



UNIVERSITÉ  
PARIS  
DESCARTES

USPC  
Université Sorbonne  
Paris Cité



Université Paris Descartes  
École doctorale Interdisciplinaire Européenne 474 Frontières du Vivant  
Microbial Evolutionary Genomic, Pasteur Institute

# Evolutionary genomics of conjugative elements and integrons

Thèse de doctorat en Biologie Interdisciplinaire

Présentée par **Jean CURY**

Pour obtenir le grade de Docteur de l'Université Paris Descartes

Sous la direction de Eduardo ROCHA

Soutenue publiquement le 17 Novembre 2017, devant un jury composé de:

|                   |                    |   |
|-------------------|--------------------|---|
| Claudine MÉDIGUE  | Rapporteuse        | CNRS, Genoscope, Évry                   |
| Marie-Cécile PLOY | Rapporteuse        | Université de Limoges                   |
| Érick DENAMUR     | Examineur          | Université Paris Diderot, Paris         |
| Philippe LOPEZ    | Examineur          | Université Pierre et Marie Curie, Paris |
| Alan GROSSMAN     | Examineur          | MIT, Cambridge, USA                     |
| Eduardo ROCHA     | Directeur de thèse | CNRS, Institut Pasteur, Paris           |



# لاعب النرد

محمود درويش

مَنْ أَنَا لِأَقُولَ لَكُمْ  
مَا أَقُولُ لَكُمْ؟  
وَأَنَا لَمْ أَكُنْ حَجَرًا صَقَلَتْهُ الْمِيَاهُ  
فَأَصْبَحَ وَجْهًا  
وَلَا قَصَبًا ثَقَبَتْهُ الرِّيحُ  
فَأَصْبَحَ نايًا ...  
أَنَا لَاعِبُ النَّرْدِ ،  
أَرَبِحُ حِينًا وَأَخْسِرُ حِينًا  
أَنَا مِثْلُكُمْ  
أَوْ أَقَلُّ قَلِيلًا ...

## The dice player

Mahmoud Darwish

Who am I to say to you  
what I am saying to you?  
I was not a stone polished by water  
and became a face  
nor was I a cane punctured by the wind  
and became a flute...

I am a dice player,  
Sometimes I win and sometimes I lose  
I am like you or slightly less...





# Contents

|  |           |
|--|-----------|
| <b>Acknowledgments</b>                               | <b>7</b>  |
| <b>Preamble</b>                                      | <b>9</b>  |
| <b>I Introduction</b>                                | <b>11</b> |
| 1 Background for friends and family . . . . .        | 13        |
| 2 Horizontal Gene Transfer (HGT) . . . . .           | 16        |
| 2.1 Mechanisms of horizontal gene transfer . . . . . | 16        |
| 2.1.1 Transformation . . . . .                       | 16        |
| 2.1.2 Conjugation . . . . .                          | 17        |
| 2.1.3 Transduction . . . . .                         | 18        |
| 2.1.4 Other mechanisms . . . . .                     | 18        |
| 2.2 Evolutionary consequences . . . . .              | 20        |
| 2.2.1 HGT and bacterial ecology . . . . .            | 20        |
| 2.2.2 HGT and bacterial chromosome . . . . .         | 21        |
| 3 Mobile Genetic Elements (MGE) . . . . .            | 25        |
| 3.1 Different types of MGE . . . . .                 | 25        |
| 3.2 Persistence of MGE . . . . .                     | 27        |
| 3.2.1 Replication . . . . .                          | 28        |
| 3.2.2 Partition system . . . . .                     | 29        |
| 3.2.3 Integration . . . . .                          | 34        |
| 4 Deciphering the bacterial chromosome . . . . .     | 36        |
| 4.1 Formal grammars . . . . .                        | 37        |
| 4.2 Hidden Markov Model profiles . . . . .           | 38        |
| 4.3 Covariance Models . . . . .                      | 40        |

|            |   |            |
|------------|---|------------|
| <b>II</b>  | <b>Contributions</b>  | <b>43</b>  |
| <b>1</b>   | <b>Conjugation</b>  | <b>45</b>  |
| 1.1        | Introduction . . . . .  | 45         |
| 1.1.1      | Background . . . . .  | 45         |
| 1.1.2      | Diversity . . . . .   | 46         |
| 1.1.3      | Mechanism . . . . .   | 47         |
| 1.1.4      | Objectives . . . . .  | 49         |
| 1.2        | Methods . . . . .   | 51         |
| 1.2.1      | <b>Article 1:</b> Identifying conjugative plasmids and integrative conjugative elements with CONJScan . . . . .   | 51         |
| 1.3        | Results . . . . .   | 79         |
| 1.3.1      | <b>Article 2:</b> Integrative and conjugative elements and their hosts: composition, distribution, and organization . . . . .                                 | 79         |
| 1.3.2      | <b>Article 3:</b> Host range expansion and genetic plasticity drive the trade-off between integrative and extra-chromosomal mobile genetic elements . . . . . | 94         |
| 1.4        | Conclusion . . . . .  | 125        |
| <b>2</b>   | <b>Integrans</b>  | <b>129</b> |
| 2.1        | Introduction . . . . .  | 129        |
| 2.1.1      | Background . . . . .  | 129        |
| 2.1.2      | Diversity . . . . .   | 129        |
| 2.1.3      | Mechanism . . . . .   | 130        |
| 2.1.4      | Objectives . . . . .  | 132        |
| 2.2        | Methods and Results . . . . .   | 133        |
| 2.2.1      | <b>Article 4:</b> Identification and analysis of integrans and cassettes arrays in bacterial genomes . . . . .  | 133        |
| 2.2.2      | <b>Article 5:</b> Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance . . . . .                 | 146        |
| 2.3        | Conclusion . . . . .  | 163        |
| <b>III</b> | <b>Conclusions and Perspectives</b>   | <b>167</b> |
|            | <b>Bibliography</b>   | <b>179</b> |
|            | <b>Annexes</b>  | <b>195</b> |

# Remerciements

Je voudrais d'abord remercier Eduardo de m'avoir proposé de faire une thèse dans son labo, alors que je comptais encore faire des manip à l'époque! Ça fera quasiment ~~trois~~ quatre cinq ans passés dans ton labo, et je dois dire que ce fût un réel plaisir d'y travailler (et pas seulement grâce à la terrasse!). Toute cette thèse ne serait pas grand chose sans ton encadrement, ta disponibilité, ton savoir encyclopédique, et ta confiance. Un grand merci aussi à Marie de m'avoir guidé dans les arcanes de la bioinfoEduardique.

Cette thèse n'aurait pas été un tel plaisir sans ce bureau 07E. Il est d'abord très bien placé, au 6<sup>ème</sup> étage d'un bâtiment neuf, il offre un accès direct à une terrasse plein sud de plus de 100m<sup>2</sup>, tout en étant suffisamment distant du bureau du boss, permettant une arrivée en douce à toute heure de la journée. Ce bureau ne serait rien sans les 3 co-thésardes qui m'ont accompagné dans ce périple et sans qui d'inestimables discussions n'auraient jamais eu lieu, au grand dam du monde probablement. Camille, Aude, Camille, merci d'avoir supporté mes bruits, mon bordel, mes pulls troués, mes plantes, la doc. Vous m'avez fait aussi perdre vachement de temps. Merci pour tout. Je voudrais aussi remercier tous ces gens que j'ai vu défiler dans ce labo, qui ont aussi participé à sa richesse, et qui d'une façon ou d'une autre ont participé à ce travail. Merci à Brigitte pour toute sa disponibilité.

Tout cette aventure est devenue possible grâce à la rencontre de deux personnes clés qui m'ont conduit à pousser la porte du labo d'Eduardo. D'abord Julian qui m'a initié à la microbiologie en toute simplicité, en me montrant quels solvants pouvait être pipetés à la bouche pour étudier un truc<sup>1</sup> trouvé en forêt. Julian m'a encouragé à aller travailler avec Didier, grâce à qui j'ai pu continuer mon apprentissage de la microbiologie, entouré d'une équipe décapante. Merci pour tout Didier, désolé pour les librairies, mais merci de m'avoir conseillé d'aller faire ce stage chez Eduardo.

Un grand merci aux Marsouins et à ce groupe de compères aimant la recherche comme la Chouffe, mainenant dispersés à travers le monde: Jehanne, Antoine, Aleks, Antoine, Clément, Ariane, Paul, Aude, Vincent, Fati, Hassen, Khalifa, les frites. Merci au CRI et à ce qu'il est et à tous les gens géniaux qu'on peut y croiser. Beaucoup ont déjà été cités précédemment. Je rajouterai juste Flora, merci d'avoir partagé ton énergie en tant que représentante des étudiants,

---

<sup>1</sup>un "slime mold", après enquête

ou en tant que co-organisatrice (avec Aleks) des Visionary & Retrospective talks. La thèse m'a rendu un peu plus solitaire que je ne le suis déjà, mais merci à tous les potos d'être là, en Albanie, à Marseille, à Essernay, à Alrance, sur l'île d'Houat, ou à Pernes. Théo, merci pour ce subtil soutien depuis bien longtemps. Nono, JibJib, NicNic merci de m'avoir laissé vous conduire à Mont de Marsan en Che Clio Sudaka. Julie merci d'avoir lancé ma carrière d'acteur. Gregoire Chéron. Merci à mon ami Meidi pour ces nombreuses années à se charier. Merci à Marjo pour la soupe aux courgettes. Ma photo officielle sur le site de l'Institut Pasteur — un jour d'été, en revenant de Bora-Bora — vous est naturellement dédié, Léo, Charlie, Antoine, Édith, Alexis, John, avec qui j'ai fait mes premiers bacs et mes premiers rappels sur arbre mort. Merci à mes parents et ma sœur pour leur soutien et encouragements. Merci à Didine d'être la meilleure grand-mère. Merci au reste de la famille pour les discussions passionnées à toute heure et à rosé. Merci à Jehanne pour tout ce que tu m'apportes depuis 9 ans, 5 mois et 1 jour. C'est inédit.

Merci aux correcteurs de cette thèse, ma mère, Aleks et Jehanne.



# Preamble

The goal of this thesis is to develop and evaluate methods to study two types of genetic elements whose role in the spread of antibiotic resistance genes is no longer to be demonstrated.

Conjugative elements on the one hand are known for their ability to spread among a wide range of bacteria and their plasticity allows them to carry numerous functions, and notably antibiotic resistance genes. On the other hand, integrons are known for their specific capacity to capture genes and rearrange them to minimize the fitness cost for the bacteria. They have been particularly successful in capturing antibiotic resistance genes. Research carried on these elements has been mainly focused on variants carrying antibiotic resistance genes, highly successful in pathogenic bacteria. However, the broader evolutionary picture of these elements remains sparse as efficient tools to detect them in bacterial genomes are lacking.

In Part I, I introduce the key concepts of horizontal gene transfer and mobile genetic elements as they play a major role in bacterial evolution. Conjugative elements and integrons are two different types of mobile genetic elements, and they spread in bacterial populations through horizontal gene transfer. I also present the basics of two methods on which the tools and methods developed in this thesis are heavily relying on.

Part II will present the contributions I made regarding these two elements. In chapter 1, I present a method that allows the detection of conjugative elements and the delimitation of integrative and conjugative elements at large scale, without imposing an a priori on the functions carried by the elements. In chapter 2, I present a tool that identifies integrons in bacterial genomes, not restricted to previously discovered integrons. Both methods allowed to produce the largest and most genetically diverse datasets of the corresponding elements so far. Each chapter also presents the subsequent analysis made on these datasets.

Together, this work aims at better understanding these elements, with the goal of providing tools and knowledge to improve their surveillance. It also sheds light on the evolution of mobile genetic elements in general, and fuels, at its scale, the global understanding of bacterial evolution.



# Part I

## Introduction



# 1 Background for friends and family

**The Resistance Awakens.** Bacteria represent one of the three kingdoms of life, together with Archaea and Eucaryotes. They are unicellular organisms, and are classed as prokaryotes with Archaea because they do not contain nuclear membrane. They are found in all major environments on earth. Their amount was evaluated as about  $10^{30}$  cells [201]. Typically, there is the same number of bacteria inhabiting a human's body as there are cells constituting it, about  $10^{13}$  cells, but they represent 0.3% of the total mass [170]. Bacteria have been on earth for a few billion of years [51], and have been discovered more than three centuries ago with the invention of the microscope [197]. Research on bacteria emerged about two centuries ago, and the inter-bacterial traffic of genetic information, the driver of bacterial evolution, was discovered 70 years ago [183].

Antibiotic-associated research has spurred the development of microbiological research, from the discovery of new compounds to the understanding of antibiotic resistance in bacteria. Antibiotic resistance proteins had already been discovered before the industrial use of penicillin [2] and warnings on the use of antibiotics were already formulated by A. Fleming in his Nobel lecture [68]:

*There may be a danger [of penicillin administration], though, in underdosage. It is not difficult to make microbes resistant to penicillin in the laboratory by exposing them to concentrations not sufficient to kill them, and the same thing has occasionally happened in the body. [...] Here is a hypothetical illustration. Mr. X. has a sore throat. He buys some penicillin and gives himself, not enough to kill the streptococci but enough to educate them to resist penicillin. He then infects his wife. Mrs. X gets pneumonia and is treated with penicillin. As the streptococci are now resistant to penicillin the treatment fails. Mrs. X dies. Who is primarily responsible for Mrs. X's death?*

Although resistance was expected as shown above, people at the time could not have anticipated the extent and velocity of the spread of antibiotic resistance. Notably because they did not expect the level of antimicrobial use [47], and the inter-bacterial exchange of genetic information was yet to be fully understood. Ever since, it has been an arms race: on the one hand researchers discovering, and later synthesizing, antibiotics, and on the other hand, bacteria becoming more and more resistant. This is exemplified by this citation from J. Davies, 1994 [47]:

*It is frightening to realize that one single base change in a gene encoding a bacterial  $\beta$ -lactamase may render useless \$100 million worth of pharmaceutical research effort.*

Nowadays, new antibiotics arrive sparingly, if at all, on the market [34]. Other alternatives to antibiotics exist and involve different methods already used in medical settings or not, like vaccines, antibodies, probiotics or phages. It has been estimated that there exists 19 such

solutions, which need further research before utilization [46]. More importantly, the lack of good practice at global scale in the use of antibiotics was, and to some extent still is, responsible for the emergence and selection of antibiotic resistance strains. It has been proposed, to avoid the spread of any resistance, to impose a strict control on the use of antibiotics or any other alternative treatment [181].

