequently inserted in hotspots of the persistent genome thus pinpointing regions of genome plasticity. The average node degree within the same partition is 2.80 for the persistent genome while the shell genome has a higher average degree (3.95, P = 5.0e-6 with a bilateral unpaired 2-sample Student's t test) and the cloud a lower one (1.97, P = 3.3e-40 with the same test). The shell genome is the most diversified in terms of network topology with many interconnections between families reflecting a mosaic composition of regions from different HGT events [29]. The major part of the cloud has a shell-like graph topology with a large connected component containing 60% of the nodes. In addition, the cloud also contains isolated components that are nearly linear (3 606 components having on average 4.25 nodes) and singletons (10 575 nodes), presumably because it includes very recently acquired genetic material. Finally, large families of mobile genes, mostly transposable elements, can be easily detected because they constitute hubs (i.e. highly connected nodes) in the graph. They vary rapidly their genetic neighborhoods and can be found in multiple loci.

As an example of a more detailed analysis that can be done using the graph, a zoom on a region containing the genes required for the synthesis of capsular polysaccharides is highlighted in Fig 2. *A. baumannii* strains are involved in numerous nosocomial infections and their capsule plays key roles in the overall fitness and pathogenicity. Indeed, it protects the bacteria against environmental stresses, host immune responses and can confer resistance to some antimicrobial compounds [30]. Over one hundred distinct capsule types and their corresponding genomic organization have been reported in *A. baumannii* [31]. A zoom on this

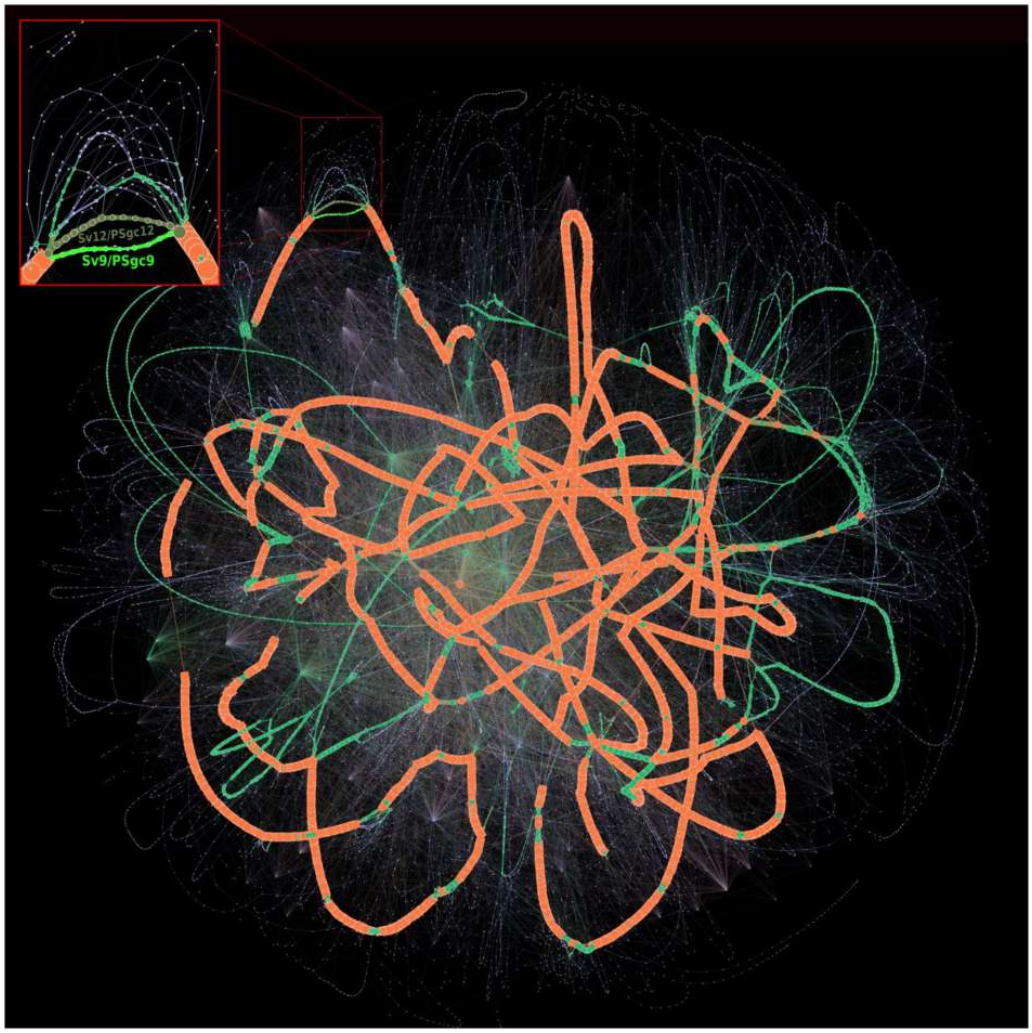


Fig 2. Partitioned pangenome graph of 3 117 *Acinetobacter baumannii* genomes. This partitioned pangenome graph of PPanGGOLiN displays the overall genomic diversity of 3 117 *Acinetobacter baumannii* strains from GenBank. Edges correspond to genomic colocalization and nodes correspond to gene families. The thickness of the edges is proportional to the number of genomes sharing that link. The size of the nodes is proportional to the total number of genes in each family. The edges between persistent, shell and cloud nodes are colored in orange, green and blue, respectively. Nodes are colored in the same way. The edges between gene families belonging to different partitions are shown in mixed colors. For visualization purposes, gene families with less than 20 genes are not shown on this figure although they comprise 84.68% of the nodes (families mostly composed of a single gene). The frame in the upper left corner shows a zoom on a branching region where multiple alternative shell and cloud paths are present in the species. This region is involved in the synthesis of the major polysaccharide antigen of *A. baumannii*. The two most frequent paths (Sv12/PSgc12 and Sv9/PSgc9) are highlighted in khaki and fluo green. The Gephi software (https://gephi.org) [32] with the ForceAtlas2 algorithm [33] was used to compute the graph layout with the following parameters: Scaling = 8000, Stronger Gravity = True, Gravity = 4.0, Edge Weight influence = 1.3.

https://doi.org/10.1371/journal.pcbi.1007732.g002

region of the graph shows a wide variety of combinations of genes for the synthesis of capsular polysaccharides. Based on the 3 117 *A. baumannii* genomes available in GenBank, we detected 229 different paths, sharing many common portions, but only a few are conserved in the species (only 24 paths are covered by more than 10 genomes). Among them, two alternative shell

paths seem to be particularly conserved (from the *gnaA* to the *weeH* genes in the figure 3 of [31]). Based on the nomenclature of [31], one (colored in khaki green in the Fig 2) corresponds to the serovar called PSgc12, contains 14 gene families of the shell genome and is fully conserved in 581 genomes. The other (colored in fluo green in the Fig 2) corresponds to the serovar PSgc9 (equivalent to PSgc7), contains 11 gene families of the shell genome and is fully conserved in 408 genomes. This analysis illustrates how the partitioned pangenome graph of PPanGGOLiN can be useful to study the plasticity of genomic regions. Thanks to its compact structure in which genes are grouped into families while preserving their genomic neighborhood information, it summarizes the diversity of thousands of genomes in a single picture and allows effective exploration of the different paths among regions or genes of interest.

Analyses of the most represented species in databanks

We used PPanGGOLiN to analyze all prokaryotic species of GenBank for which at least 15 genomes were available. This is the minimal number of genomes we recommend to ensure a relevant partitioning. The quality of the genomes was evaluated before their integration in the graph to avoid taxonomic assignation errors and contamination that can have a major impact on the analysis of pangenomes (see Materials and methods). This resulted in a dataset of 439 species pangenomes, whose metrics are available in S1 File. We focused our analysis on the 88 species containing at least 100 genomes (Fig 3). This data was used for in-depth analysis of persistent and shell genomes (see the two next sections). Proteobacteria, Firmicutes and Actinobacteria are the most represented phyla in this dataset and comprise a variety of species, genome sizes and environments. In contrast, Spirochaetes, Bacteroidetes and Chlamydiae phyla are represented by only one or two species (*Leptospira interrogans*, *Bacteroides fragilis*, *Flavobacterium psychrophilum* and *Chlamydia trachomatis*). For each species, we computed the median and interquartile range of persistent, shell and cloud families in the genomes. As expected, we observed a large variation in the range of these values: from pathogens with reduced genomes such as *Bordetella pertussis* or *C. trachomatis* which contain only a small fraction of variable gene families (less than ≈5% of shell and cloud genomes) to commensal or environmental bacteria such as *Bifidobacterium longum* and *Burkholderia cenocepacia* whose shell represents more than ≈35% of the genome. Furthermore, for a few species the number of estimated partitions (K) is greater than 3 (11 out of 88 species), especially for those with a higher fraction of shell genome. Hence, our method provides a statistical justification for the use of three partitions as a default in pangenome analyses, while indicating that species with large shell content might be best modeled using more partitions (see 'Shell structure and dynamics' section).

Estimation of the persistent genome in comparison to the soft core approach

To demonstrate the added value of PPanGGOLiN, we compared our statistical method to a classical approach where persistent genes are those present in at least 95% of the genomes (generally called the soft core approach). Indeed, this threshold is very often used in pangenomic studies probably because it is the default parameter in Roary [34] which is to date the most cited software to build bacterial pangenomes. In the 88 studied species, the number of persistent gene families is greater than or equal to the soft core with an average of 11% (SD = 9%) of additional families (see Fig 3 and S1 File). Furthermore, persistent gene families include those of the soft core with the exception of very few gene families (12 families in total for all studied species). The gene family frequencies in each of the 88 pangenomes are available in S1 Fig. For four species, *Pseudomonas stutzeri*, *Clostridium perfringens*, *Clostridium*