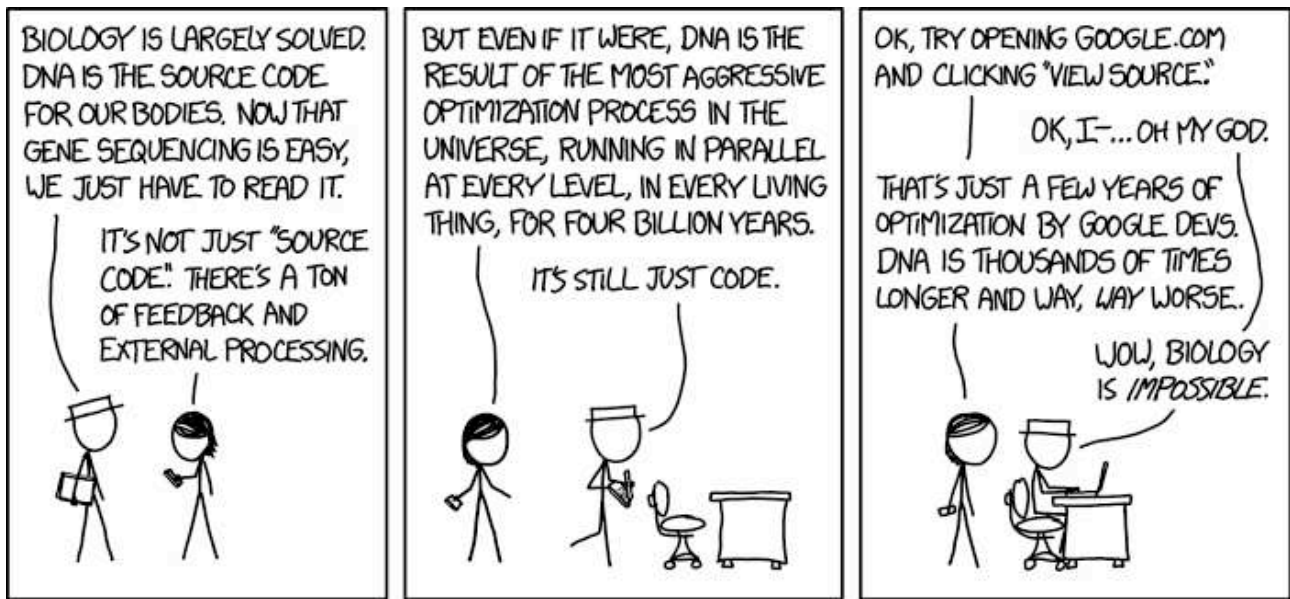


Figure 1: DNA. *Title text: Researchers just found the gene responsible for mistakenly thinking we've found the gene for specific things. It's the region between the start and the end of every chromosome, plus a few segments in our mitochondria.* From <https://xkcd.com/1605/>.

**From the base pair to the binary digit.** In the meantime, bioinformatics, as we know it today, was a by-product of the invention of the sequencing of protein and later DNA by F. Sanger [168, 169]. These double Nobel prize discoveries initiated the need to use computers in genetics research. Indeed, the combination of the digital representation of these sequences (a repetition of the four nucleotides A/T/C/G for the DNA or of the 20 amino-acids for proteins) and the size of a sequence render impossible for a human to treat that information (Figure 1). Ironically, the power of computers is growing slower than the decrease in cost of DNA sequencing [200]. One can grasp this price drop with the evolution of the price for sequencing a complete human genome, from 100 million US\$ in 2001 to about 1000 US\$ today, and it is still going down. The result, as seen in Figure 2, is an exponential increase of available genomes. But most of them are partially assembled (they are called “draft genomes”), meaning that part of the information is missing. Handling this amount of data requires computers, but also efficient algorithms to transform the data into meaningful information for biologists. This defines two stereotypical types of bioinformaticians. On the one hand, those who develop algorithms and

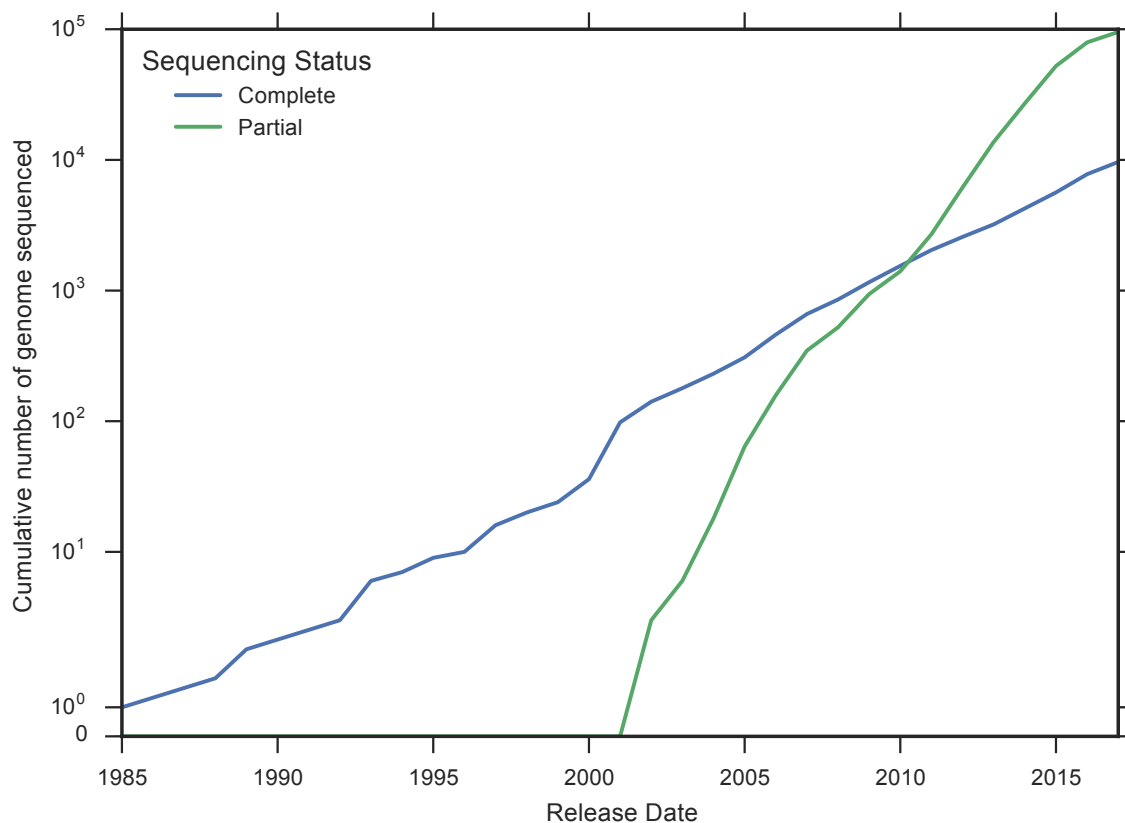


Figure 2: Number of prokaryotic genomes sequenced over the years. The lines represent the cumulative number of genome sequenced, depending on the type of sequencing. Data obtained from [ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME\\_REPORTS/prokaryotes.txt](ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt)

tools to analyze biological data. They aim at answering computational questions. And on the other hand, those who use these tools to analyze biological data. They aim at answering biological questions. My work leans toward the latter type. During my PhD, I developed tools and methods on top of those created by the former type to gather biological data, from complete bacterial genomes. I have asked evolutionary questions on genetic elements favoring the spread of antibiotic resistance genes and other adaptive functions, and I have been trying to answer these questions with the tools I developed.

## 2 Horizontal Gene Transfer (HGT)

Bacterial populations expand by clonal (or asexual) reproduction. Early in the history of bacterial research, it has been shown that bacteria could acquire exogenous genes by transformation [83, 9], transduction [206] or by conjugation [183]. These processes were at the time considered as a classical sexual phase of the bacterial life cycle by analogy with eukaryotes. Following this idea, it was proposed to group these mechanisms under the term *meromixis*, which is the process characterized by a partial genetic transfer leading to the creation of a partial zygote or *merozygote* [203]. However, it was realized in the late 1950s that these mechanisms could transfer antibiotic resistance genes between different species [4, 146]. These experiments showed the transfer of antibiotic resistance genes between a strain of *Escherichia coli* and a strain of *Shigella*. This was the first time that genetic flow between species was observed. Ironically, strains of the *Shigella* genus was later shown to actually belong to the *Escherichia coli* species [22, 23, 65]. Yet, the results hold. The bacterium *Agrobacterium tumefaciens* was described first in 1907 [175] as a bacterium inducing the production of tumor in plants. In the early 1970s it was understood that a plasmid was transferred to the plant and expressed there to facilitate bacterial infection (and induction of plant tumor) [112]. Thus, horizontal gene transfer (HGT) can occur between different kingdom of life, and has been actually observed early in the history of microbiology [175, 83].

Finally, in every kingdom of life, there are processes known as horizontal or lateral gene transfer (HGT or LGT), by opposition to vertical gene transfer (where the genes follow the traditional genealogy). These processes are important in bacteria as they represent a major source of acquisition of new genes or new functions. These processes are also important in molecular biology, genetics and in biotechnology. For instance, the first genetic map of a bacterial chromosome was established using conjugation [203], and since the 1950s, researchers use the properties of HGT to introduce genetic systems in organisms to study them [36].

In the following section, I will describe the different mechanisms of HGT and their evolutionary consequences on bacterial evolution.

### 2.1 Mechanisms of horizontal gene transfer

#### 2.1.1 Transformation

Transformation was the first mechanism of horizontal gene transfer to be discovered. In 1928, a “principle” able to *transform* bacteria’s phenotype was discovered [83]. This “principle” was later shown to be genetic information by the Avery-MacLeod-McCarty experiment [9]. The mechanism allowing the acquisition of genetic information was then called transformation. More precisely, transformation is the process characterized by the uptake of DNA from the environment by the bacteria. It depends entirely on the bacterial cell and does not require mobile



genetic elements. The genes encoding the competence system are often present in all strains of a species but few bacterial species have been shown to be naturally transformable [110]. All transformable bacteria share the same set of genes, usually referred as *com* regulons [110], at the exception of *Helicobacter pylori* [104]. Transformation consists in three steps. First, donation or DNA release in the environment upon the lysis of a bacterial cell. Certain bacteria can also secrete DNA to the environment [95]. Second, the DNA is up-taken by the bacterial cell if the competence system is activated; the bacterium is said competent. The state of competence may not be permanent. Regulation of competence tends to evolve faster than the molecular machinery, which results in a variety of regulating signals, like cell-cell signaling, stressful condition or nutritional depletion [110]. The DNA enters the cell in a single stranded form and is often integrated into the chromosome by homologous recombination, or by illegitimate recombination if not degraded (Figure 3). The imported material involves small pieces of DNA [44], which tends to reduce genomes' size. This property of transformation led to the hypothesis that transformation could inhibit vertical transmission of mobile genetic elements [45]. Other roles have been proposed for the imported DNA, among DNA as a nutrient source, genome maintenance, genome diversification [110].

### 2.1.2 Conjugation

The mechanism of conjugation was discovered three years after transformation [119, 183]. The term conjugation pre-existed this discovery, and was used to refer to isogamy (sexual reproduction between two gametes of similar morphology) events between eukaryotes [8]. Contrary to transformation, conjugation is a mechanism that occurs between two cells that are in close contact, and the transfer is independent of the cells. The conjugation machinery is encoded on a mobile genetic element, which is either integrated into the chromosome, or extra-chromosomal (Figure 3). Bacterial conjugation refers essentially to two distinct mechanisms, involving the transfer of either a single stranded DNA (ssDNA) or a double stranded DNA (dsDNA). But another type of conjugation has been observed recently and the mechanism is not well understood yet. It is specific to *Mycobacteria* and called distributive conjugal transfer [49]. It transfers DNA fragments of size ranging from 50bp to more than 200kb, which will be integrated at different positions in the recipient's chromosome [49].

Conjugation relying on ssDNA will be the subject of the chapter 1 of this thesis. Briefly, ssDNA conjugation has been found in all major bacterial clades, and involves conjugative elements either integrated or not [90]. It is composed of a multi-component system called type 4 secretion system (T4SS), a relaxase, and a type 4 coupling protein (T4CP). The relaxase nicks the dsDNA at the origin of transfer (*oriT*), attaches covalently to one of the DNA strands, and this nucleo-protein complex, with the help of the T4CP, will be secreted through the T4SS to enter the recipient cell. The dsDNA conjugation, on the other hand, is restricted to

*Actinobacteria*, which can form mycelia while growing. This conjugation corresponds to an intermycelial transfer. A single protein is essential for this mechanism of transfer [184] and does not involve a T4SS.

Finally, ssDNA conjugation, hereafter called conjugation, is the mechanism of horizontal gene transfer that can transfer DNA over the longest phylogenetic distances, and as stated previously, it can notably transfer DNA from bacteria to plants [112].

### 2.1.3 Transduction

Transduction was discovered in 1952 by Zinder and Lederberg [206]. These researchers were looking for recombination events like those occurring after conjugation in a bacterium where conjugation was not observed yet. They found recombination events, which occurred despite the presence of a filter between the two bacterial cultures preventing conjugation. The size of the filter's pores allowed them to determine that the agent of the genetic transfer was of the size of bacteriophages that had already been discovered for almost forty years at the time [196]. Phages (or bacteriophages) are bacterial viruses. Different types of phages exist, defining different types of transduction mechanisms. Virulent phages upon infection of a cell, replicate extensively, produce phage particles (virions) and lyse the cell to infect other cells — this is the lytic cycle [27]. Temperate phages can either enter the lytic cycle or a quiescent state, where they integrate the host's chromosome or remain in a plasmid-like form — this is the lysogenic cycle, and the phage becomes a prophage [27]. The expression of the prophages' genes can lead to phenotypic changes for the host, in a process called lysogenic conversion [189]. Upon an inducing signal, temperate phages excise from the chromosome and enter the lytic cycle [157].

Transduction is the process of transferring the genetic material of a bacterium to another by the intermediate of phages (Figure 3). Different types of transduction exist. Transduction is said generalized when, during the lytic cycle, phages degrade and encapsulate its host's DNA. Any part of the host's genome can be captured and further transferred. This event is relatively rare, it has been estimated for the phage P22 that about 2% of virions contain DNA from the donor chromosome [59]. The amount of DNA transferred is proportional to the phage's size, and can be substantial, up to 93kb for phage P1, for instance [123]. Specialized transduction occurs when an integrated temperate phage enters the lytic cycle, and upon excision with secondary recombination sites, captures neighboring chromosome genes [27]. These genes will be then transferred into a new host.

### 2.1.4 Other mechanisms

There are other mechanisms of horizontal gene transfer (Figure 3) but their impact is less clear and might be less important than the previously mentioned mechanisms and are often

specific to certain clades. The literature is accordingly much sparser. Gene transfer agents (GTAs) are thought to derive from phages after partial deletion of the phage's genomes. They transfer random pieces of DNA, and the amount of DNA transferred is insufficient to package the locus encoding the GTA [117]. Other elements with intermediate properties between GTA and transducing phages have been described, and are named phage-like particles [117]. The transfer of genetic information between organisms can take place through direct or indirect cell membrane contacts. It has been shown that bacteria could form nanotubes between adjacent cells. These nanotubes form a direct contact between the cells and might allow the transfer of DNA, proteins, or metabolites [57, 56, 180, 152]. Membrane cell contact can also occur indirectly due to the secretion of outer membrane vesicles (OMV) in diderm bacteria. OMV can deliver a wide range of substance to its recipient, including DNA or proteins [113]. Antibiotic resistance genes have been found to be transferred this way, notably in the major nosocomial pathogen *Acinetobacter baumannii* [30].

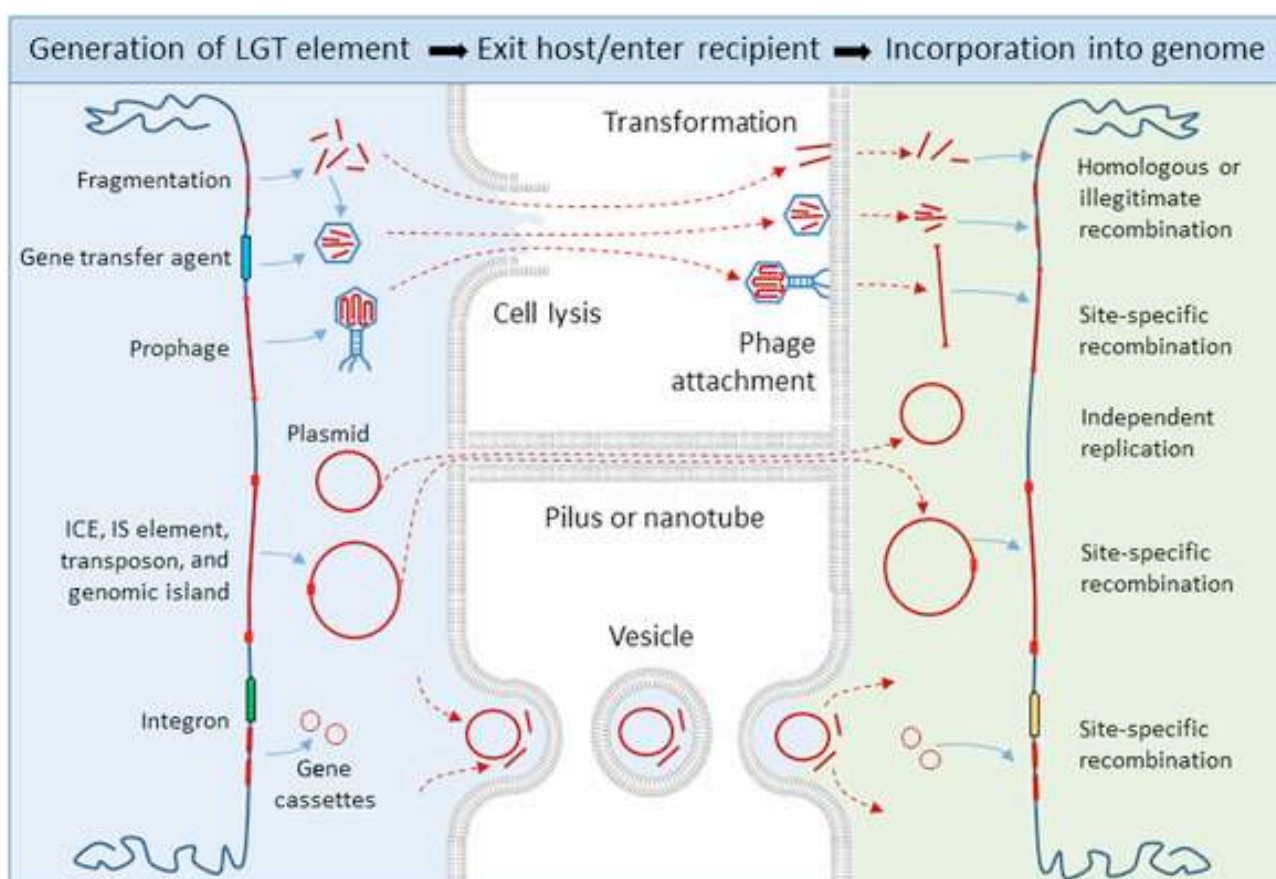


Figure 3: The various mobile genetic elements and corresponding transfer mechanisms. Figure from [79]

## 2.2 Evolutionary consequences

### 2.2.1 HGT and bacterial ecology

HGT resembles sexual reproduction of eukaryotes in that it allows recombination of genetic information between two individuals. But it goes beyond sexual reproduction, as HGT can involve two individuals from different species. HGT occurs between closely related species, or between any two individuals, be they a bacterium and a plant.

Horizontal transfer of homologous genes between related bacteria occurs mainly through conjugation or transformation [176, 125] but possibly with any of the above-mentioned mechanisms. This can introduce genetic variation through allelic recombination between related strains creating novel combinations of alleles [67], but this is less likely to create radically novel traits [147]. This is however a powerful process to remove deleterious mutation [192], notably those hitchhiking along adaptive ones [150].

In contrast, transfer of non-homologous genetic elements mediated through transduction, conjugation or transformation will incorporate a whole set of new functions, and can take place between closely or distantly related bacteria. These evolutionary leaps [86] allow bacteria to adapt rapidly to a newer ecological niche, in case of environmental change or stressful conditions. The most famous example being the massive spread of antibiotic resistance genes, such that it became a threat for human health [204]. Another example consists of the heavy-metal resistance genes found in the early described plasmids in the 1950s which were selected due to the industrial pollutions generated during the 18<sup>th</sup> and 19<sup>th</sup> centuries [173]. Because of these evolutionary leaps, a bacterium can become a pathogen upon a single transfer event. For instance, certain strains of *E. coli* can convert into a uropathogen or an enteropathogen, depending on the integrated element they receive [86]. Similarly, in *Shigella* the pathogenicity locus is not integrated into the chromosome but is on a virulence plasmid and permits the entry and dissemination of the bacteria within intestinal cells [65]. This radical phenotypic change is such that it led people to think that *Shigella* bacteria were a genus when they actually belong the *E. coli* species [22, 23, 65]. Another infamous example concerns the lysogenic conversion of *Vibrio cholerae* by the phage CTX $\Phi$  that encodes the cholera toxin and thus transforms this bacterium in a deadly pathogen [199]. Combined with antibiotic resistance genes, which are more and more widespread among non-pathogenic bacteria, future pathogens might already be multi-resistant to antibiotics before acquiring the virulence factors. This situation has been observed in 2016 in a Chinese hospital where a multi-resistant strain of *Klebsiella pneumoniae* acquired a virulence plasmid [87].

HGT can also involve functions having more positive outcomes from human perspective. Indeed, nitrogen fixation in leguminous plants, one of the most important ecological process, requires the formation of root nodules by a *Rhizobial*, whose genes encoding the formation of

the nodule and the nitrogen fixation are on a conjugative element, integrated or extrachromosomal [43]. Bacteriophages of cyanobacteria also play an important role in the carbon and oxygen cycles [153]. *Cyanobacteria* have been the major contributor of oxygen to the atmosphere for billions of years. Through lytic or lysogenic cycles, phages modulate the levels of carbon dioxide by modifying the population structure of *Cyanobacteria* and by altering bacteria's phenotype. Some of these phages also encode core photosynthetic genes, allowing the photosynthesis to occur even during infection [153].

Bacteria interact, directly or indirectly, with each other by exchanging DNA, but also by secreting molecules in the environment. These molecules have a wide range of functions, from enzymes allowing the acquisition of nutrient or toxins, to molecules triggering antibiotic production, increasing adhesion to a host or favoring biofilm formation [102, 167]. These molecules are shared with the bacterial community at the cost of individual bacterium secreting it. In such a situation, cheaters — bacteria not producing the molecule but taking advantage of it — can emerge. It has been shown that gene encoding these secreted protein were carried by mobile genetic elements [143], but also that the cooperative behavior is favored with HGT by increasing the relatedness among the bacterial community [143, 50]. Indeed, the transfer of a MGE to physically close bacteria will homogenize the alleles of this spatially structured subpopulation, thus, cooperative alleles are favored by kin selection [50].

Thus, HGT has a deep impact on bacterial communities by allowing the rapid spread of numerous functions, removing deleterious alleles and favoring cooperation among bacteria.

### 2.2.2 HGT and bacterial chromosome

HGT impacts deeply the bacterial genome. From all the genes transferred, many remain as part of the chromosome, and they actually represent a sizable proportion of newly acquired genes. Gene families expand due to horizontal gene transfer rather than by gene duplication. It has been estimated that HGT accounts for between 88% to 98% of all gene family expansions within the eight different bacterial clades analyzed [193]. These expansions increase the total genome size (size of the chromosome(s) and the extra-chromosomal elements like plasmids or phages). Conversely, it exists a deletional bias in bacterial genome [138] which tends to suppress DNA fragment lacking of adaptive value. Thus, the size of the bacterial genome depends on the equilibrium between processes of gene accretion, notably through HGT events, and gene deletion [190].

These fluxes of genes have an enormous impact on the gene repertoire of a species or genus. For a given clade the gene repertoire is often characterized by three metrics [136, 188, 187]. The average genome size, which here refers to the number of genes in a genome, gives an order of magnitude for the size of the genome in this clade. The genome size can vary among bacterial clade from 50kb to 13Mb [190], but varies little within a species [147]. The pan-genome which

represents the number of gene families found in the entire clade. It considers the plasticity of the genome and represents the gene pool available for any strain in the clade. The pan-genome varies with the number of strains in the clade. It is said open if every new strain will add specific gene to the pan-genome, this is typically the case for strains that can colonize multiple environment [136]. Conversely, the pan-genome is closed if its size plateaus and adding new genome does not reveal new gene families. This is the case of bacteria living in isolated niches which have difficulty to acquire foreign genes [136]. In contrast, the core-genome describes the number of gene families shared by all the strains in a given clade. This metric decreases with the number of genomes until it reaches a plateau. The core-genome is often approximated with the persistent genomes when using large dataset. The persistent genome corresponds to the set of gene families present in more than 90% or 95% of the genome. It is a less stringent metric than the core-genome which can quickly decrease to zero gene families if a single genome is too divergent (*e.g.* due to bad clade assignation) or if a few genomes are from drafts of poor quality.

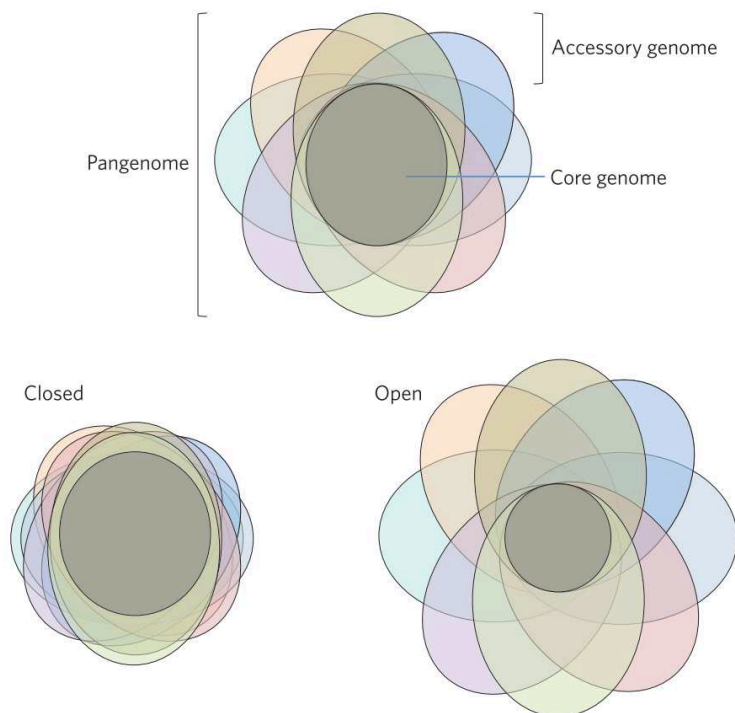


Figure 4: Schematic representation of the Pan- and Core-genomes. Each circle represents a genome whose area indicates the genome size. The overlap of these circles represents the core-genome. The union of these circles represents the pan-genome. The bottom two panels represent a closed and an open pan-genome. The figure is from [135].