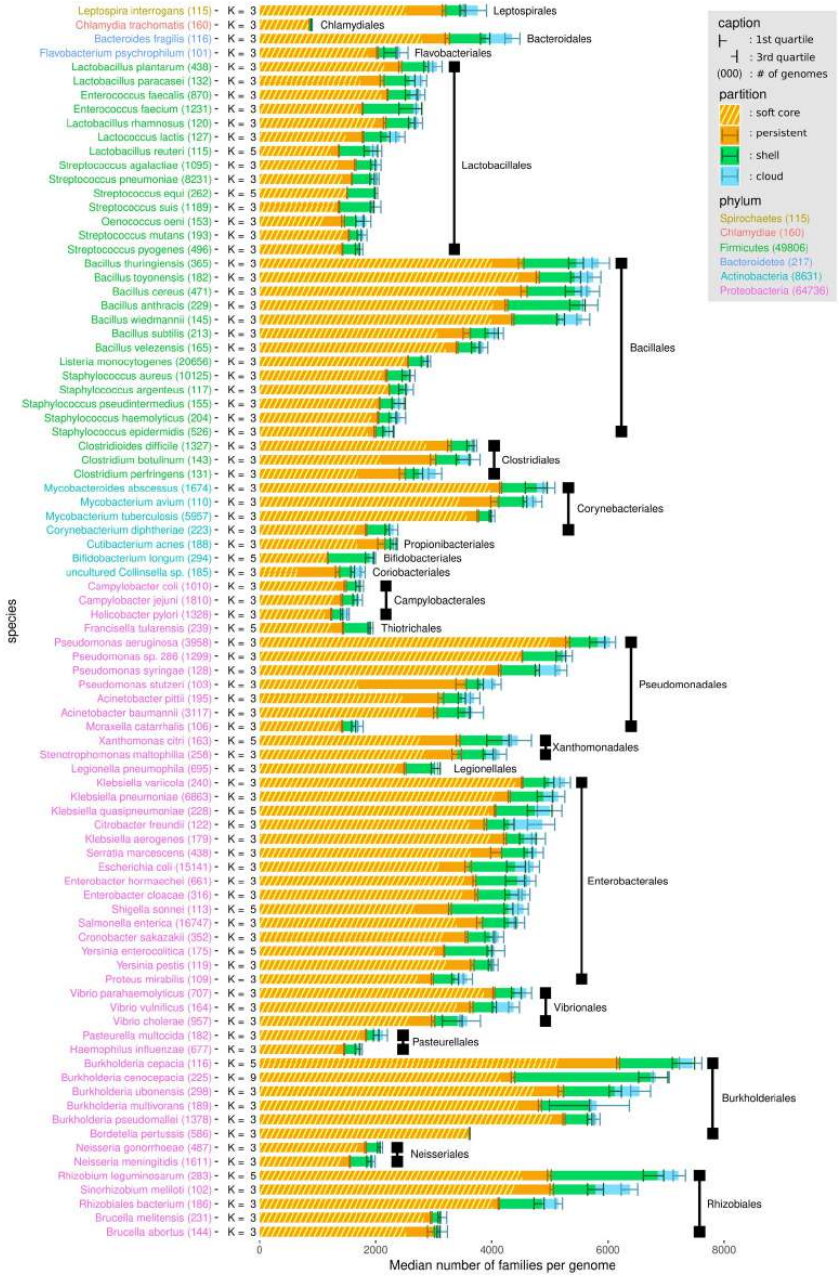


Fig 3. Distribution of PPanGGOLiN partitions in the genomes of the most represented species in GenBank. Each horizontal bar shows the median number of gene families per genome among the different PPanGGOLiN partitions (persistent, shell and cloud) in the 88 most represented species in GenBank (having at least 100 genomes). The error bars represent the interquartile ranges. Hatched areas on the persistent genome bars show the median number of gene families for the soft core (⩾95% of presence). The species names are colored according to their phylum and sorted by taxonomic order and then by decreasing cumulative bar size. Next to the species names, the number of genomes is indicated in brackets and the number of partitions (*K*) that was automatically determined by PPanGGOLiN is also shown.

https://doi.org/10.1371/journal.pcbi.1007732.g003

botulinum and *Colinsella sp*., the size of the soft core genome is unexpectedly small and represents less than 55% of the genomes whereas it is above 75% for the PPanGGOLiN persistent. For the first three species, this could be due to sampling effects and species heterogeneity. For the last one (*Colinsella sp*.), this could be explained by the fact that the species is made of incomplete genomes from metagenomes (i.e. MAGs) that were submitted as complete genomes in GenBank.

For an in-depth comparison of these approaches, we performed multiple resamplings of the genome dataset for each species in order to measure the variability of the pangenome metrics and the impact of genome sampling according to an increasing number of genomes considered in the analyses (hereafter called rarefaction curves, see Materials and methods for more details and S2 Fig as an example for *Lactobacillus plantarum*). These rarefaction curves indicate whether the number of families tends to stabilize, increase or decrease. To this end, the curves were fit with the Heaps' law where γ represents the growth tendency [35] (hereafter called γ-tendency). The persistent component of a pangenome is supposed to stabilize after the inclusion of a certain number of genomes, which means it has a γ-tendency close to 0. In addition, interquartile range (IQR) areas along the rarefaction curves were computed to estimate the variability of the predictions in relation to the sampling. Small IQR areas mean that the predictions are stable and resilient to sampling. Using these metrics, the PPanGGOLiN predictions of the persistent genome were evaluated in comparison to the soft core approach.

We observed that the γ-tendency of the PPanGGOLiN persistent is closer to 0 than that of the soft core approach (mean of absolute γ-tendency = 9.1e-3 versus 2.5e-2, P = 1.5e-9 with a one-sided paired 2-sample Student's t-test) with a lower standard deviation error too (mean = 5.3e-04 versus 2.1e-03, P = 9.5e-11 with one-sided paired 2-sample Student's t-test) (see Fig 4 and S1 File). A major problem of the soft core approach is that the γ-tendency is high for many species (32 species have a γ-tendency above 0.025), suggesting that the size of the persistent genome is not stabilized and tends to be underestimated. Besides, the IQR area of the PPanGGOLiN prediction is far below the one of the soft core genome (mean = 4906.6 versus 11645.9, P = 8.9e-07 with a unilateral paired 2-sample Student's t test). It can be partially explained because the threshold used in the soft core method induces a 'stair-step effect' along the rarefaction curves depending on the number of genomes sampled. This is illustrated on S2 Fig showing a step every 20 genomes (i.e. corresponding to $20 = \frac{100}{100-95}$ where 95% is the threshold of presence used) on the soft core curve of *L. plantarum*. We found a total of 20 species having atypical values of γ-tendency (absolute value above 0.05) and/or IQR area (above 15 000) for the soft core and only 2 species for the persistent genome of PPanGGOLiN, which are *Bacillus anthracis* and *Burkholderia cenocepacia*. For *B. cenocepacia*, it could be explained by the high heterogeneity of its shell (see next section), which is made of several partitions and complicates its distinction from the persistent genome during the process of partitioning. For *Bacillus anthracis*, the source of variability to define the persistent genome is a result of an incorrect taxonomic assignation in GenBank of about 17% of the genomes that are, according to the Genome Taxonomy DataBase (GTDB) [36], actually *B. cereus* or *B. thuringiensis*. This issue was not detected by our taxonomy control procedure because these species are at the boundary of the conspecific genomic distance threshold used (see Materials and methods). Some of persistent gene families of *bona fide B. anthracis* may therefore shift between persistent or shell partitions depending on the resampling. Excluding these misclassified genomes, we predicted a larger persistent genome than the one of the initial full set of genomes (about a thousand gene families more) with a γ-tendency much closer to 0 (-0.017 versus a γ-tendency of 0.036 for the soft core genome) and a lower IQR area (8367.0 vs 32167.1). Altogether, these results suggest that our approach provides a more robust partitioning of gene families in the

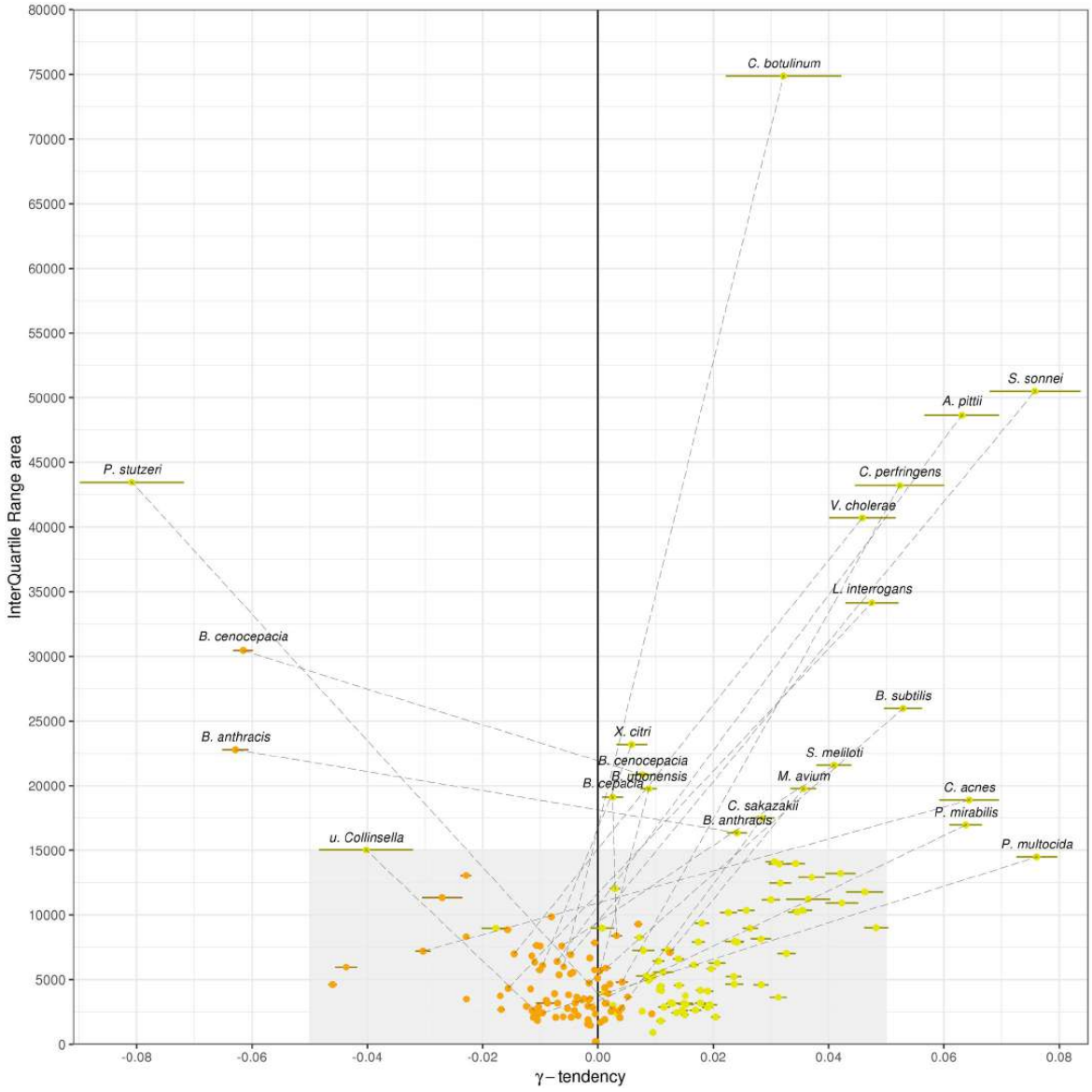


Fig 4. γ-tendencies and IQR areas of the persistent and the soft core rarefaction curves. Each of the 88 most abundant species in GenBank are represented by two points: orange points correspond to the PPanGGOLiN persistent values and yellow points to the ones of the soft core (⩾95% of presence). A dashed line connects the 2 points if either the soft core or the persistent values are not in the range of the grey area ($-0.05 \leqslant \gamma \leqslant 0.05$ and $0 \leqslant IQR_{area} \leqslant 15000$). The colored horizontal bars show the standard errors of the fitting of rarefaction curves via the Heaps' law.

https://doi.org/10.1371/journal.pcbi.1007732.g004

persistent genome than the use of arbitrary thresholds. Indeed, the statistical method behind PPanGGOLiN uses directly the information of the gene family P/A whereas the soft core is based only on frequency values. PPanGGOLiN can then classify families with similar frequencies in different partitions by distinguishing them according to their pattern of P/A in the

matrix and their genomic neighborhood. The main drawback of using family frequency to partition pangenomes is that even if it was possible to determine the best threshold for each species it would still not take into account that some persistent gene families may have atypically low frequency. This may be due to high gene losses in the population or technical reasons like belonging to a genomic region that is difficult to assemble (i.e. genes that are missing or fragmented in draft genome assemblies).