Thus, the pan-genome is the union of the gene families of a clade, while the core-genome is the intersection of the gene families (Figure 4). The difference between the pan-genome and the core-genome is called the accessory genome and typically corresponds to the genes horizontally transferred or environment-specific genes. For instance, the pan-genome in *E. coli* has been estimated as four times as big as the average number of gene in a genome, while the core-genomes is twice as small [188]. This result was based on 20 genomes of *E. coli*. Interestingly,

with a dataset of 2085 genomes, the size of the pan-genome for this species increases to be about 20 times as big as the average genome size [116], which is characteristic of an open pan-genome. In *E. coli* the persistent genome stays stable, whether with 20 genomes or 2085, at around 3000 genes [188, 116].

The bacterial chromosome is a highly organized structure [190]. Its organization is shaped by the continuous cellular processes acting on it (Figure 5). Most bacterial genes are organized

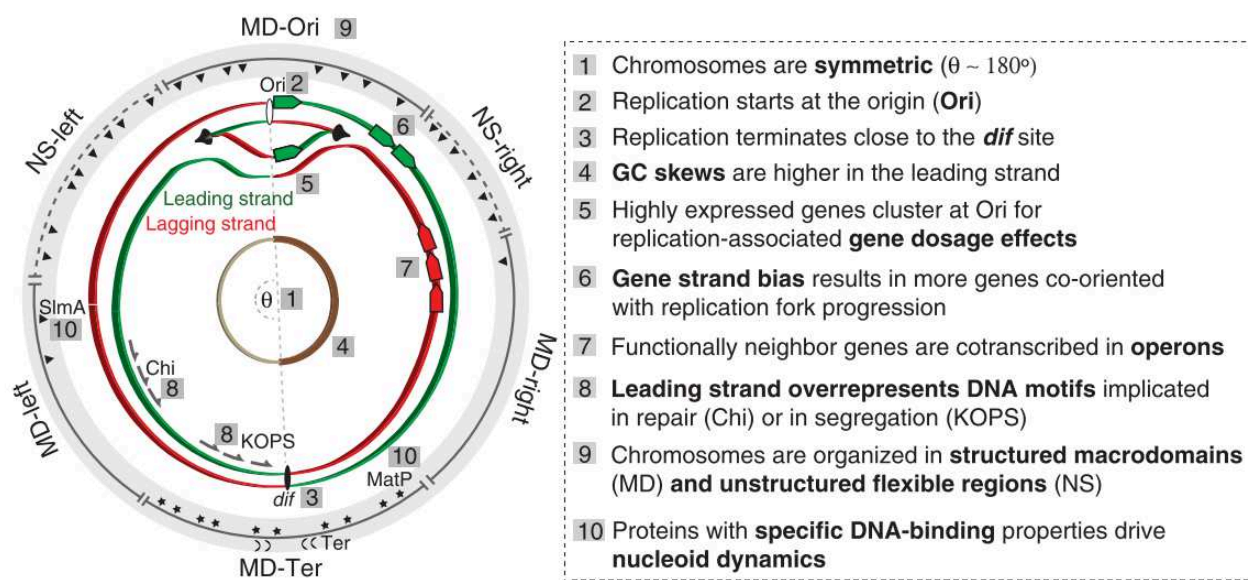


Figure 5: Schema of an organized bacterial genome. The figure is from [190] and represents the different outcomes of cellular processes shaping the bacterial chromosome

in operons, which form a condensed array of co-transcribed and co-regulated genes, leading to high genetic linkage among them. Genes in an operon tend to be more conserved [161] and share similar functions [162]. At a larger scale, other features associated with constraints imposed on the chromosome by its interactions with cellular processes can be observed. For instance, highly expressed genes tend to accumulate toward regions near the origin of replication, profiting from a gene dosage effect in fast-growing bacteria [198]. The gene dosage effect is created by the need for fast-growing bacteria to undergo many chromosomal replications at the same time since it takes more time to replicate a chromosome than to duplicate a bacterial cell [38]. Genes also tend to be located preferentially on the leading strand such that their transcription is co-oriented with the replication of the chromosome, which minimizes the impact of the collision between the DNA and RNA polymerases [179]. These organization constraints are selected to diminish the cost of maintenance and expression of the cellular machinery.

Similarly, the cost imposed by the massive flux of horizontally transferred genes can be diminished by organizational features. Indeed, the relative fitness induced by the transfer of exogenous material on the chromosomal organization depends on where it integrates, and

whether it disrupts or not the organization. In *E. coli*, most accessory genes were found in a few conserved regions of the chromosome [188]. A recent study with 80 bacterial species found that most of the horizontally transferred genes were indeed found in a very few number of loci along the chromosomes [148]. This compromise between diversification and organization was even more accentuated for genomes with high rate of HGT. Extra-chromosomal elements may also help to accommodate the accumulation of horizontally transferred genes.

Overall, HGT is at the core of bacterial evolution, and has profound consequences on bacterial life style. These consequences translate into global health issues with the spread of virulence and antibiotic resistance genes [204]. It also plays an important role in the biotechnology sectors [133]. The potential of gene transfer has been used widely to insert exogenous elements in bacterial chromosomes, but also across kingdom of life, to obtain a desired phenotype. It has notably a wide range of application, from the bioremediation of polluted environments [73], to the production of a chemical or protein of interest [133]. All these applications take advantages of the ease to transfer genetic material, but face the other side of the coin: this genetic material can leave as easily as it enters. In environmental applications of biotechnology such as bioremediation, biofuels, biofilters, biotransformation, etc. . . , it raises the issue of the pollution with genetic materials [74], while in drug production, the issues relate on the stability of the system and on preventing infections from other mobile genetic elements [126]. Thus, HGT is also a challenge for the biotechnological companies.



## 3 Mobile Genetic Elements (MGE)

### 3.1 Different types of MGE

HGT is often driven by mobile genetic elements moving between organisms. But other elements are said mobile because of their ability to move within a chromosome, or more generally between replicons. Hence, the word “mobile” in mobile genetic elements has a two-fold meaning.

First, mobile refers to the ability of the element to mobilize itself from an organism to another, and thereby leads to a HGT event. Conjugative elements and phages belong to this category (Figure 3). Both contain all genes necessary for their transfer, they are autonomous in that respect. These elements can be integrated into the host chromosome or they can be extrachromosomal. Both forms are frequently found in conjugative elements [90]. Lytic phages are extrachromosomal by nature, while lysogenic phages are more often integrated than extrachromosomal [27, 123]. Conjugative elements will be introduced more exhaustively in chapter 1. Briefly, they are transferred through conjugation and are widespread among bacteria [90]. They transfer through the mating pair formation system used as a channel between the donor and the recipient cells (Figure 3, see section 2.1.2 above for more details). In contrast, upon transfer, phages form infectious particles or virions, a physical protection enveloping the genetic information. Virions have different morphologies of their structural component, which can be made of proteins or lipids [3]. Phages’ genetic information can be in any form of DNA (double or single stranded, linear, circular, superhelical or segmented arrangements) or RNA [3]. These morphological variations are used to class phages in more than 100 families, even if actually about 96% of phages described belong to the three families of the *Caudovirales* order [3]. Conjugative elements are thought to be broader host range than phages, although the fact that phages are narrow host range is being challenged [189]. Both phages and conjugative elements are widespread among bacteria and carry a wide range of adaptive functions like virulence genes. However, the presence of antibiotic resistance genes in phages less clear [64, 37].

Other mobile elements do not have the possibility to transfer to another organism without the help of self-transmissible mobile genetic elements, or by using a mechanism of horizontal transfer like transformation or outer membrane vesicles. Integrons are genetic elements composed of a specific integrase and an array of cassettes. They can capture cassettes and rearrange the array upon expression of the integrase. The integron cassette could be the smallest element fitting this definition of mobility, which has for unique defining feature, a recombination site ranging from about 50 to 150bp, but often contains a gene [93]. A cassette is mobile because it can move from an integron to another (Figure 3). Integrons and their cassettes will be described in depth in chapter 2. Transposable elements constitute another class of elements with this type of mobility. Insertion sequences (IS) and transposons are part of this class. IS are larger than integron cassettes, their size is generally smaller than 2.5 kb [128]. They encode

nothing else than what is necessary for their intra-genomic mobility, namely a transposase (or two) flanked by inverted repeats, recognized as recombination sites. The transposases are usually DDE-transposases [128]. IS can be associated with passenger genes, which are often those found in MGE like antibiotic resistance genes, methyltransferases, or regulators [172]. A transposase and passenger genes flanked by inverted repeats form a transposon.

For most of these elements, mobilizable or non-autonomous derivatives exist. They do not encode all the genes necessary for their transfer or displacement, the full system is coded in *trans*. For instance, there are mobilizable elements that can conjugate by using the conjugative system from a complete conjugative system present in the same cell [174]. Miniature Inverted repeats Transposable Elements (MITEs) and Mobile Insertion Cassettes (MICs) are also in this category, and similarly to integron cassettes, they encode two recombination sites of the IS with passenger in the case of MICs but not the transposase [172]. Integron-cassettes are different in that they integrate near their integrase, while MITEs and MICs can transpose into many different sites within a genome due to the low specificity of their recombination site [172]. However, we will characterize in chapter 2 the existence of clusters of cassettes lacking integrase nearby, whose some of them have probably inserted at secondary recombination sites. Importantly, although these elements harbor different levels of autonomy in term of mobility, they need the bacterial machinery to be replicated and to have their proteins expressed. They very rarely encode their own polymerase, albeit some phages have their own RNA or DNA polymerase [13, 27].

At the crossroad of these two types of mobility is the workhorse of molecular biology: the plasmid. Some are conjugative or mobilizable. In contrast, others encode no proteins allowing them to transfer horizontally, even in *trans*. Conjugative plasmids tend to be larger than mobilizable ones [174]. Non-transmissible are on average smaller, and can encode down to a single protein, often a replication protein. Some plasmids can be found without any genes [174]. However, some non-transmissible plasmids can also be larger than certain bacterial chromosomes. In addition, some secondary chromosomes are thought to be ancient plasmids such as the secondary chromosome of *Vibrio cholerae* [100]. The term chromid has been dubbed to distinguish these elements, at the frontier between chromosomes and plasmids [97]. Chromids may not engage in horizontal transfer as unit anymore, and they can carry essential genes. The constraints allowing the transition of a plasmid into a chromid are not clear yet.

Overall one can comprehend the matryoshka's nature of mobile genetic element [12]: MGE with inter-genomic mobility, often smaller than MGE with inter-organisms mobility, can integrate within the latter. A famous example is the class 1 integrons, which is captured by a transposon, allowing the transposition within other transposons or plasmids (Figure 6) [80].

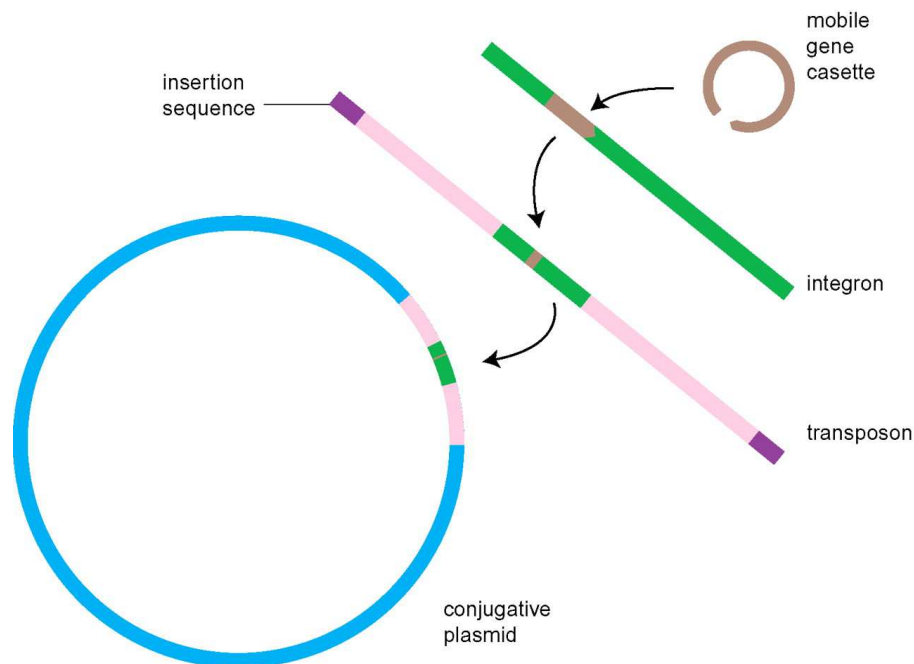


Figure 6: The matryoshka's nature of MGEs. A mobile integron-cassette is captured by an integron (in class 1 integrons, it is often an antibiotic resistance gene). This integron is then captured by a transposon allowing intra-genomic mobility. The transposon can insert into a conjugative plasmid, conferring to the whole system, an inter-organism mobility. Figure from [144].

### 3.2 Persistence of MGE

Upon transfer into a new host, a mobile genetic element encounters many obstacles [185]. First, different mechanisms controlling the infection of the bacteria are found and remind of the immune system because of their ability to distinguish between the self from the non-self. Today's famous example of defense system is the CRISPR-Cas system, but other systems exist like restriction modification (RMS) or DNA phosphorothioation (DND) systems [129]. Briefly, these systems aim at being able to distinguish the host DNA (self) from exogenous DNA (non-self) and degrade the latter. They are often compared to innate immunity (RMS, DND) or adaptive immunity systems (CRISPR). Their role as a way to strictly defend against HGT event has been discussed [186, 149].

Once this barrier is passed, the MGE must be stably maintained in the cell and in the population. To be maintained in a cell it has to endure similar events than the chromosome, which replicates and is partitioned between the two daughter cells. To do so, a MGE can either integrate into the chromosome or replicate and segregate independently. Mechanisms exist to compensate failure to be present in the two daughter cells, named post segregational killing (PSK) systems [205]. These mechanisms encode a pair of proteins or a protein and a RNA, whose one is toxic and the other is the antidote [76, 108]. The antidote is degraded

faster such that if its production ceases (if the MGE is lost), the toxic protein will kill the cell. However, PSK strategies are not ubiquitous among MGEs, and their role in maintenance has been debated [40, 39]. Finally, if the MGE brings no adaptive value to the cell, only a high rate of HGT will prevent its loss from the population.

This section will review the mechanisms of replication, partition and integration with an emphasis on plasmid replication and partition systems, since they represent the majority of elements encoding these functions and are the most relevant ones for the forthcoming analyses.

### 3.2.1 Replication

Three main types of replication mechanisms exist for plasmids. Namely, theta, strand-displacement and rolling-circle replication.

Theta replication is the type of replication used by the bacterial chromosome. Both strands are replicated at the same time forming a replication bubble recalling the greek letter  $\theta$  when seen under electron microscopy [122]. The various theta replication systems contain four main elements that are encoded on the plasmid. A replication initiator protein, often referred to as Rep, and at least three important motifs or regions, namely iterons, DnaA boxes and AT-rich region [114]. Other motifs can be present like integration host factor (IHF) binding sites, GATC motifs, or specific protein binding sites, but their role in replication is not essential or unclear [114]. The three essential motifs are part of the origin of replication, and are recognized by the replication initiator protein and host factors. The iterons are short direct repeated sequences whose number is variable. The Rep protein will bind to them, which destabilizes locally the pairing between the two complementary strands of the DNA. This action is often followed by the binding of a host factor — DnaA — on DnaA-boxes near the iterons. Similarly to iterons, the number of DnaA-boxes varies [114]. The destabilization of the AT-rich region by the fixation of Rep and DnaA nearby will allow the host helicase to further separate the strand and permit the host DNA polymerase to start the replication. The interaction of the Rep protein with the host proteins may be not stable enough to trigger the replication, leading to the loss of the plasmid. This dependency on host factor for theta replication limits the host range in which a plasmid can replicate. Replication of the plasmid also influences its copy number in a cell, which has a cost for the bacterium. Regulation of the replication controls the copy number and thus can decrease the cost. It involves physical inhibition of the replication (“handcuffing” mechanism), anti-sense RNA inhibition, or negative feedback loop on the Rep promoter by the Rep itself [114, 122].

During strand-displacement replication, the two strands are replicated continuously [122]. This mechanism involves a different initiator Rep protein, but also specific helicases and primases encoded by the plasmids. Consequently, plasmids replicated with this mechanism do not depend on host factors, which broadens their host range [122].

Rolling circle replication plasmids encode another type of Rep protein. They are usually encoded on small plasmids, from as small as less than 1kb to up to about 30kb [165]. The Rep protein nicks the DNA at the double-strand origin of replication, which leaves a space for the host replisome to process the leading strand [165]. The leading strand is replicated and the Rep protein will free the new double stranded plasmid and an intermediate single stranded DNA, corresponding to the parental lagging strand. The latter is then replicated using the host machinery from a single strand origin of replication that adopt a particular secondary structure. This mechanism of replication is less dependent of host proteins to initiate the replication than theta replication and hence is thought to enable broader host range [165].

### 3.2.2 Partition system

This section exists as a Wikipedia article, available here: [https://en.wikipedia.org/wiki/Plasmid\\_partition\\_system](https://en.wikipedia.org/wiki/Plasmid_partition_system). I created this page (pseudonym: Jeanrjc) while doing the bibliographical work on partition systems. The present work includes the last version available online at the time of writing. My work represents about 68% of the text added so far. Interestingly, the process behind creating a wikipedia page has some similarities with a peer-review process in academia. Indeed, when creating a new Wikipedia page, the article goes through a process called “Article for creation”. Wikipedia’s peer will then review the article and assess whether it fits the format of a wikipedia article. This article has been visited about 500 times per month for the last year. The article has been formatted by Wikipedia’s pdf export tool.

# Plasmid partition system

A **plasmid partition system** is a mechanism that assures the stable transmission of plasmids during bacterial cell division. Each plasmid has its independent replication system which controls the number of copies of the plasmid in a cell. The higher the copy number is, the more likely the two daughter cells will contain the plasmid. Generally, each molecule of plasmid diffuses randomly, so the probability of having a plasmid-less daughter cell is  $2^{1-N}$ , where N is the number of copies. For instance, if there are 2 copies of a plasmid in a cell, there is a 50% chance of having one plasmid-less daughter cell. However, high-copy number plasmids have a cost for the hosting cell. This metabolic burden is lower for low-copy plasmids, but those have a higher probability of plasmid loss after a few generations. To control vertical transmission of plasmids, in addition to controlled-replication systems, bacterial plasmids use different maintenance strategies, such as multimer resolution systems, post-segregational killing systems (addiction modules), and partition systems.<sup>[1]</sup>

## 1 General properties of partition systems

Plasmid copies are paired around a centromere-like site and then separated in the two daughter cell. Partition systems involve three elements, organized in an autoregulated operon.<sup>[2]</sup>

- centromere-like DNA site
- Centromere binding protein (CBP)
- Motor protein

Centromere-like DNA site is required in *cis* for plasmid stability. It often contains one or more inverted repeats which are recognized by multiple CBPs. This forms a nucleoprotein complex termed partition complex. This complex recruits the motor protein, which is a nucleotide triphosphatase (NTPase). The NTPase uses energy from NTP binding and hydrolysis to directly or indirectly move and attach plasmids to specific host location (e.g. opposite bacterial cell poles).

The partition systems are divided in four types, based primarily on the type of NTPases:<sup>[3]</sup>

- Type I : Walker type P-loop ATPase

- Type II : Actin-like ATPase
- Type III : tubulin-like GTPase
- Type IV : No NTPase

## 2 Type I partition system

This system is also used by most bacteria for chromosome segregation.<sup>[3]</sup> Type I partition systems are composed of an ATPase which contains Walker motifs and a CBP which is structurally distinct in type Ia and Ib. ATPases and CBP from type Ia are longer than the ones from type Ib, but both CBPs contain an arginine finger in their N-terminal part.<sup>[1][4]</sup> ParA proteins from different plasmids and bacterial species show 25 to 30% of sequence identity to the protein ParA of the plasmid P1.<sup>[5]</sup> The partition of type I system uses a “diffusion-ratchet” mechanism. This mechanism works as follows:<sup>[6]</sup>

1. Dimers of ParA-ATP dynamically bind to nucleoid DNA<sup>[7][8]</sup>
2. ParB bound to *parS* stimulates the release of ParA from the nucleoid region surrounding the plasmid<sup>[9]</sup>
3. The plasmid then chases the resulting ParA gradient on the perimeter of the ParA depleted region of the nucleoid
4. The ParA that was released from the nucleoid behind the plasmid’s movement redistributes to other regions of the nucleoid after a delay<sup>[10]</sup>
5. After plasmid replication, the sister copies segregate to opposite cell halves as they chase ParA on the nucleoid in opposite directions

It should be noted that there are likely to be differences in the details of type I mechanisms.<sup>[4]</sup>

Type I partition has been mathematically modelled with variations in the mechanism described above.<sup>[11][12][13]</sup>

### 2.1 Type Ia

The CBP of this type consists in three domains:<sup>[4]</sup>

- N-terminal NTPase binding domain
- Central Helix-Turn-Helix (HTH) domain
- C-terminal dimer-domain

## 2.2 Type Ib

The CBP of this type, also known as *parG* is composed of:<sup>[4]</sup>

- N-terminal NTPase binding domain
- Ribon-Helix-Helix (RHH) domain

For this type, the *parS* site is called *parC*.

## 3 Type II partition system

This system is the best understood of the plasmid partition system.<sup>[4]</sup> It is composed of an actin-like ATPase, ParM, and a CBP called ParR. The centromere like site, *parC* contains two sets of five 11 base pair direct repeats separated by the *parMR* promoter. The amino-acid sequence identity can go down to 15% between ParM and other actin-like ATPase.<sup>[5][14]</sup>

The mechanism of partition involved here is a pushing mechanism:<sup>[15]</sup>

1. ParR binds to *parC* and pairs plasmids which form a nucleoprotein complex, or partition complex
2. The partition complex serves as nucleation point for the polymerization of ParM; ParM-ATP complex inserts at this point and push plasmids apart
3. The insertion leads to hydrolysis of ParM-ATP complex, leading to depolymerization of the filament
4. At cell division, plasmids copies are at each cell extremity, and will end up in future daughter cell

The filament of ParM is regulated by the polymerization allowed by the presence the partition complex (ParR-*parC*), and by the depolymerization controlled by the ATPase activity of ParM.

## 4 Type III partition system

The type III partition system is the most recently discovered partition system. It is composed of tubulin-like GTPase termed TubZ, and the CBP is termed TubR. Amino-acid sequence identity can go down to 21% for TubZ proteins.<sup>[5]</sup>

The mechanism is similar to a treadmill mechanism:<sup>[16]</sup>

1. Multiple TubR dimer binds to the centromere-like region *stbDRs* of the plasmids.
2. Contact between TubR and filament of treadmilling TubZ polymer. TubZ subunits are lost from the - end and are added to the + end.

3. TubR-plasmid complex is pulled along the growing polymer until it reaches the cell pole.
4. Interaction with membrane is likely to trigger the release of the plasmid.

The net result being transport of partition complex to the cell pole.

## 5 Other partition systems

### 5.1 R388 partition system

The partition system of the plasmid R388 has been found within the *stb* operon. This operon is composed of three genes, *stbA*, *stbB* and *stbC*.<sup>[17]</sup>

- StbA protein is a DNA-binding protein (identical to ParM) and is strictly required for the stability and intracellular positioning of plasmid R388 in *E. coli*. StbA binds a *cis*-acting sequence, the *stbDRs*.

The StbA-*stbDRs* complex may be used to pair plasmid the host chromosome, using indirectly the bacterial partitioning system.

- StbB protein has a Walker-type ATPase motif (ParA-like), it favors for conjugation but is not required for plasmid stability over generations.
- StbC is an orphan protein of unknown function. StbC doesn't seem to be implicated in either partitioning or conjugation.













StbA and StbB have opposite but connected effect related to conjugation.

This system has been proposed to be the type IV partition system.<sup>[18]</sup> It is thought to be a derivative of the type I partition system, given the similar operon organization. This system represents the first evidence for a mechanistic interplay between plasmid segregation and conjugation processes.<sup>[18]</sup>

### 5.2 pSK1 partition system (reviewed in [1])

pSK1 is a plasmid from *Staphylococcus aureus*. This plasmid has a partition system determined by a single gene, *par*, previously known as *orf245*. This gene does not effect the plasmid copy number nor the grow rate (excluding its implication in a post-segregational killing system). A centromere-like binding sequence is present upstream of the *par* gene, and is composed of seven direct repeats and one inverted repeat.