Shell structure and dynamics

Two types of pangenome evolution dynamics are generally distinguished: open pangenomes and closed ones [1, 2, 35]. From rarefaction curves, the dynamics of pangenomes can be assessed using the γ-tendency of a Heaps' law fitting (see Materials and methods). A low γ-tendency means a rather closed pangenome whereas a higher γ-tendency means a rather open pangenome. A closed pangenome rigorously means a stabilized pangenome and we found no species obeying this strict criterion (that is to say $\gamma = 0$). This suggests that instead of using binary classifications for pangenomes, it is more useful to quantify the degree of openness of pangenomes given the flux of horizontal gene transfers and gene loss [7]. We computed rarefaction curves for the 88 studied species and determined the γ-tendency for different pangenome components (see S1 File and S3 Fig). The distribution of γ values of the PPanGGOLiN shell genome shows a greater amplitude of values than the other components of the pangenome such as the whole pangenome or the accessory component. This indicates that the main differences in terms of genome dynamics between species seem to reside in the shell genome.

As expected, we found a positive correlation (Spearman's $\rho = 0.46$, P = 8.2e-06) between the total number of shell gene families in a species and the γ-tendency of the shell (S4 Fig). This means that species with high γ-tendency do accumulate genes that are maintained and exchanged in the population at relatively low frequencies, suggesting they may be locally adaptive. More surprisingly, although one could expect that larger genomes have a larger fraction of variable gene repertoires, the fraction of shell and cloud genes per genome does not correlate with the genome size (Spearman's $\rho = 0.007$, P = 0.95, Fig 5). The results remain qualitatively similar when analyzing the shell or the cloud separately (see S5 and S6 Figs). During this analysis, we noticed that, among host-associated bacteria with relatively small genomes (between ≈2000 and ≈3000 genes), three species (*Bifidobacterium longum*, *Enterococcus faecium* and *Streptococcus suis*) have a high fraction of shell genes ($>$ 28%) but low shell γ-tendency. Two of them (*B. longum* and *E. faecium*) are found in the gut of mammals and the third (*S. suis*) in the upper respiratory tract of pigs. They differ from other host-associated species in our dataset that are mainly human pathogens (e.g. bacteria of the genus *Corynebacterium*, *Neisseria*, *Streptococcus*, *Staphylococcus*) and have a low fraction of shell genes ($<$ 20%). It is possible that these three species have specialized in their ecological niches while maintaining a large and stable pool of shell genes for their adaptation to environmental stress. Further analysis would be required to confirm this hypothesis.

We then investigated the importance of the phylogeny of the species on the patterns of P/A of the shell gene families (shell structure). To this end, Spearman's rank correlations were computed between a Jaccard distance matrix generated on the basis of patterns of P/A of the shell gene families and a genomic distance obtained by Mash pairwise comparisons between genomes [37]. Mash distances were shown to be a good estimate of evolutionary distances for closely related genomes [38]. This correlation was examined in relation to the fraction of gene families that are part of the shell genome for each species (Fig 6). We observed that species with a high fraction of shell ($>$ 20% of their genome) have a shell structure that is mainly explained by the species phylogeny (i.e. shell P/A are highly correlated with genomic distances,

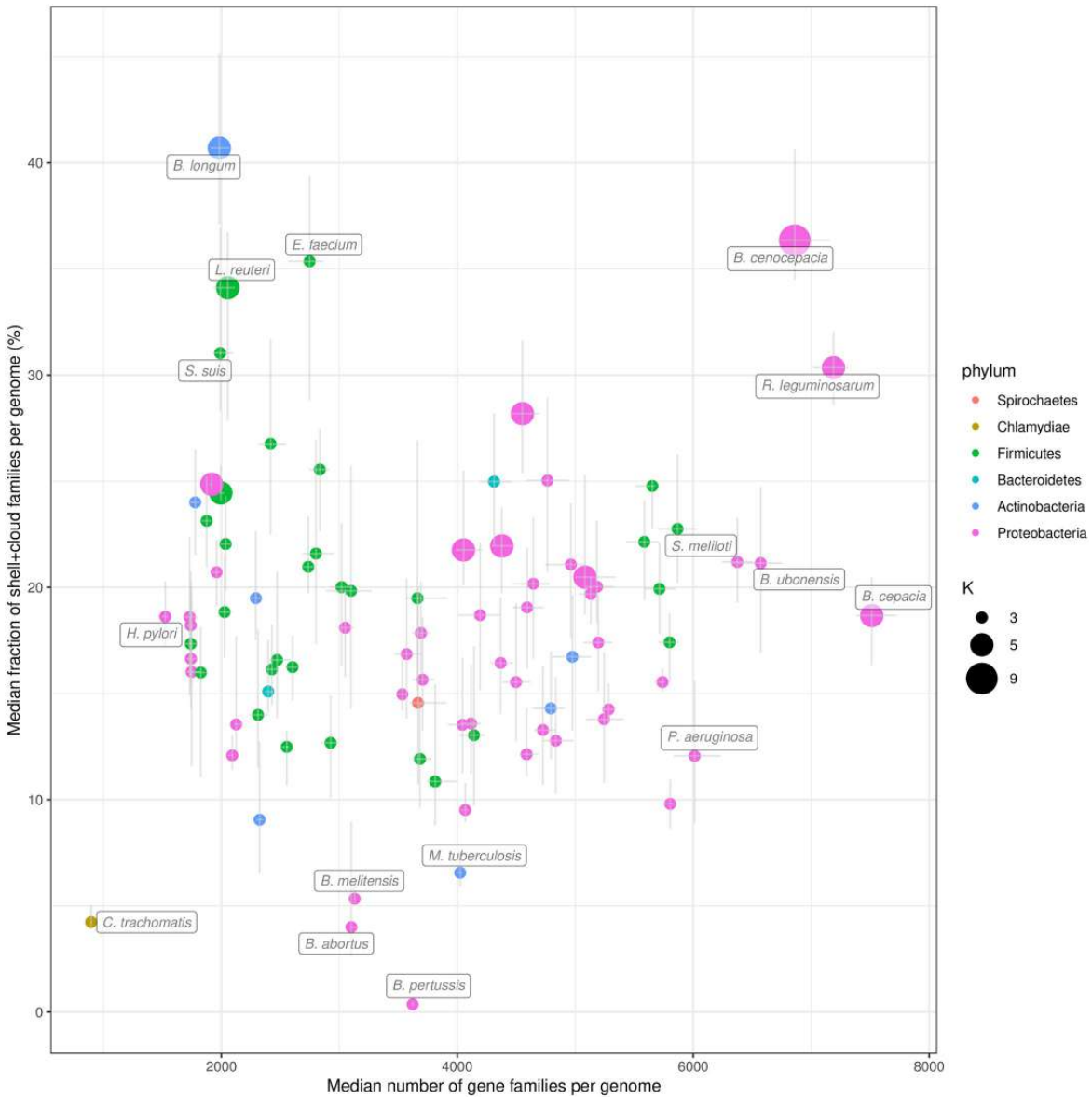


Fig 5. Fraction of the variable (shell + cloud) families per genome compared to the number of gene families. The results for the 88 most abundant species in GenBank are represented. The error bars show the interquartile ranges of the two variables. The points are colored by phylum and their size corresponds to the number of partitions (*K*) used.

https://doi.org/10.1371/journal.pcbi.1007732.g005

Spearman's $\rho > 0.75$). In addition, PPanGGOLiN predicts a number of partitions (K) for these species often greater than 3. Hence, their shell is more heterogeneous between subclades and becomes structured in several partitions whereas for species with a single shell partition the shell is less structured, possibly indicating many gene exchanges between strains from different

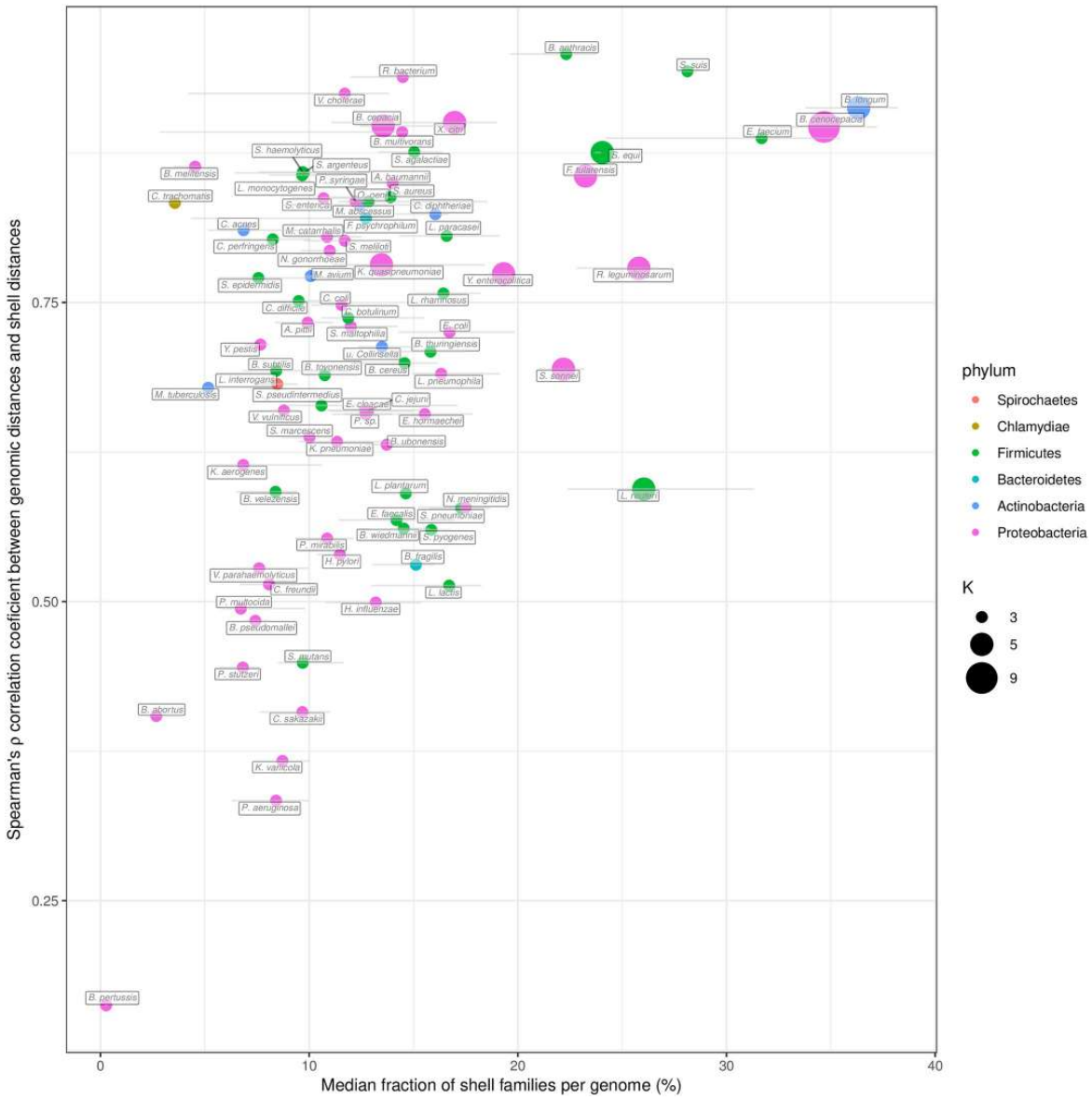


Fig 6. Spearman's ρ correlation coefficients between the shell genome presence/absence patterns and the MASH genomic distances compared with the shell fraction per genome. The results for the 88 most abundant species in GenBank are represented. The error bars show the interquartile ranges of the shell fraction. The points are colored by phylum and their size corresponds to the number of partitions (K) used.