## 6 References

- [1] Dmowski M, Jagura-Burdzy G (2013). “Active stable maintenance functions in low copy-number plasmids of Gram-positive bacteria I. Partition systems” (PDF). *Polish Journal of Microbiology / Polskie Towarzystwo Mikrobiologów = the Polish Society of Microbiologists*. **62** (1): 3–16. PMID 23829072.
- [2] Friedman SA, Austin SJ (1988). “The P1 plasmid-partition system synthesizes two essential proteins from an autoregulated operon”. *Plasmid*. **19** (2): 103–12. PMID 3420178. doi:10.1016/0147-619X(88)90049-2.
- [3] Gerdes K, Møller-Jensen J, Bugge Jensen R (2000). “Plasmid and chromosome partitioning: surprises from phylogeny”. *Molecular Microbiology*. **37** (3): 455–66. PMID 10931339. doi:10.1046/j.1365-2958.2000.01975.x.
- [4] Schumacher MA (2012). “Bacterial plasmid partition machinery: a minimalist approach to survival”. *Current Opinion in Structural Biology*. **22** (1): 72–9. PMC 4824291 . PMID 22153351. doi:10.1016/j.sbi.2011.11.001.
- [5] Chen Y, Erickson HP (2008). “In vitro assembly studies of FtsZ/tubulin-like proteins (TubZ) from Bacillus plasmids: evidence for a capping mechanism”. *The Journal of Biological Chemistry*. **283** (13): 8102–9. PMC 2276378 . PMID 18198178. doi:10.1074/jbc.M709163200.
- [6] Badrinarayanan, Anjana; Le, Tung B. K.; Laub, Michael T. (2015-11-13). “Bacterial Chromosome Organization and Segregation”. *Annual Review of Cell and Developmental Biology*. **31**: 171–199. ISSN 1530-8995. PMC 4706359 . PMID 26566111. doi:10.1146/annurev-cellbio-100814-125211.
- [7] Hwang, Ling Chin; Vecchiarelli, Anthony G.; Han, Yong-Woon; Mizuuchi, Michiyo; Harada, Yoshie; Funnell, Barbara E.; Mizuuchi, Kiyoshi (2013-05-02). “ParA-mediated plasmid partition driven by protein pattern self-organization”. *The EMBO Journal*. **32** (9): 1238–1249. ISSN 1460-2075. PMC 3642677 . PMID 23443047. doi:10.1038/emboj.2013.34.
- [8] Vecchiarelli, Anthony G.; Hwang, Ling Chin; Mizuuchi, Kiyoshi (2013-04-09). “Cell-free study of F plasmid partition provides evidence for cargo transport by a diffusion-ratchet mechanism”. *Proceedings of the National Academy of Sciences of the United States of America*. **110** (15): E1390–1397. ISSN 1091-6490. PMC 3625265 . PMID 23479605. doi:10.1073/pnas.1302745110.
- [9] Vecchiarelli, Anthony G.; Neuman, Keir C.; Mizuuchi, Kiyoshi (2014-04-01). “A propagating ATPase gradient drives transport of surface-confined cellular cargo”. *Proceedings of the National Academy of Sciences of the United States of America*. **111** (13): 4880–4885. ISSN 1091-6490. PMC 3977271 . PMID 24567408. doi:10.1073/pnas.1401025111.
- [10] Vecchiarelli, Anthony G.; Han, Yong-Woon; Tan, Xin; Mizuuchi, Michiyo; Ghirlando, Rodolfo; Biertümpfel, Christian; Funnell, Barbara E.; Mizuuchi, Kiyoshi (2010-10-01). “ATP control of dynamic P1 ParA-DNA interactions: a key role for the nucleoid in plasmid partition”. *Molecular Microbiology*. **78** (1): 78–91. ISSN 1365-2958. PMC 2950902 . PMID 20659294. doi:10.1111/j.1365-2958.2010.07314.x.
- [11] Hu, Longhua; Vecchiarelli, Anthony G.; Mizuuchi, Kiyoshi; Neuman, Keir C.; Liu, Jian (2015-12-08). “Directed and persistent movement arises from mechanochemistry of the ParA/ParB system”. *Proceedings of the National Academy of Sciences of the United States of America*. **112** (51): E7055–64. ISSN 1091-6490. PMC 4697391 . PMID 26647183. doi:10.1073/pnas.1505147112.
- [12] Vecchiarelli, Anthony G.; Seol, Yeonee; Neuman, Keir C.; Mizuuchi, Kiyoshi (2014-01-01). “A moving ParA gradient on the nucleoid directs subcellular cargo transport via a chemophoresis force”. *Bioarchitecture*. **4** (4–5): 154–159. ISSN 1949-100X. PMID 25759913. doi:10.4161/19490992.2014.987581.
- [13] Ietswaart, Robert; Szardenings, Florian; Gerdes, Kenn; Howard, Martin (2014-12-01). “Competing ParA structures space bacterial plasmids equally over the nucleoid”. *PLOS Computational Biology*. **10** (12): e1004009. ISSN 1553-7358. PMC 4270457 . PMID 25521716. doi:10.1371/journal.pcbi.1004009.
- [14] Gunning PW, Ghoshdastider U, Whitaker S, Popp D, Robinson RC (2015). “The evolution of compositionally and functionally distinct actin filaments”. *J Cell Sci*. **128** (11): 2009–19. PMID 25788699. doi:10.1242/jcs.165563.
- [15] Møller-Jensen J, Borch J, Dam M, Jensen RB, Roepstorff P, Gerdes K (2003). “Bacterial mitosis: ParM of plasmid R1 moves plasmid DNA by an actin-like insertional polymerization mechanism”. *Molecular Cell*. **12** (6): 1477–87. PMID 14690601. doi:10.1016/S1097-2765(03)00451-9.
- [16] Ni L, Xu W, Kumaraswami M, Schumacher MA (2010). “Plasmid protein TubR uses a distinct mode of HTH-DNA binding and recruits the prokaryotic tubulin homolog TubZ to effect DNA partition”. *Proceedings of the National Academy of Sciences of the United States of America*. **107** (26): 11763–8. PMC 2900659 . PMID 20534443. doi:10.1073/pnas.1003817107.
- [17] Guynet C, Cuevas A, Moncalián G, de la Cruz F (2011). “The stb operon balances the requirements for vegetative stability and conjugative transfer of plasmid R388”. *PLoS Genetics*. **7** (5): e1002073. PMC 3098194 . PMID 21625564. doi:10.1371/journal.pgen.1002073.
- [18] Guynet C, de la Cruz F (2011). “Plasmid segregation without partition”. *Mobile Genetic Elements*. **1** (3): 236–241. PMC 3271553 . PMID 22312593. doi:10.4161/mge.1.3.18229.



---

## 7 Text and image sources, contributors, and licenses

### 7.1 Text

- **Plasmid partition system** *Source:* [https://en.wikipedia.org/wiki/Plasmid\\_partition\\_system?oldid=803668722](https://en.wikipedia.org/wiki/Plasmid_partition_system?oldid=803668722) *Contributors:* R. S. Shaw, Rjwilmsi, PrimeHunter, Levskaya, Headbomb, DGG, Boghog, Yobot, L235, Dcirovic, Hewhoamareismyself, Sir Electron, Citation-CleanerBot, JSHuisman, Jeanrjc, ToonLucas22, CellfOrganized, Quinton Feldberg, Harsh Pinjani India and Anonymous: 2

### 7.2 Images

- **File:Lock-green.svg** *Source:* <https://upload.wikimedia.org/wikipedia/commons/6/65/Lock-green.svg> *License:* CC0 *Contributors:* en:File:Free-to-read\_lock\_75.svg *Original artist:* User:Trappist the monk

### 7.3 Content license

- Creative Commons Attribution-Share Alike 3.0

### 3.2.3 Integration

An alternative way for an exogenous element to remain in a cell lineage, is to integrate into the host's chromosome. The latter already encodes replication and partition machineries needed for its segregation into the daughter cells. There are two types of site specific recombinases, the tyrosine-recombinase and the serine-recombinase, named after their catalytic residue used for DNA cleavage. During recombination, both need two dsDNA partners having specific recombination sites. The recombinases form a complex which will cleave the DNA partner and join them after strand exchanges [85]. The outcome of this recombination event depends on the relative positions of the recombination sites, and can lead to either integration, excision or inversion (Figure 7).

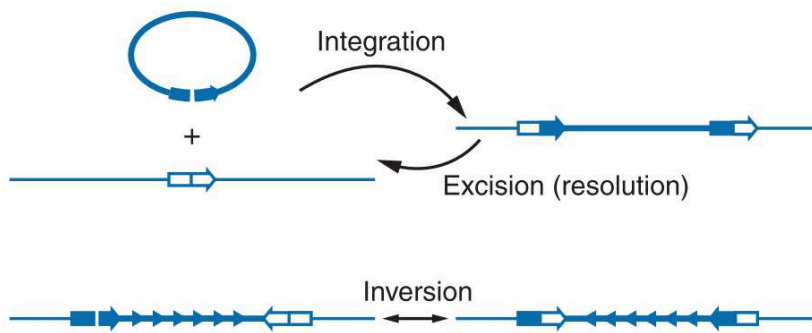


Figure 7: Different possible outcomes of site-specific recombination. The outcome depends on the relative orientation of the recombination sites (large arrows). Figure from [85].

The difference between tyrosine and serine recombinases is in the mechanism to achieve the recombination event. Briefly, for the tyrosine-recombinase, there is the formation of an intermediate Holliday junction which is the result of the cleavage of one strand of each partner, followed by their junction before the cleavage of the second strand [91] (Figure 8). For the serine-recombinase, both strands are cut and the strand exchange happens afterwards [160]. Tyrosine recombinases are widespread among bacteria, archaea and phages and can be found in eukaryotes [85]. In bacteria, they can be associated with the bacterial chromosome maintenance, for instance, XerCD are essential proteins allowing the resolution of chromosome dimers that can occur after replication [16]. Other tyrosine-recombinases are associated with MGE, notably with prophages [145] and Integrative and Conjugative elements (ICEs) [21], but also with integrons [26]. Tyrosine recombinases are genetically diverse and form a vast group of recombinases. The different types of recombinase among which integrases, resolvases, invertases and transposases do not define different subgroups of tyrosine recombinases [145, 177]. In contrast, serine-recombinases form three large monophyletic groups corresponding to their specific function among which resolvases/invertases, IS-like transposases and large serine recombinases (for integration) [177]. Serine recombinases are found associated with mobile genetic elements like phages or transposons [85].

Integration of MGEs can also be through transposition mechanisms which is mainly used

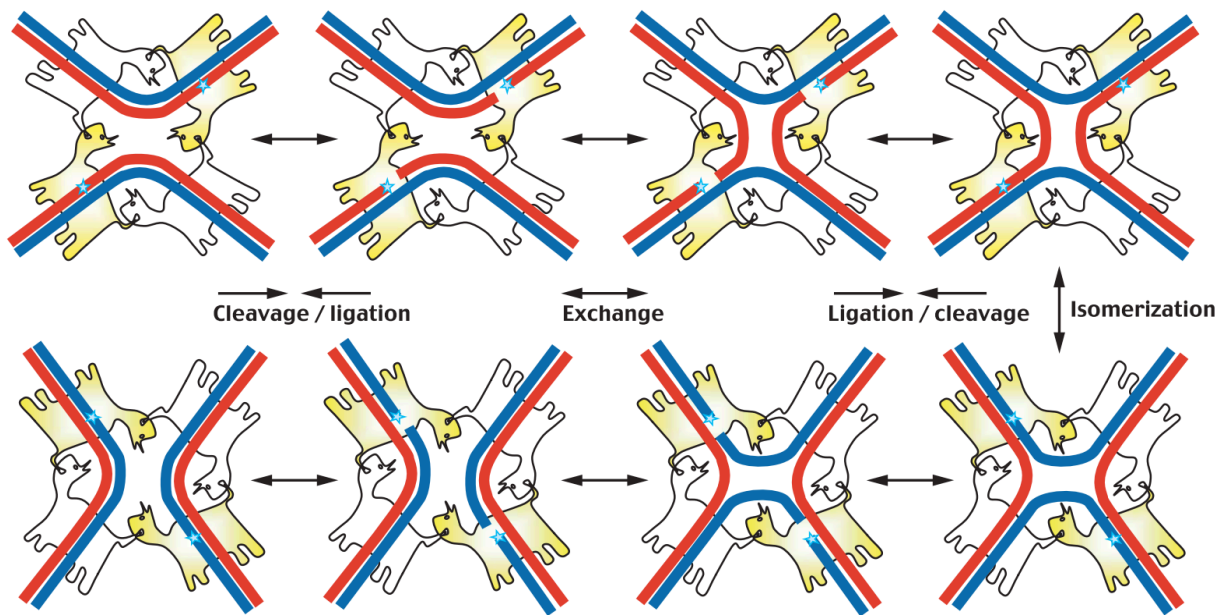


Figure 8: Schema of a tyrosine-recombinase strand exchange mechanism. The recombination complex is composed of two dsDNA bound to four recombinases. Yellow integrases are active, and blue stars represent their catalytic center. A first strand (red) is cleaved, exchanged and ligated with the other partner. It forms a Holliday junction (rightmost two panels). After isomerization, the other two recombinases proceed with the remaining strands. Figure from [85].

by transposable elements. The specificity of the site is often lower than with site-specific recombinase [172]. There are “copy-paste”, “cut-and-paste” and “peel-and-paste” mechanisms, which depend on the transposase [172]. Other MGEs than transposable elements like ICEs, have been found to be integrated with DDE-transposases [53].

Importantly, integration in the chromosome can disrupt essential genes or its organization. Thus, MGE evolved different strategies to minimize the cost of their integration. For instance, integrative MGE tend to integrate at the edges of operons to prevent the disruption of their expression [155]. Likewise, phages’ site specific recombinases target few conserved sites to avoid highly transcribed regions [18]. In addition, most genes are repressed in certain integrated elements, be it an ICE [109] or a prophage [156], which further reduces their fitness cost.

## 4 Deciphering the bacterial chromosome

One of the goal of bioinformatics is to succeed in understanding the information encoded in the genome. For example, the genetic code was discovered in the 1960s [42, 142], and this knowledge was later applied to detect the presence of protein-coding sequences in genomes. In general, a start codon, and  $N$  nucleotides later, a stop codon, with  $N$  being a multiple of 3. This can be implemented easily and allows the rapid identification of the potential number of proteins in a genome. Of course this is a simplistic method to find coding sequences, and more powerful ones exist nowadays [106, 48]. Once we identified proteins, a much more complicated problem in bacterial genomics is to know what function they encode.

In experimental genetics, one can infer protein's function without prior knowledge on the system. In bioinformatics, to infer the function of a protein, there is no other solution than that of comparing to something we know. For instance, one can compare a protein to another protein sequence for which we have experimental data demonstrating its function. It is also possible to infer a possible function of a gene given the presence of a known genetic context [151]. Thus, all the existing methods boil down to implementing the best algorithm for “*comparing to something we know*”. The most famous one being the BLAST algorithm [6] which allows the rapid detection of a given sequence in a database, allowing the transfer of annotations from very similar sequences. The problem with BLAST, especially when doing evolutionary genomics, is that protein sequences may evolve fast and remain functional homologs while lacking any significant sequence similarity, and BLAST cannot identify homologous proteins that are too divergent.

The problem is more complex for the detection of RNA genes since they cannot be translated into amino-acid sequences which are more conserved due the degeneracy of the genetic code. RNAs adopt structure and the interactions governing the secondary structures are often stronger than those determining the tertiary structure [131]. It has been estimated that for most RNAs, about half of their residues engage in base-pairing [54, 131]. Consequently, their primary sequence is not important as long as the secondary structure is conserved. Hence, RNA genes are often very poorly conserved, albeit some regions might be highly conserved like it is the case for the 16S ribosomal RNA [202, 35].

During my PhD I have used a wide range of tools and methods like protein annotation and alignment, phylogenetic tree reconstruction, pan- and core-genome construction, among others, which constitute the toolbox of the bioinformatician, and allow to perform a wide range of analysis that will be encountered in this thesis. In the following section I will describe the basics of two of the most powerful methods existing to annotate proteins and RNA in genomes. These methods are particularly elegant because they relate on the theoretical framework developed by linguists in the 1950s to study natural languages. They have been also used extensively during this PhD.

## 4.1 Formal grammars

**Colorless green ideas sleep furiously.** This grammatically correct sentence exists since 1955 when Noam Chomsky developed his theory on the description of languages [31]. It is believed that this sentence does not exist in any context not referring to Chomsky’s work. Chomsky was interested in understanding how a human can determine whether a given sentence is grammatically correct or not, even when one encounters this sequence of word for the first time, and especially given that the sentence is meaningless. How human parse sentences, and how this could be modelled with a mathematical framework? Hence, Chomsky provides a theoretical framework for linguistic analysis of languages, where a language is the set of sentences (finite or infinite), constructed from a finite alphabet of symbols. The grammar of this language is defined as “a device of some sort that (it) produces all of the strings that are sentences of [the language] and only these” [32]. To the question: “is this sentence grammatically correct?”, which could be rephrased as “does the language contain this sentence?”, linguistic theory offers the possibility to ask instead: “Can the grammar generate this sequence?” [58]. A probabilistic model of the grammar allows to quantify this question by providing a probability as an answer to this question. Chomsky later defined four types of formal grammars with increasing sets of restrictions, known as the Chomsky hierarchy [33]. The more constrained grammar is the easiest to parse, but is bad at describing a natural language. The two most constrained grammars found applications in deciphering biological data, namely regular grammar and context-free grammar. Interestingly, many bioinformatics methods based on sequence alignments are using implicitly regular grammars [58].

Formal grammars are based on productions rules, which generates two types of symbols: terminal symbols (*e.g.* a “word” in a sentence) and non-terminal symbols (*e.g.* the “part-of-speech” in a sentence — noun, verb, verb phrase, etc. . . ). Non-terminal symbols can be replaced by terminal symbols (*e.g.* **Noun** → **Cat**), or another non-terminal symbols (*e.g.* **Prepositional phrase** → **Preposition** + **Noun**). The production rules have two sides (separated by the arrow in the previous example) and are read from left to right. The left part always contains at least a non-terminal symbol. An example of a regular grammar is shown below:

$$S \rightarrow aS \mid bS \mid \epsilon \quad (1)$$

$S$  is a non-terminal symbol that produces strings of  $a$  and  $b$  until it stops ( $\epsilon$ ) with equal probability among the three outputs<sup>2</sup>. This grammar has an infinite number of sentences, and one derivation from this grammar is shown below:

$$S \Rightarrow aS \Rightarrow abS \Rightarrow abbS \Rightarrow abbaS \Rightarrow abba. \quad (2)$$

---

<sup>2</sup>One could read the above rule as “ $S$  generates  $aS$  or  $bS$  or stops”

Regular grammars can only be of the form  $W \rightarrow aW$  or  $W \rightarrow a$ , with  $W$  being any non-terminal symbol and  $a$  any terminal symbol. Context-free grammar can contain any combination of symbols on the right-hand side of the production rule, like  $W \rightarrow aWa$ . An example of a context-free grammar is shown below:

$$S \rightarrow aSa \mid bSb \mid c \quad (3)$$

A derivation of this grammar can be:

$$S \Rightarrow aSa \Rightarrow abSba \Rightarrow abaSaba \Rightarrow abacaba. \quad (4)$$

Thus, context-free grammar has the possibility to generate sequences from outside in, unlike regular grammar that can only generate from left to right. Hence, context-free grammar can generate sequences where it exists dependencies between distant positions, like it is the case for a palindromic sequence. It is important to note that regular grammar can generate palindromes, as seen in the first derivation (*abba*). Regular grammar will fail at generating only palindromes. Thus, one cannot distinguish a palindromic sequence from a non-palindromic sequence using probabilistic model of regular grammar [32, 58]. On the other hand, context-free grammar can generate palindromes efficiently, and their associated probabilistic model can distinguish a palindromic sequence from a non-palindromic sequence, because the probability of a palindrome being generated by a context-free grammar can be higher than a non-palindromic sequence. This has important consequences on the detection of RNA genes since they harbor palindromic-like sequences.

Hidden Markov Models and Covariance Models are probabilistic models for regular grammar and context-free grammar respectively, which are used to compute the probability of a sequence being generated by the corresponding grammar.

## 4.2 Hidden Markov Model profiles

Hidden Markov Model (HMM) profiles allow the detection of remote protein homologous [61]. They can model the fact that parts of a protein are more conserved while other parts are more prone to mutation, deletion or insertion. Taking this into account is the goal of profile based approaches [61]. HMM turned out to be the most efficient mathematical model for this approach [63]. HMM is a probabilistic model generating sequences. An HMM is composed of a finite number of states ( $\Leftrightarrow$  non-terminal symbol), each state can generate symbols ( $\Leftrightarrow$  terminal symbol), according to different probabilities, and after emission of a symbol, the state changes according to state transition probabilities [60]. The sequence is then generated by moving through the states (Figure 9, left). The model depicted in Figure 9, top-left, is the

graphical representation of a Markov Model, which happens to correspond to a graphical representation of the grammatical rules of a regular grammar seen in rule (1) above (although the state 1 would have been sufficient). The alphabet of symbols here is composed of  $\{a, b\}$ , and the language is the infinite set of symbol sequences (or sentences) generated by this grammar. The sentence  $aba$  (the observed symbol sequence, Figure 9, left) generated with states  $1 \rightarrow 1 \rightarrow 2$  could have been obtained with the states  $1 \rightarrow 2 \rightarrow 2$ , but with a different probability. This marks the difference between a Markov Model and a Hidden Markov Model. In the latter, the observer does not have access to the state that generated the observed symbol or character. The utility of using HMM resides in its ability to answer the question mentioned above: “Can a given HMM generate this observed sentence?”

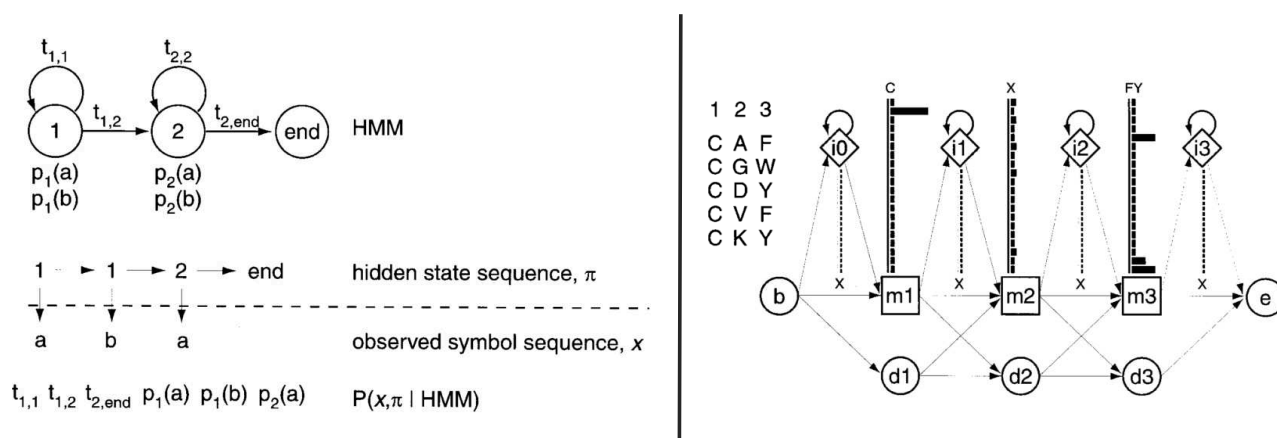


Figure 9: Schema of a toy HMM (Hidden Markov Model) and of a small HMM profile. On the left-hand side, a hidden Markov model. The Markov model is on the top. Circles represent states, and an arrow represent a transition from one state to another.  $t_{i,j}$  represent the probability of transition from state  $i$  to state  $j$ . Each state can generate (emit) characters, here  $a$  or  $b$  with different probabilities  $p_i$ , depending on the state. The model starts from left and continues until it reaches the *end* state.

On the right-hand side, a small HMM profile is depicted. It is parametrized with a multiple alignment of three amino-acids and five sequences. It exists three kinds of state in addition to begin ( $b$ ) and end ( $e$ ) states. There are the match states ( $m_j$ ), the insertion states ( $i_j$ ) and the deletion states ( $d_j$ ). Insertion and match states have different emission probabilities, parametrized after the alignment. Since there is no insertion in this alignment, all insertion states can emit any amino-acid with equal probability, likewise for the state  $m_2$  where each amino-acids is different in the alignment. In contrast, there is a high probability of generating a  $C$  on the first state  $m_1$  since there is a consensus on the alignment. The figure is adapted from [61].

The notion of profile refers to the application of HMM to biological sequence data, where the question now becomes, for instance : “Can a given HMM profile generate this observed protein sequence?” The transition and emission probabilities parameters are based on a multiple sequence alignment. Hence, to build an HMM profile, the first step is to build a multiple

alignment from a set of homologous proteins with a given function. This will define the grammar for the language, which corresponds to the possible protein space. All the sentences of this language correspond to variants or likely variants of the protein. Consequently, from the HMM profile's perspective, homologous are the set of proteins whose sequences could have been generated by it with sufficient probability. When searching on entire databases, this probability is turned into a score. The E-value corresponds to the expected number of hits with a similar score on a random database without real homologous and is used to assess the significance of a hit.

This framework is implemented in the **HMMER** suite [63], which allows building and uses HMM profiles. During this thesis, most of the protein annotations are done with these method and tool.

### 4.3 Covariance Models

Covariance Models (CM) allow the detection of any sequences having a palindromic-like structure. RNA genes typically fall into that category. RNAs adopt secondary structures, important for their function [84]. Secondary structures of single stranded DNA or RNA exist because the folded molecule is stabilized by Watson-Crick base-pairing, or with less stable non-canonical pairs, the most common being the wobbles G-U pairs [58]. This base-pairing implies the presence of a palindrom in the primary sequence and thus in the DNA sequence. Palindromic sequences in DNA are different from text palindromes, as the notion of backward reading is different. A DNA palindrome is read similarly when reading it on the direct strand or on the complementary strand. For instance **ATTA** is a palindromic word and not a DNA palindrome, but **ATAT** is a DNA palindrome ( $\begin{matrix} \overrightarrow{\text{ATAT}} \\ \overleftarrow{\text{TATA}} \end{matrix}$ ). There are dependencies in a palindromic sequence: if one position mutates, the corresponding palindromic position in the sequence has to mutate to maintain the palindrome and the secondary structure. There will be covariation of these mutations. Covariance Models are to context-free grammar what HMM profiles are to regular grammar [84]. CM can model both the sequence and the dependencies of some positions in the sequence. The probabilistic models of context-free grammars are called stochastic context-free grammars (SCFG). They can estimate the probability for a sequence to have been generated by a context-free grammar. A CM is a SCFG parametrized with a multiple sequence alignment of RNAs on their consensus secondary structure. Covariance Models are good at finding RNA sequences in genomes because SCFG add informations related to the covariation of distant positions that the HMM cannot apprehend (see Figure 10). This gain of information is proportional to the E-value for high scores [62] ( $Evalue \propto 2^{-x}$ , with  $x$  being the bit-score), such that approximately, every three bits of gained information improves the E-value by a factor of ten [141].

This framework is implemented in the **INFERNAL** suite [140], which allows building and uses



Covariance Models. CM are used in chapter 2 for the detection of the *attC* recombination sites. Their DNA can adopt a secondary structure by extruding from the double stranded DNA or during replication when they are transiently single stranded. Although INFERNAL is built to detect RNA genes, we have used it here to detect DNA recombination sites.