https://doi.org/10.1371/journal.pcbi.1007732.g006

lineages. Among the nine species with a large shell genome (excluding *B. anthracis* due to taxonomic assignation errors), only two of them (*Shigella sonnei* and *Lactobacillus reuteri*) showed a relatively low correlation of their shell structure with the phylogeny (Fig 6). For *S. sonnei*, this could be explained by a high number of gene losses in the shell of this species that result

from convergent gene loss mediated by insertion sequences (preprint: [39]). For *L. reuteri*, these bacteria colonize the gastrointestinal tract of a wide variety of vertebrate species and have diversified into distinct phylogenetic clades that reflect the host where the strains were isolated, but not their geographical provenance [40]. As illustrated in S7 Fig, the shell of *L. reuteri* shows patterns of P/A that are only partially explained by the species phylogeny. Indeed, we observed clusters of families present across strains from distinct lineages that could contain factors for adaptation to the same host. In contrast, the shell structure of *B. longum* strongly depends on phylogenetic distances showing a clear delineation of adult and infant strains that have specialized into two subspecies (see S8 Fig).

We would like to stress the importance of the shell in the study of the evolutionary dynamics of bacteria. The shell content reflects the adaptive capacities of species through the acquisition of new genes that are maintained in the population. We found that the proportion of shell genes does not increase with the genome size. Instead, the shell accounts for a large fraction of the genomes of species when it is structured in several partitions. We can assume that those species are made of non-homogeneous subclades harboring specific shell genes which contribute to the specialization of the latter. Finally, it could be of interest to associate phenotypes to patterns of shell families that co-occur in different lineages independently of the phylogeny.

Analysis of Metagenome-Assembled Genomes in comparison with isolate genomes

The graph approach should make our tool robust to gaps in genome data, making it a useful tool to analyze pangenomes obtained from MAGs. To test this hypothesis, we built the pangenomes of the Species-level Genome Bins (SGBs, clusters of MAGs that span a 5% genetic diversity and are assumed to belong to the same species) from the recent paper of Pasolli *et al.* [41]. This study agglomerated and consistently built 4 930 SGBs (154 723 MAGs) from 13 studies focussed on the composition of the human microbiome. We skipped the quality control step (already performed by the authors), and computed the pangenomes following the procedure we used for the GenBank species. The only parameter which differs is the *K* value which is set to 3 as the detection of several shell partitions is difficult for MAGs because of their incompleteness. To make the comparison with GenBank species, SGBs were grouped according to their estimated species taxonomy (provided by the supplementary table S4 of [41]). In this table, we noticed potential errors in the taxonomic assignation of two species (*Blautia obeum* and *Chlamydia trachomatis* corresponding to SGBs 4844 and 6877, respectively) and thus excluded them from the analysis. Keeping the same constraint as previously, only species with at least 15 genomes in both MAGs and GenBank were used for the comparison. A total of just 78 species (corresponding to 151 SGBs) could be analyzed as a lot of microbiome species are laborious to cultivate and thus less represented in databanks (see S2 File). Then, we compared the MAG pangenome partitions predicted by PPanGGOLiN with those obtained with GenBank genomes. To perform this, we aligned MAG and GenBank families for each species and computed the percentage of common families for each partition (see Materials and methods for details and S2 File for detailed results).

We observed that the size of the estimated persistent genome of MAGs is similar to the one of GenBank genomes for most species (Fig 7). In 55 out of the 78 species, the absolute fold change of persistent size is less than 1.2 and 90% (SD = 5%) of its content is common between MAGs and GenBank genomes. The 23 other species with more important differences showed smaller persistent genomes with only 60% (SD = 15%) of the persistent genome of GenBank being found in MAGs. For these species, the PPanGGOLiN method missed a fraction of the persistent genome due to the incompleteness of MAGs. Indeed, in such cases, the

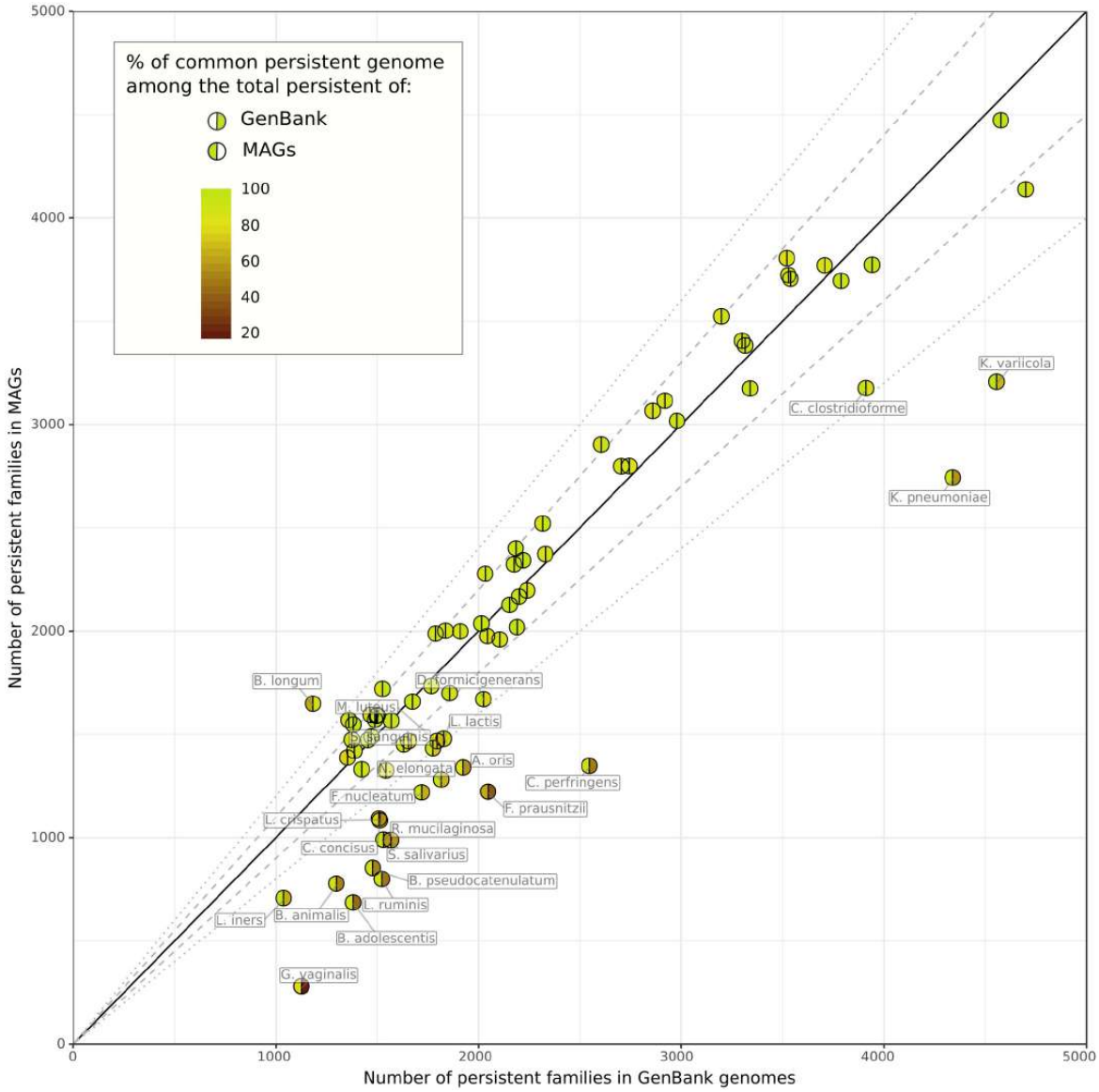


Fig 7. Illustration of the persistent genome overlaps between GenBank genomes and MAGs. Results for 78 species are represented. The colors of the hemispheres provide the percentage of common persistent gene families among the total persistent of MAGs (left hemisphere) or GenBank genomes (right hemisphere). The solid, dashed and dotted lines indicate the identity, a fold change of 1.1 and a fold change of 1.2 between the persistent genome sizes.

https://doi.org/10.1371/journal.pcbi.1007732.g007

missing gene families are mostly classified in the shell of the MAGs which contains 32% (SD = 11%) of the GenBank persistent families. Nevertheless, 89% (SD = 9%) of the MAG persistent families match the GenBank ones, meaning that PPanGGOLiN correctly assigned persistent families for MAGs even if the persistent genome of these 23 species is incomplete.

However, two species, *Bifidobacterium longum* and *Faecalibacterium prausnitzii*, have less than 75% of their MAG persistent families in common with GenBank ones. For *B. longum*, this could be explained by the fact that the MAGs were obtained mostly from human adult samples while this species in databanks are from a broader host range (infants and pigs). It means that the MAG persistent might contain additional genes related to host-specificity. As a matter of fact, 412 gene families from the MAG persistent (25% of the total MAG persistent) are found in the GenBank shell which supports our hypothesis. For *F. prausnitzii*, the differences might be explained by a poor estimation of the persistent using GenBank data due to the low number of considered genomes (17 genomes versus 4232 MAGs). As expected, the soft core (based on the usual threshold of 95% presence) is unrealistically low in the MAG species with only ≈98 gene families on average and only 4 species out of 78 having more than 500 families classified in the soft core (see S2 File). Hence, the soft core approach is not well adapted to the analysis of MAGs. Furthermore, using lower thresholds of presence is not adequate because defining a unique threshold for all the families misses the heterogeneity of gene family presence in MAGs.

To explore the diversity within the pangenome of each species, we compared the shell of GenBank genomes and MAGs for the 55 ones with similar persistent genomes. Interestingly, we observed for all the 55 species only a partial overlap between the MAGs and GenBank shells (see S9 Fig). Indeed, as the MAGs are obtained only from a specific environment (i.e. the human microbiome), the diversity of GenBank is not fully captured by MAGs. It is especially the case for most of the Firmicutes and Proteobacteria. Conversely, most of the MAGs of Bacteroidetes phylum cover more than half of GenBank diversity while containing a large fraction of shell genes that are lacking in the shell of isolate genomes (i.e. less than 45% of the families are represented in the shell of GenBank). As already reported by Pasolli *et al.* [41], this confirms that the MAGs considerably improve the estimate of the genetic diversity of Bacteroidetes which are key players in the gut microbiome.

In summary, we have shown that PPanGGOLiN is able to provide an estimation of the persistent genome even using MAGs, which may miss significant numbers of genes and be contaminated by fragments from other genomes. This is especially the case for the accessory genome because its assembly coverage and nucleotide composition generally differ from those of the persistent genome making the binning of these regions more difficult. Nevertheless, PPanGGOLiN is able to find shell gene families in MAGs bringing new genes that may be important for species adaptation in the microbiome. Hence, it enables further analyses, even for uncultured species lacking reference genomes, such as the reconstruction of the core metabolism from the persistent genome to predict culture media or the study of the landscape of horizontally transferred genes within species.

Conclusion

We have presented here the PPanGGOLiN method that enables the partitioning of pangenomes in persistent, shell and cloud genomes using a gene family graph approach. This compact structure is useful to depict the overall genomic diversity of thousands of strains highlighting variable paths made of shell and cloud genes within the persistent backbone. The statistical model behind PPanGGOLiN makes a more robust estimation of the persistent genome in comparison to classical approaches based on gene family frequencies in isolate genomes and also in MAGs. The definition of shell partitions based on statistical criteria allowed us to understand genome dynamics within species. We observed different patterns of shell with regard to phylogeny that may suggest different adaptive paths for the diversification of the species. It should be stressed that genome sampling is one of the main limitations of pangenome

studies and can therefore influence PPanGGOLiN partitioning especially for the shell genome. An improvement in the method could be to normalize the data to remove sampling bias. But as suggested by Brockhurst *et al*. [42], this issue should first be examined from a biological perspective by collecting and analyzing genomes from ecologically coherent microbial populations or ecotypes.

Future applications of PPanGGOLiN could include the prediction of genomics islands within the shell and cloud genomes. A first version of this application (Bazin *et al*., in preparation) is already integrated in the MicroScope genome analysis platform [43]. Next, it would be interesting to determine the architecture of these variable regions by predicting conserved gene modules using information on the occurrence of families and their genomic neighborhood in the pangenome graph. Regarding metagenomics, pangenome graphs of PPanGGOLiN could be used as a reference (i.e. instead of individual genomes) for species quantification by mapping short or long reads on the graph to compute the coverage of the persistent genome. Indeed, each gene families of the partitioned pangenome graph could embed a variation graph as an alignment template [19]. Moreover, coverage variation in the shell or cloud genomes could allow the detection of strain-specific paths in the graph that are signatures of distinctive traits within microbiotes.

To conclude, the graph-based approach proposed by PPanGGOLiN provides an effective basis for very large scale comparative genomics and we hope that drawing genomes on rails like a subway map may help biologists navigate the great diversity of microbial life.

Materials and methods

To explain the partitioning of pangenomes, we first need to describe the method based on the P/A matrix only (BinEM) and then the method built upon it that uses the pangenome graph to improve the partitioning (NEM).

Modeling the P/A matrix via a multivariate Bernoulli Mixture Model

PPanGGOLiN aims to classify patterns of P/A of gene families into K partitions ($K \in \mathbb{N}$; $K \geqslant 3$). Input data consists of a binary matrix X in which a x_{ij} entry is 1 if family i is present in a genome j and 0 otherwise (Fig 1) where $1 \leqslant i \leqslant F$ in each of the F gene families and $1 \leqslant j \leqslant N$ in each of the N genomes. A first approach for partitioning the data relies on a multivariate Bernoulli Mixture Model (BMM) estimated through the Expectation-Maximization (EM) algorithm [44] (named the BinEM method). The number of partitions K may be greater than 3 (persistent, shell and cloud) due to the possible presence of antagonist P/A patterns among the different strains of a species. Therefore, two of the partitions will correspond to the persistent and cloud genome and a number of $K - 2$ partitions will correspond to the shell genome. The value of K can be either provided by the user or determined automatically (see next section).

In the BMM, the matrix comprises data vectors $X_i = (x_{ij})_{1 \leqslant j \leqslant N}$ describing P/A of families, which are assumed to be independent and identically distributed with a mixture distribution given by:

$$P(X_i = (x_{ij})_{1 \leqslant j \leqslant N}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{N} \epsilon_{kj}^{|x_{ij} - \mu_{kj}|} (1 - \epsilon_{kj})^{1 - |x_{ij} - \mu_{kj}|}$$

where $\pi = (\pi_1, \ldots, \pi_k, \ldots, \pi_K)$ denotes the mixing proportions satisfying $\pi_k \in [0, 1]$; $(\sum_{k=1}^{K} \pi_k) = 1$ and where π_k is the unknown proportion of gene families belonging to the k^{th} partition. Moreover, $\mu_k = (\mu_{kj})_{1 \leqslant j \leqslant N} \in \{0; 1\}^N$ are the centroid vectors of P/A of the k^{th} partition representing the most probable binary states and $\epsilon_k = (\epsilon_{kj})_{1 \leqslant j \leqslant N} \in \left[0, \frac{1}{2}\right]^N$ are the unknown

vectors of dispersion around μ_k. The default values of the dispersion vector ϵ_k associated to each centroid vector μ_k are constrained to be identical for all the ϵ_{kj} of a specific k partition (for all the genomes of a specific partition) in order to avoid over-fitting but it is possible to release this constraint. The parameters of this model, as well as corresponding partitions, are estimated by the EM algorithm. To speed up the computation of the EM algorithm, a heuristic is used to initialize the BMM parameters in order to converge to a relevant partitioning using fewer EM-steps. This heuristic consists in setting π_k with equiprobable proportions equal to $1/K$ while the ϵ_{kj} and μ_{kj} parameters are initialized triangularly.

Given $s = 1/\lceil K/2 \rceil$, the triangular initialization consists of:

$$\{\mu_{kj}\}_{1 \leqslant k \leqslant K/2, 1 \leqslant j \leqslant N} = 1$$
$$\{\mu_{kj}\}_{K/2 < k \leqslant K, 1 \leqslant j \leqslant N} = 0$$
$$\{\epsilon_{kj}\}_{1 \leqslant k \leqslant K/2, 1 \leqslant j \leqslant N} = s \cdot k$$
$$\{\epsilon_{kj}\}_{K/2 < k \leqslant K, 1 \leqslant j \leqslant N} = s \cdot (K - k + 1)$$

An interesting consequence of this initialization is that the persistent genome will be the first partition ($k = 1$) while the cloud genome will correspond to the last partition ($k = K$). This particular initialization solves the classical label switching problem in our context.

Partitioning of the P/A matrix

To perform the partitioning of the P/A matrix, each gene family i must be allocated to a single partition. The variables $\{Z_i\}_{1 \leqslant i \leqslant F}$ with a state space $\{1, \ldots, K\}$ indicate the partition to which each gene family i belongs. Therefore, once the NEM parameters are optimized, the method automatically assigns the gene families to their most probable partition z_i according to the model if their estimated posterior probability is above 0.5. If no partition can be assigned in this way, then the gene family is assigned to the shell (partition with intermediate frequency).

Selection of the optimal number of partitions (K)

To determine the optimal K, named $\hat{K}$, the algorithm runs multiple partitionings with increasing values of K. After a few steps of the EM algorithm (10 steps by default), the Integrated Completed Likelihood (*ICL*) [45] is computed for each K. The *ICL* corresponds to the Bayesian Information Criterion (*BIC*) [46] penalized by the estimated mean entropy and is calculated as:

$$ICL(K) = BIC(K) - \sum_{k=1}^{K}\sum_{i=1}^{F} p(z_i \mid X, \hat{\theta}, k) \log\left(p(z_i \mid X, \hat{\theta}, k)\right); \forall p(z_i \mid X, \hat{\theta}, k) > 0$$

and

$$BIC(K) = \log \mathbb{P}_K(X \mid \hat{\theta}) - 1/2 dim(K) \log F$$

where $\log \mathbb{P}_K(X \mid \theta)$ is the data log-likelihood under a multivariate BMM with K partitions and $\theta = (\{\pi_k\}_{1 \leqslant k \leqslant K}, \{\mu_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N}, \{\epsilon_{kj}\}_{1 \leqslant k \leqslant K, 1 \leqslant j \leqslant N})$. This log-likelihood can be calculated as follows:

$$\log \mathbb{P}_K(X \mid \theta) = \sum_{i=1}^{F} \log\left(\sum_{k=1}^{K} \pi_q \prod_{j=1}^{N} \epsilon_{kj}^{|x_{ij} - \mu_{kj}|} (1 - \epsilon_{kj})^{1 - |x_{ij} - \mu_{kj}|}\right)$$

Moreover, $\hat{\theta}$ is the maximum likelihood estimator (approximated through the BinEM

algorithm) and $dim(K)$ is the dimension of the parameter space for this model. Here, $dim(K) = K(N + 2)$ if the dispersion vector ϵ_k associated to each centroid vector μ_k is constrained to be identical for all the ϵ_{kj} of a specific k partition and $dim(K) = K(2N + 1)$ if the dispersion vector ϵ_k is free. Relying on this criterion, the best number of partitions is selected as $\hat{K} = \arg\min_K ((1 - \delta_{ICL})ICL(K))$ where δ_{ICL} is a sufficiently small margin to avoid choosing a too high K value that would provide no significant gain compared to a lower value of K (by default $\delta_{ICL} = 0.05 \times (max(ICL) - min(ICL))$).

Generation of the pangenome graph

PPanGGOLiN uses a graph-based representation to store and visualize pangenomes. In this graph, the nodes correspond to gene families and the edges to genetic contiguity (i.e. genes that are direct neighbors in a genome). Two nodes are connected if the corresponding gene families contain at least one pair of genes that are adjacent in a genome. Edges are labeled with the corresponding genome identifiers and weighted by the proportion of genomes sharing that link. This process results in a pangenome graph (see Fig 2 as an example).

Formally, a pangenome graph $G = (V, E)$ is a graph having a set of vertices $V = \{(v_i)_{(1 \leqslant i \leqslant F)}\}$ where F is the number of gene families in the pangenome associated with a set of edges $E = \{e_{i \sim i'}\} = \{(v_i, v_{i'})\}$, $v_i \in V$, $v_{i'} \in V$ where the couple of vertices $(v_i, v_{i'})$ are gene families having their genes $(v_{i,j}, v_{i',j})$ adjacent on the genome j and where the function $countNeighboringGenes$ $(v_i, v_{i'})$ counts the adjacency occurrences in the N genomes. Each edge $\{e_{i \sim i'}\}$ has a weight $w_{i \sim i'}$ where $w_{i \sim i'} = \frac{1}{N} \sum_{j=1}^{N} countNeighboringGenes(v_{i,j}, v_{i',j})$.

Partitioning via Neighboring Expectation-Maximization

From the graph previously described, the neighborhood information of the gene families is used to improve the partitioning results. Indeed, the BinEM approach described above is extended by combining the P/A matrix X with the pangenome graph G. This relies on a hidden Markov Random Field (MRF) model whose graph structure is given by G. In this model, each node belongs to some unobserved (hidden) partitions which are distributed among gene families according to a MRF which favors two neighbors to be more likely classified in the same partition. Conditional on this hidden structure, the binary vectors of P/A are independent and follow a multivariate Bernoulli distribution with proportion vectors depending on the associated partition. This approach is called NEM, as it relies on the Neighboring Expectation-Maximization algorithm [47–49]. As such, NEM tends to smooth the partitioning by grouping gene families that have a weighted majority of neighbors belonging to the same partition. The previously introduced latent variables $\{Z_i\}_{1 \leqslant i \leqslant F}$, that indicate the partition to which each gene family belongs are now distributed according to a MRF. More precisely, they have the following Gibbs distribution:

$$\mathbb{P}(\{Z_i\}_{1 \leqslant i \leqslant F}) = W_\beta^{-1} \exp\left(\sum_{i=1}^{F} \sum_{k=1}^{K} \pi_k 1_{Z_i = k} + \beta \frac{F}{\sum_{i \sim i'} w_{i \sim i'}} \sum_{i \sim i'} w_{i \sim i'} 1_{Z_i = Z_{i'}}\right)$$

where 1_A is the indicator function of event A and the second sum concerns every pair $(i \sim i')$ of neighbor gene families. The parameter $\beta \geq 0$ corresponds to the coefficient of spatial regularity. The $\frac{F}{\sum_{i \sim i'}^{E} w_{i \sim i'}}$ is a corrector term ensuring that the strength of the spatial smoothing is balanced regardless of the number of gene families. Indeed, when the number of genomes (N) increases, the number of gene families (F) tends to be higher than the sum of the edge weights.

Finally,

$$W_\beta = \sum_{\{\tilde{z}_i\}\in\{1\dots K\}^F} \exp\Big(\sum_{i=1}^{F}\sum_{k=1}^{K}\pi_k 1_{\tilde{z}_i=k} + \beta\frac{F}{\sum_{i\sim i'} w_{i\sim i'}}\sum_{i\sim i'} w_{i\sim i'} 1_{\tilde{z}_i=\tilde{z}_{i'}}\Big)$$

is a normalizing constant. Note that W_β cannot be computed, due to a large number of possible configurations. The degree of dependence between elements is controlled by the parameter β. Neighboring elements will be more inclined to belong to the same group with a higher value of this parameter. Here, the data vectors $(X_i)_{1\leqslant i\leqslant F}$ are not independent anymore. However, conditional on the latent groups $(Z_i)_{1\leqslant i\leqslant F}$, they are independent and follow the multivariate Bernoulli distribution:

$$\mathbb{P}(\{X_i\}_{1\leqslant i\leqslant F}|\{Z_i\}_{1\leqslant i\leqslant F}) = \prod_{i=1}^{F}\prod_{j=1}^{N}\epsilon_{Z_i j}^{|x_{ij}-\mu_{Z_i j}|}(1-\epsilon_{Z_i j})^{1-|x_{ij}-\mu_{Z_i j}|}.$$

Many different techniques may be used to approximate the maximum likelihood estimator in the hidden MRF. NEM relies on a mean-field approximation for the distribution of the latent random variables $Z_{i1\leqslant i\leqslant F}$ conditional on the observations. It should be noted that the optimal number of partitions (K) is not determined automatically using NEM and is therefore first estimated using the BinEM approach.

Issues resulting from high-dimensional statistics and parallelization

As plenty of statistical approaches, NEM is not adapted to high dimensional settings (i.e. whenever the condition $F >> N$ is not satisfied). This can occur in pangenomics as the discovery rate of new families in the pangenome slightly decreases when new genomes are added. Mathematical solutions to this problem seem to exist [50–52] for example via the weighting of genomes (based on their respective contribution to the pangenome diversity) or via sparse partitioning methods. An improvement of NEM should include these solutions and could be a perspective of this work.

Pangenome software must be designed to scale up to thousands of genomes. NEM scales quadratically with the number of genomes and is hard to parallelize. Thus, it leads to intensive computations when thousands of genomes are included in the analysis.

Our solution to the mentioned issues is to sample the genomes in chunks and to perform multiple partitioning in parallel. Each family must be involved in at least $N_{total}/N_{samples}$ samplings and will be partitioned only if it is classified in the same partition in at least 50% of the samplings where it is present (absolute majority). If some families do not respect this condition, we continue sampling until all gene families have been partitioned. Chunks have to be large enough to be representative, therefore a size of at least 500 genomes is advised.

Analysis of isolate genomes and Metagenome-Assembled Genomes

To obtain the set of isolate genomes to be analyzed, we downloaded all archaeal and bacterial genomes (220 561 genomes) of the GenBank database at the date of the 17th of April 2019. We removed genome assemblies that do not respect quality control criteria defined by GenBank. They correspond to entries with an assembly status flag different from "status = latest" in the "assembly_status.txt" files. In addition, genomes were discarded if they had more than 1000 contigs or a L90 > 100. These filters allowed us to exclude poor quality assemblies, some of which may correspond to contaminated genomes and others to incomplete ones. For each species (identified by its NCBI species taxid), a pairwise genomic distance matrix was computed using Mash (version 2.0) [37]. To avoid redundancy, if several genomes are at a Mash

distance < 0.0001, only one was kept (the one having the lowest number of contigs). A single linkage clustering using SiLiX (version 1.2.11) [53] was then performed on the adjacency graph of the Mash distance matrix considering only distances below or equal to 0.06. This Mash distance corresponds to a 94% Average Nucleotide Identity (ANI) cutoff which is a usual value to define species [54]. Genomes that were not in the largest connected component were discarded to remove potential taxonomic assignation errors. Only species having at least 15 remaining genomes were then considered for the analysis. The list of all the GenBank assembly accessions used after filtering is available in S3 File. This dataset consists of 439 species encompassing 136 287 genomes (see S1 File). MAGs from the Pasolli *et al*. study [41] were downloaded from https://opendata.lifebit.ai/table/SGB. In this dataset, the genomes are already grouped in Species Genome Bins. These SGBs do not exactly match the GenBank taxonomy. Thus, SGBs assigned with the same species name (column "estimated taxonomy" in the supplementary table S4 of [41]) were merged to allow comparison with GenBank. SGBs that do not have a taxonomy assigned at the species level were not considered. A total of 583 species encompassing 698 SGBs and 71 766 MAGs were analyzed but only MAGs from 78 species were finally compared to GenBank genomes. To avoid introducing a bias in our analysis due to heterogeneous gene calling, GenBank annotations were not considered as they were obtained using a variety of annotation workflows. Genomes from GenBank and Pasolli *et al* were consistently annotated using the procedure implemented in PPanGGOLiN. Prodigal (version 2.6.2) [55] is used to detect the coding genes (CDS). tRNA and tmRNA genes are predicted using Aragorn (version 1.2.38) [56] whereas the rRNA are detected using Infernal (version 1.1.2) [57] with HMM models from Rfam [58]. In the case of overlaps between a RNA and a CDS, the overlapping CDS are discarded. Homologous gene families were determined using MMseqs2 (version 8-fac81) [59] with the following parameters: coverage = 80% with cov-mode = 0, minimal amino acid sequence identity = 80% and cluster-mode = 0 corresponding to the Greedy Set Cover clustering mode. PPanGGOLiN partitioning was executed on each species using the NEM approach with a parameter $\beta = 2.5$. The nodes having a degree above 10 (which is the default parameter) were not considered to smooth the partitioning via the MRF. The number of partitions (K) was determined automatically for each NCBI species using a $\delta_{ICL} = 0.05$ and iterating between 3 and 20 for the possible values of K. K was fixed at 3 for the MAG analysis. The partitioning was done using chunks of 500 genomes when there were more than 500 genomes in a species. To compare PPanGGOLiN results between MAGs and GenBank genomes for each species, the representative sequences of each MAG gene family (extracted using the mmseqs2 subcommand: "result2repseq") were aligned (using mmseqs2 "search") on those of GenBank genomes. If the best hit of the query had a sequence identity $> 80\%$ and a coverage $> 80\%$ of the target, the 2 corresponding gene families of each dataset were associated.

Rarefaction curves

To represent the pangenome evolution according to the number of sequenced genomes, a multiple resampling approach was used. For each species with at least 100 genomes, 8 rarefaction curves showing the evolution of the pangenome and the persistent, shell, cloud, soft core, soft accessory, exact core and exact accessory components were computed for sample sizes of 1 to 100 genomes randomly drawn from the set of all genomes of the species. Each sample size was analyzed using 30 different samples. For each sample, the number of partitions K is automatically determined between 3 and the K obtained on all the genomes of the species. A non-linear Least Squares Regression was performed to fit the rarefaction curves with Heaps' law $F = \kappa N^{\gamma}$ where F is the number of gene families, N the number of genomes, γ the tendency of

the evolution and κ a proportional factor [35]. Subset sizes ≤ 15 were not used for the fitting as they are sometimes too variable to ensure a good fitting. The function "scipy.optimize.curve_fit" of the Python scipy package (version 1.0.0), based on the Levenberg-Marquardt algorithm, was used to fit the rarefaction curves. For each subset size, the median and quartiles were calculated to obtain a ribbon of interquartile ranges (IQR) along the rarefaction curves. We call the area of this ribbon the IQR area (see S2 Fig as an example).

PPanGGOLiN software implementation

PPanGGOLiN was designed to be a software suite performing the annotation of the genomic sequences, building the gene families and the pangenome graph before partitioning it. Users can also provide their own annotations (GFF3 or GBFF format) and gene families. The application stores its data in a compressed HDF5 file but can also return the graph in GEXF or JSON formats and the P/A matrix with the partitioning in CSV or Rtab files (similarly to the ones provided by Roary [34]). It also generates several illustrative figures, some of which are presented in the article. PPanGGOLiN was developed in the Python 3 and C languages and is intended to be easily installable on Linux and Mac OS systems via a BioConda package [60] (see https://bioconda.github.io/recipes/ppanggolin/README.html). The code is also freely available on the GitHub website at the following address: https://github.com/labgem/PPanGGOLiN.

Supporting information

S1 Fig. Density distributions of the gene family frequencies of each partition. Results for the 88 most abundant species in GenBank are represented in addition with a global distribution of the gene family frequencies from all the species. Density values of the cloud genome above 100 (y-axis) were trimmed for visualization purpose. The dashed yellow vertical bars indicate the threshold of frequency (⩾95%) used to delimit the soft core genome.
(PDF)

S2 Fig. Evolution of the persistent, shell, soft core and exact core metrics of Lactobacillus plantarum compared to the number of genomes. The rarefaction curves represent the evolution of the partition sizes as a function of an increasing number of genomes in random subsets of genomes. Plain lines connect the medians while colored areas represent the interquartile ranges. A regression curve (bold dashed line) is drawn fitting all the points of each partition by the Heaps' law ($F = \kappa N^{\gamma}$). The total area of the interquartile ranges (IQR) is indicated for each partition.
(TIF)

S3 Fig. Density distributions of the Heaps' law γ-tendencies. These γ-tendencies were obtained by fitting a Heaps' law on rarefaction curves between subset sizes of 15 to 100 genomes in the 88 most abundant species in GenBank. The exact core median and exact accessory are not shown.
(TIF)

S4 Fig. Shell γ-tendency compared to the total number of shell families normalized by the median number of gene families per genome in each species. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (K) used.
(TIF)

S5 Fig. Fraction of shell families per genome compared to the number of gene families. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (*K*) used.
(TIF)

S6 Fig. Fraction of cloud families per genome compared to the number of gene families. Results for the 88 most abundant species in GenBank are represented. The points are colored by phylum and their size corresponds to the number of partitions (*K*) used.
(TIF)

S7 Fig. Presence/Absence matrix of the shell genome of *L. reuteri* ordered by a Neighbor Joining tree based on the MASH distances. The leaves of the tree are colored by host or origin. This information was obtained from the metadata in GenBank files (host and isolation source qualifiers).
(TIF)

S8 Fig. Presence/Absence matrix of the shell genome of *B. longum* ordered by a Neighbor Joining tree based on the MASH distances. The leaves of the tree are colored by species clusters defined by the GTDB database (release R04-RS89), namely (*B. infantis* or *B. longum*). "NA" values correspond to genomes not available in GTDB.
(TIF)

S9 Fig. Illustration of the shell genome overlaps between MAGs or GenBank of 55 species. The x-axis represents the percentage of common shell of the GenBank shell while the y-axis corresponds to the percentage of common shell of the MAGs shell. Diamonds and squares represent MAGs and GenBank genomes, respectively. They are colored by phylum and their size indicates the number of genomes.
(TIF)

S1 File. Table compiling all the metrics obtained from the pangenomes of the 439 GenBank species. This is a CSV file.
(CSV)

S2 File. Table compiling all the metrics obtained from the comparison of PPanGGOLiN results between MAGs and GenBank genomes in 78 species. This is a CSV file.
(CSV)

S3 File. List of GenBank assembly accessions for the 439 studied species. This is a TSV file where each line corresponds to all the GenBank assembly accession used in this study for each 'species id' in the NCBI taxonomy.
(TSV)

Acknowledgments

We acknowledge Alexandre Renaux and Jonathan Mercier for their preliminary insights on pangenome graphs. We thank Mélanie Buy for drawing the PPanGGOLiN logo. Finally, we thank Guilhem Royer, Valentin Sabatet, Johan Rollin, Mohammed-Amin Madoui, Tom Delmont, Nicolas Pons and Pierre Peterlongo for all their advice along this work.

Author Contributions

Conceptualization: David Vallenet.

Data curation: Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin.

Formal analysis: Guillaume Gautreau, Adelme Bazin.

Investigation: Guillaume Gautreau, Adelme Bazin.

Methodology: Guillaume Gautreau, Catherine Matias, Christophe Ambroise.

Software: Guillaume Gautreau, Adelme Bazin.

Supervision: David Vallenet.

Visualization: Guillaume Gautreau.

Writing – original draft: Guillaume Gautreau, Adelme Bazin, Eduardo P. C. Rocha, David Vallenet.

Writing – review & editing: Guillaume Gautreau, Adelme Bazin, Mathieu Dubois, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P. C. Rocha, David Vallenet.

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA. 2005; 102(39):13950–13955. https://doi.org/10.1073/pnas.0506758102 PMID: 16172379
2. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005; 15(6):589–594. https://doi.org/10.1016/j.gde.2005.09.006 PMID: 16185861
3. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics. 2011; 7(1):1–12. https://doi.org/10.1371/journal.pgen.1001284
4. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 2010; 60(4):708–720. https://doi.org/10.1007/s00248-010-9717-3 PMID: 20623278
5. Acevedo-Rocha CG, Fang G, Schmidt M, Ussery DW, Danchin A. From essential to persistent genes: a functional approach to constructing synthetic life. Trends Genet. 2013; 29(5):273–279. https://doi.org/10.1016/j.tig.2012.11.001 PMID: 23219343
6. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 2013; 79(24):7696–7701. https://doi.org/10.1128/AEM.02411-13 PMID: 24096415
7. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends Genet. 2009; 25(3):107–110. https://doi.org/10.1016/j.tig.2008.12.004 PMID: 19168257
8. Bolotin E, Hershberg R. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. Front Microbiol. 2017; 8:1536. https://doi.org/10.3389/fmicb.2017.01536 PMID: 28890711
9. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, Ussery DW. On the Origins of a Vibrio Species. Microbial Ecology. 2010; 59(1):1–13. https://doi.org/10.1007/s00248-009-9596-7 PMID: 19830476
10. Periwal V, Patowary A, Vellarikkal SK, Gupta A, Singh M, Mittal A, et al. Comparative whole-genome analysis of clinical isolates reveals characteristic architecture of Mycobacterium tuberculosis pangenome. PLoS ONE. 2015; 10(4):e0122979. https://doi.org/10.1371/journal.pone.0122979 PMID: 25853708
11. Livingstone PG, Morphew RM, Whitworth DE. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. Front Microbiol. 2018; 9:3187. https://doi.org/10.3389/fmicb.2018.03187 PMID: 30619233
12. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008; 36(21):6688–6719. https://doi.org/10.1093/nar/gkn668 PMID: 18948295
13. Baumdicker F, Hess WR, Pfaffelhuber P. The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol. 2012; 4(4):443–456. https://doi.org/10.1093/gbe/evs016 PMID: 22357598

14. Collins RE, Higgs PG. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol Biol Evol. 2012; 29(11):3413–3425. https://doi.org/10.1093/molbev/mss163 PMID: 22752048

15. Lobkovsky AE, Wolf YI, Koonin EV. Gene frequency distributions reject a neutral model of genome evolution. Genome Biology and Evolution. 2013;. https://doi.org/10.1093/gbe/evt002 PMID: 23315380

16. Bolotin E, Hershberg R. Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. Genome Biol Evol. 2015; 7(8):2173–2187. https://doi.org/10.1093/gbe/evv135 PMID: 26163675

17. Moldovan MA, Gelfand MS. Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. Frontiers in Microbiology. 2018; 9:428. https://doi.org/10.3389/fmicb.2018.00428 PMID: 29593678

18. Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y, Huang XZ, et al. A novel method of consensus pan- chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. Genome Biol. 2015; 16:143. https://doi.org/10.1186/s13059-015-0701-6 PMID: 26195261

19. Garrison E, Siren J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018; 36(9):875–879. https://doi.org/10.1038/nbt.4227 PMID: 30125266

20. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019; 37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807

21. Rakocevic G, Semenyuk V, Lee WP, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet. 2019; 51(2):354–362. https://doi.org/10.1038/s41588-018-0316-4 PMID: 30643257

22. Consortium TCPG. Computational pan-genomics: status, promises and challenges. Brief Bioinformatics. 2016.

23. Zekic T, Holley G, Stoye J. Pan-Genome Storage and Analysis Techniques. Methods Mol Biol. 2018; 1704:29–53. https://doi.org/10.1007/978-1-4939-7463-4_2 PMID: 29277862

24. van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, Farmer CL, et al. Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol. 2014; 10(8): e1003788. https://doi.org/10.1371/journal.pcbi.1003788 PMID: 25144616

25. Gumiere T, Meyer K, Burns AR, Gumiere SJ, Bohannan BJM, Andreote FD. A probabilistic model to identify the core microbial community. bioRxiv. 2018.

26. Snipen L, Almøy T, Ussery DW. Microbial comparative pan-genomics using binomial mixture models. BMC Genomics. 2009; 10:385. https://doi.org/10.1186/1471-2164-10-385 PMID: 19691844

27. Snipen L, Liland KH micropan: an R-package for microbial pan-genomics BMC Bioinformatics. 2015; 16:79. https://doi.org/10.1186/s12859-015-0517-0 PMID: 25888166

28. Fang G, Rocha EP, Danchin A. Persistence drives gene clustering in bacterial genomes. BMC Genomics. 2008; 9(1):4. https://doi.org/10.1186/1471-2164-9-4 PMID: 18179692

29. Oliveira PH, Touchon M, Cury J, Rocha EPC. The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. 2017; 8(1):841. https://doi.org/10.1038/s41467-017-00808-w PMID: 29018197

30. Singh JK, Adams FG, Brown MH. Diversity and Function of Capsular Polysaccharide in Acinetobacter baumannii. Front Microbiol. 2018; 9:3301. https://doi.org/10.3389/fmicb.2018.03301 PMID: 30687280

31. Hu D, Liu B, Dijkshoorn L, Wang L, Reeves PR. Diversity in the major polysaccharide antigen of Acinetobacter baumannii assessed by DNA sequencing, and development of a molecular serotyping scheme. PLoS ONE. 2013; 8(7):e70329. https://doi.org/10.1371/journal.pone.0070329 PMID: 23922982

32. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks; 2009. Available from: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

33. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE. 2014; 9(6):1–12. https://doi.org/10.1371/journal.pone.0098679

34. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015; 31(22):3691–3693. https://doi.org/10.1093/bioinformatics/btv421 PMID: 26198102

35. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008; 11(5):472–477. https://doi.org/10.1016/j.mib.2008.09.006 PMID: 19086349

36. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018; 36(10):996–1004. https://doi.org/10.1038/nbt.4229 PMID: 30148503

37. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016; 17(1):132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

38. Criscuolo A. A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies. Research Ideas and Outcomes. 2019; 5:e36178. https://doi.org/10.3897/rio.5.e36178

39. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of Shigella species. bioRxiv. 2019.

40. Oh PL, Benson AK, Peterson DA, Patil PB, Moriyama EN, Roos S, et al. Diversification of the gut symbiont Lactobacillus reuteri as a result of host-driven evolution. ISME J. 2010; 4(3):377–387. https://doi.org/10.1038/ismej.2009.123 PMID: 19924154

41. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell. 2019; 176(3):649–662. https://doi.org/10.1016/j.cell.2019.01.001 PMID: 30661755

42. Brockhurst MA, Harrison E, James PJ, Richards T, McNally A, MacLean C The Ecology and Evolution of Pangenomes Current Biology. 2019; 29(20):1094–1103.

43. Vallenet D, Calteau A, Dubois M, Amours P, Bazin A, Beuvin M, et al. MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. Nucleic Acids Research. 2019. https://doi.org/10.1093/nar/gkz926

44. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B. 1977; 39(1):1–38.

45. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000; 22(7):719–725. https://doi.org/10.1109/34.865189

46. Schwarz G. Estimating the Dimension of a Model. Ann Statist. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136

47. Ambroise C, Dang M, Govaert G. Clustering of Spatial Data by the EM Algorithm. In: Soares A, Gómez-Hernandez J, Froidevaux R, editors. geoENV I—Geostatistics for Environmental Applications. Dordrecht: Springer Netherlands; 1997. p. 493–504.

48. Ambroise C, Govaert G. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters. 1998; 19(10):919–927. https://doi.org/10.1016/S0167-8655(98)00076-2

49. Dang M, Govaert G. Spatial Fuzzy Clustering using EM and Markov Random Fields. In: International-Journal of System Research and Information Science; 1998. p. 183–202.

50. Bouguila N. On multivariate binary data clustering and feature weighting. Computational Statistics and Data Analysis. 2010; 54(1):120–134. https://doi.org/10.1016/j.csda.2009.07.013

51. Yamamoto M, Hayashi K. Clustering of multivariate binary data with dimension reduction via L1-regularized likelihood maximization. Pattern Recognition. 2015; 48(12):3959–3968. https://doi.org/10.1016/j.patcog.2015.05.026

52. Śmieja M, Hajto K, Tabor J. Efficient mixture model for clustering of sparse high dimensional binary data. Data Mining and Knowledge Discovery. 2019.

53. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011; 12:116. https://doi.org/10.1186/1471-2105-12-116 PMID: 21513511

54. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA. 2005; 102(7):2567–2572. https://doi.org/10.1073/pnas.0409727102 PMID: 15701695

55. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11:119. https://doi.org/10.1186/1471-2105-11-119 PMID: 20211023

56. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004; 32(1):11–16. https://doi.org/10.1093/nar/gkh152 PMID: 14704338

57. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29(22):2933–2935. https://doi.org/10.1093/bioinformatics/btt509 PMID: 24008419

58. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018; 46(D1):D335–D342. https://doi.org/10.1093/nar/gkx1038 PMID: 29112718

59. Steinegger M, Soeding J. Sensitive protein sequence searching for the analysis of massive data sets. Nat Biotech. 2017. https://doi.org/10.1038/nbt.3988

60. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature methods. 2018; 15(7):475–476. https://doi.org/10.1038/s41592-018-0046-7 PMID: 29967506

PhD defense keywords

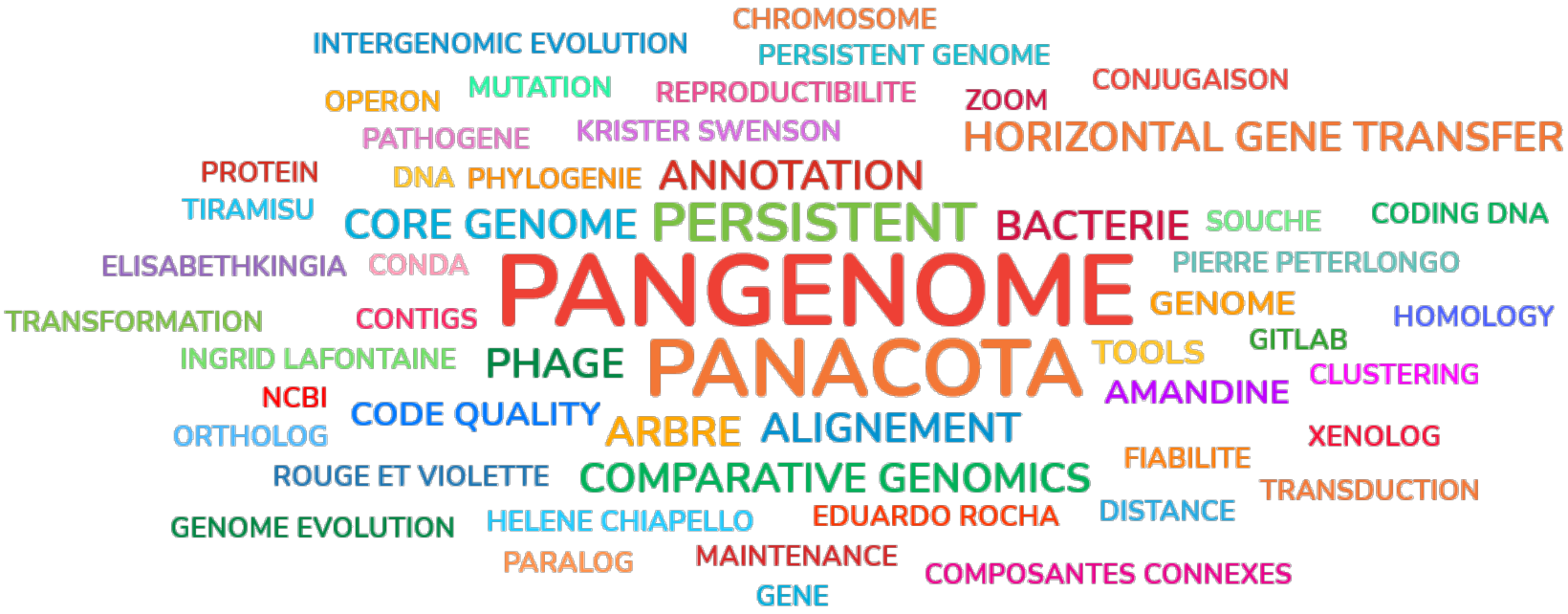


Figure 8.2: During my PhD defense, I asked the audience to write words they consider as important, noteworthy and/or simply "nice". Here is the result. The size of the word is proportional to the number of times it has been written.

Abbreviations

aa amino-acid. 25

ANI Average Nucleotide Identity. 66

ARG Antibiotic Resistance Gene. 46

CDS Coding DNA Sequence. 25

CI Continuous Integration. 193

CP Conjugative Plasmid. 34

DDBJ DNA Data Bank of Japan. 57

DDH DNA-DNA hybridization. 65

DNA Deoxyribonucleic acid. (dsDNA: double stranded; ssDNA: single stranded). 21, 23, 249

DNase Deoxyribonuclease. 46

DT Doubling Time. 28

ENA European Nucleotide Archive. 57

GWAS Genome-Wide Association Studies. 193

HGT Horizontal Gene Transfer. 33

HMM Hidden Markov Model. 57

HR Homologous Recombination. 48

ICE Integrative and Conjugative Element. 34, 42

ICNP International Code of Nomenclature of Prokaryotes. 30

ICTV International Committee on Taxonomy of Viruses. 36

INSDC International Nucleotide Sequence Database Collaboration. 57

IR Inverted Repeat. 39

IS Insertion Sequence. 39

LPS Lipopolysaccharides. 21

LPSN List of Prokaryotic names with Standing in Nomenclature. 30

LRS Long-Read Sequencing. 54

MGE Mobile Genetic Element. 33

MIC Mobile Insertion Cassettes. 40

MITE Miniature Inverted repeats Transposable Elements. 40

MMR Mismatch repair system. 28

mRNA messenger RNA. 26

MSA Multiple Sequence Alignment. 63

NCBI National Center for Biotechnology Information. 57

NGS Next-Generation Sequencing. 54, 58

NTP Nucleoside Triphosphate. 26

NW Needleman-Wunch algorithm. 59

RBH Reciprocal Best Hits. 72

RNA Ribonucleic acid. 23, 250

SNP Single Nucleotide Polymorphism. 47

SSR site-specific recombinase. 49

SW Smith and Waterman algorithm. 59

T4CP Type 4 Coupling Protein. 35, 42

T4SS Type 4 secretion system. 35, 42

TE Transposable Element. 39

Tn Transposon. 39

tRNA transfer RNA. 27

RÉSUMÉ/ABSTRACT

Résumé / Abstract

Outils pour la génomique comparative des bactéries à large échelle : développement et applications.

Résumé:

La génomique comparative bactérienne consiste à comparer les contenus en gène des différentes souches : leur pangenome. Avec le nombre croissant de séquençages, les logiciels existants au début de cette thèse arrivaient à leurs limites en termes de temps de calcul et de mémoire. L'enjeu était de passer à l'échelle de milliers de génomes dans un temps raisonnable, en gardant une précision correcte. De plus, à notre connaissance, aucun logiciel ne permettait d'effectuer toutes les étapes clés d'une étude de génomique comparative. C'est dans ce contexte que nous avons développé PanACoTA, un outil ayant pour but de standardiser et automatiser la préparation de données pour ces études, depuis le téléchargement des génomes et leur contrôle qualité jusqu'à l'inférence de l'arbre phylogénétique du core génome (gènes communs à tous les génomes). Son implémentation sous forme de modules a été pensée pour permettre de s'adapter aux besoins spécifiques de certaines études (exploration de paramètres, étapes supplémentaires). Concernant le module « pangenome », nous avons développé une nouvelle méthode, s'appuyant sur des outils récents de comparaison et clustering de séquences. Robuste aux changements d'échelle, elle permet de calculer un pangénome de 4000 souches en 30 minutes. Au cours de son développement, nous avons appliqué PanACoTA dans différents contextes. Nous avons montré l'utilité de l'outil sur des études à court terme (recherche de la particularité d'une souche épidémique d'*E. anophelis*), sur du long terme (étude de la diversité génomique de l'espèce *E. coli*), ou encore pour différencier différentes espèces d'un genre peu connu (*Morganella*).

Mots clés : `[bactéries, génomique comparative, pangenome, développement et génie logiciel]`

Tools for massive bacterial comparative genomics: Development and Applications.

Abstract:

Bacterial comparative genomics consists in comparing the gene contents of different strains: their pangenome. With the increasing number of strains sequenced, the tools available when I started this PhD were reaching their limits in terms of computation time and space. The aim was to develop a method able to handle thousands of genomes, accurately and in a reasonable amount of time. Besides, to our knowledge, no tool was able to do all key steps of any comparative genomics study. This spurred the development of PanACoTA, a tool to standardize and automatize the process to build the key collections of data needed for these studies. This includes all steps from downloading genomes with a quality control until the inference of a phylogenetic tree based on the core genome (genes shared by all strains). In order to be able to adapt to specific needs (exploration of parameters, additional steps), we implemented it in a modular way. For the "pangenome" module, we developed a new method, based on recent tools of genome comparison and clustering. Robust to changes in sampling size, this method can infer a pangenome of 4000 strains in 30 minutes. During its development, we applied PanACoTA to different kinds of studies. We showed its usefulness for short-term studies (find specificity of a pathogenic strain of *E. anophelis*), long-term (genomic diversity of *E. coli* species), or to identify different species in an little-known genus (*Morganella*).

Keywords: `[bacteria, comparative genomics, pangenome, software development]`