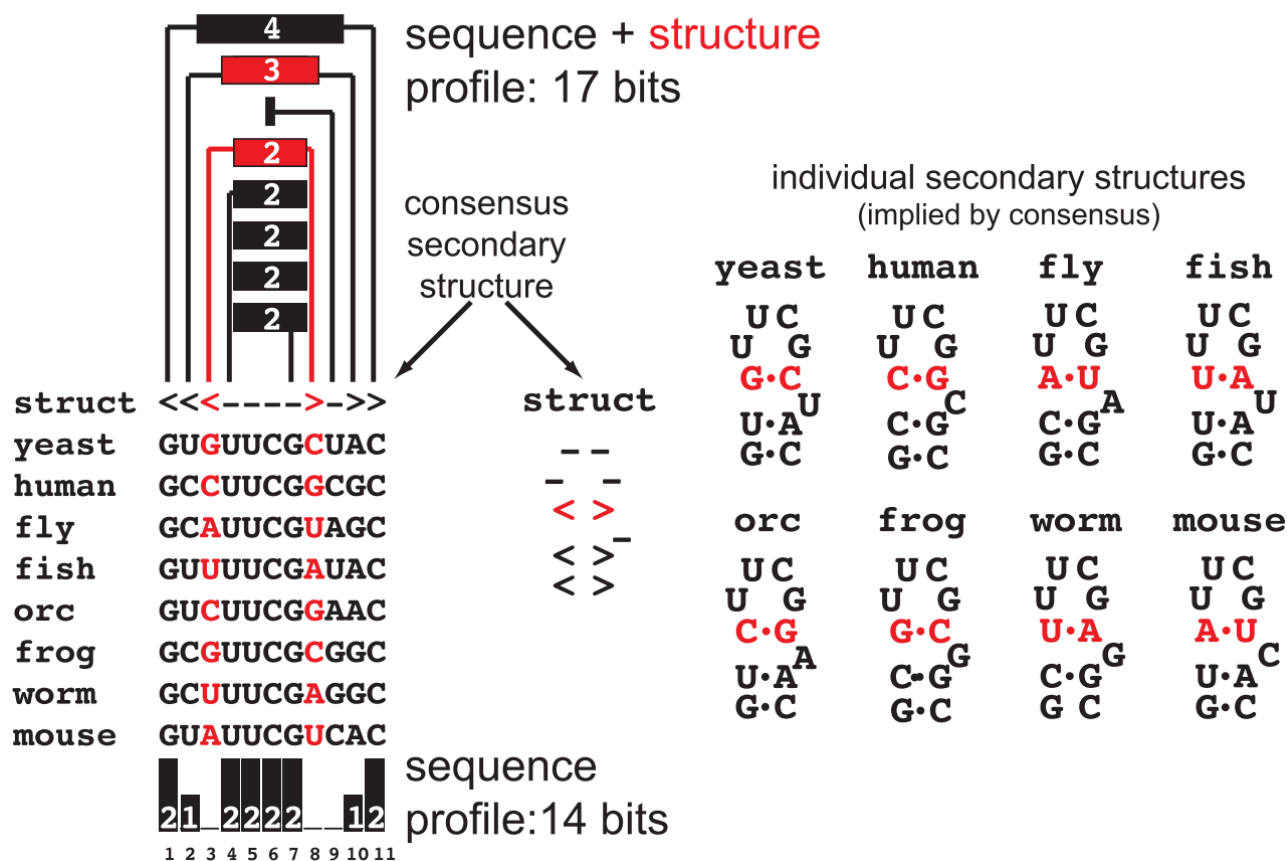


Figure 10: Covariance model example. On the left is depicted a multiple alignment of RNA sequences that are used to build a HMM profile. The amount of information (bits score) that the HMM profile has for each position is depicted beneath the alignment. The information here is  $\log_2\left(\frac{p_O}{p_E}\right)$ , where  $p_O$  is the observed probability (or frequency) of a residue, and  $p_E$  is the expected probability of a residue, here  $\frac{1}{4}$ . On top of it, there is the consensus secondary structure (**struct**) of the sequence where brackets indicate pairing in the secondary structure, as illustrated on the right side of the alignment. The amount of information using a Covariance Model is shown above the secondary structure consensus sequence, in red. The Covariance Model allows to gain 3 bits of information, which improves the E-value by a ten-fold factor. It takes into account both the sequence and the structure. Note that if the position is well conserved (*e.g.* positions 1 and 11), there is no gain of information with the secondary structure. On the right-hand side of the figure is shown the different sequences used in the alignment. Figure from [141].

## Outline

The following sections will focus on the work I have done during the last four years. Each chapter introduces in depth the corresponding element by describing the state of the art in this field. Then I present the work in the form of scientific papers, either in preparation, submitted or published. The chapter 1 focuses on conjugative systems. I present a method to detect and delimit Integrative and Conjugative Elements (ICEs) in bacterial genomes, followed by the analysis of the output of this method. I finally present an analysis of both types of conjugative elements, the integrative and extra-chromosomal one, and try to understand the whys and wherefores of this dichotomy. The chapter 2 concentrates on integrons. The tool and the analysis of the results are merged in a same paper. An other analysis made in collaboration with experimental biologists illustrates the interest of this tool in providing complementary information to experimental data and hypothesis.

**Part II**  
**Contributions**



# Chapter 1

## Conjugation

### 1.1 Introduction

#### 1.1.1 Background

Together with transduction and transformation, conjugation is one of the major mechanisms of horizontal gene transfer [178]. Unlike transduction or transformation, an event of conjugation is quantitatively more important because it transfers large amount of genetic information [24, 92]. It can occur between distant lineages [144], up to *eukaryotes* for instance [75]. Conjugation is a property of the eponym genetic element, and not a property of the bacteria. Conjugative elements have two distinct forms, integrated into the chromosome or extrachromosomal, called respectively Integrative and Conjugative Elements (ICEs) and Conjugative Plasmids (CPs). For a long time, the two forms have been treated separately, although both were discovered in the 1950s [120].

Integrative elements were at the time named episome (“agents with traffic in and out of the chromosome”) [107, 120]. But because of the difficulty to demonstrate the chromosomal location of episomes, people used the term episome instead of plasmid, despite the lack of evidence of their integration [120]. The term episome was then proposed to be abandoned to avoid further confusion [99]. Since then, the term episome has rarely been used, and researchers focused mainly on conjugative plasmids for practical reasons (they are easier to detect and work with). Thus, CPs are much more studied for historical reasons, but also due to their role in the spread of antibiotic resistance and virulence genes and their extensive use as a powerful genetic engineering tool for molecular biology and biotechnologies.

It is only in the 1980s that Integrative and Conjugative Elements (ICEs) were first described. They were found associated with transposons and thus named conjugative transposons [70]. They gained in interest because of their role in the spread of antibiotic resistance genes. The term ICE was coined only recently, in 2002 [25] with the objective of unifying the different terms

used since the 1980s (*e.g.* conjugative transposon, integrative plasmids, ...) used to refer to the same type of element “*that excise[s] by site-specific recombination into a circular form, self-transfer[s] by conjugation and integrate[s] into the host genome, whatever the specificity and the mechanism of integration and conjugation*” [25]. The evolutionary importance of ICEs was highlighted by the fact that they were actually more numerous than conjugative plasmids [90], despite the difference of coverage in the literature. Recently, it has been shown that some ICEs had plasmid-like properties like replication or partition [121, 29], upon which it was suggested that ICEs and CPs could be more similar than anticipated. Additionally, it has been suggested that ICEs and CPs might be two faces shown by a very similar type of element [90], where sufficient selective pressure for the presence of ICEs would lead to the conversion of a CP into an ICE, and vice versa. However, the question of the evolutionary forces driving this possible interconversion remains.

### 1.1.2 Diversity

**Diversity of conjugative systems.** There are different types of conjugation machinery, and part of the diversity can be explained by the different taxa in which they are found. Notably, bacteria have different types of cell envelopes, classically split in two main groups, the diderm (formerly gram negative) and the monoderm (formerly gram positive). The former has two cell membranes with a thin cell wall of peptidoglycan in between, while the latter has one cell membrane and a thicker cell wall. It has been shown that only one protein has a recognizable homolog among all the known conjugative systems, the VirB4 protein. This protein forms a monophyletic group among the AAA+ ATPases involved in DNA translocation, a large class of proteins, which has other members found in the conjugative system, but also in key cellular processes like cell division [88]. VirB4 is divided in eight different monophyletic groups forming eight different types of mating pair formation systems (MPF). The MPF types are associated with bacterial clades, among which four are mainly in Proteobacteria (MPF<sub>T</sub>, MPF<sub>G</sub>, MPF<sub>I</sub>, MPF<sub>F</sub>), two in monoderm bacteria, including mainly Firmicutes and Actinobacteria (MPF<sub>FA</sub>, MPF<sub>FATA</sub>), one in Cyanobacteria (MPF<sub>C</sub>) and one in Bacteroidetes (MPF<sub>B</sub>). These groups represent highly sampled bacterial clades, but other poorly sample clades can also contain these systems, for instance MPF<sub>T</sub> are found in Chlorobi or Acidobacteria, which are diderm bacteria like Proteobacteria [88]. Importantly, different MPF types can be found in the same bacterial species, and other undescribed conjugation systems might exist [130]. These eight different groups are the basis for the classification of conjugative systems [88] (Figure 1.1). Within each group, at short evolutionary time scale compared to the evolution of the MPF systems, the conjugative system diversifies in term of replicon, some types are more often associated with ICEs while others are more associated with CPs. The conjugative system can evolve toward other use than conjugation. This process is called exaptation [82]. For instance, exaptations

of the conjugative system into protein secretion systems or DNA uptake machinery have been observed [90, 69, 104].

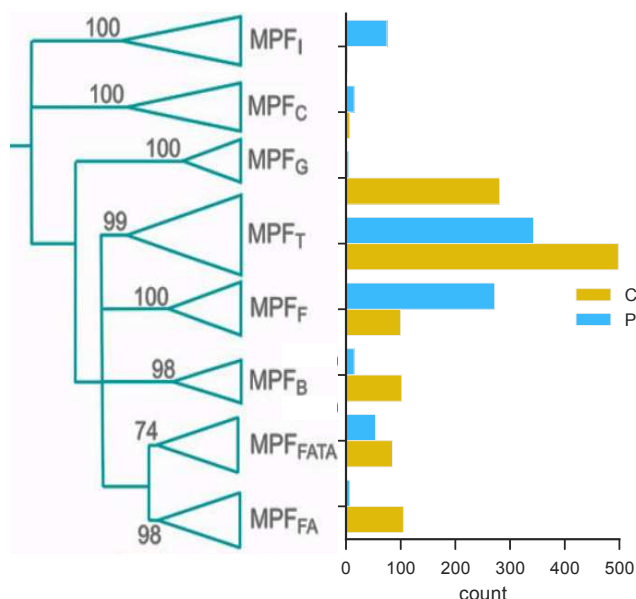


Figure 1.1: Phylogeny of VirB4 defining eight MPF types and their associated replicon type. On the left-hand side, the cartooned phylogenetic tree of the VirB4 protein. On the right-hand side, the type of replicon on which is detected the *virB4* gene (C: Chromosome, P: Plasmid). The conjugative systems were searched on a dataset of 5776 complete genomes downloaded in November 2016. The tree is from [88].

**Diversity of conjugative elements.** Conjugative elements are large mobile genetic elements compared to others, and can reach sizes as large as more than 500kb for ICEs [182] and up to more than 1 Mb for CP [174]. These sizes overlap the size range of the smallest bacterial chromosomes, which makes conjugative elements the largest type of self-transmissible mobile genetic element. In contrast, both elements can be as small as less than 20kb [70, 174]. They encode various phenotypical functions like antibiotic resistance, heavy metal resistance, carbon source utilization, virulence factor, proteins involved in mutualism [109, 174]. ICEs and CPs are modular structures, and endure different recombination events generating diversity among them [109, 96]. Notably, other mobile genetic elements can insert within ICEs or CPs [144, 109]. Likewise, ICEs can insert into other ICEs or MGEs [10]. ICEs and CPs can also recombine with another similar element nearby to generate diversity. ICEs can capture neighboring genes upon excision with secondary recombination site [109].

Finally, ICEs and CPs are widespread among Bacteria, carry dozens of genes with various functions, and most of them are unknown.

### 1.1.3 Mechanism

The conjugative system is composed of four main different components (Figure 1.2) [89]. The largest one in size is the type 4 secretion system (T4SS), which is a multi-molecular system whose primary goal is to allow the transfer of DNA through the bacterial cell membranes of

the donor and the receptor [71]. The second and third elements, are the relaxase (often named *mob* gene, for mobilization) and the origin of transfer (*oriT*). The MOB protein will nick the DNA at the origin of transfer (*oriT*) and bring the DNA to the T4SS. The last element is the type 4 coupling protein (T4CP), it intermediates between the relaxase and the T4SS.

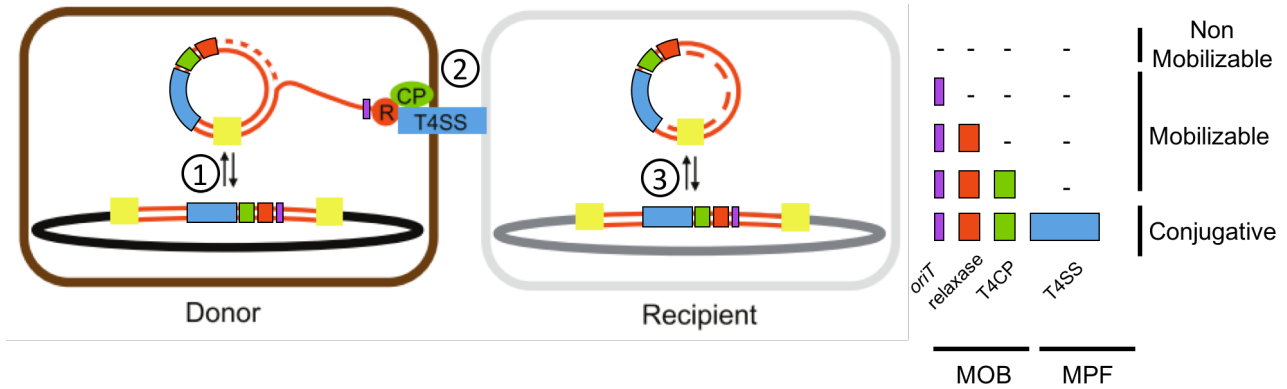


Figure 1.2: Schema of the conjugation mechanism and its main components: the relaxase (red), the type 4 coupling protein (green), the type 4 secretion system (blue) and the origin of transfer, *oriT* (purple). The ICEs have to excise before engaging into the conjugative transfer (①). The conjugative transfer is identical between ICEs and CPs (②). Upon transfer, the ICEs integrate into the host chromosome (③). The diagram on the right summarizes the elements needed for a conjugative or a mobilizable system. The figure is adapted from [90] and the diagram is adapted from [174].

Both ICEs and CPs share the same life cycle, at the exception of two additional steps of integration and excision for the ICEs. These steps involve different types of recombinases, often site-specific tyrosine recombinases, but it could also be serine recombinases or DDE transposases [109]. The conjugation mechanism proceeds as follow:

1. The ICE excises from the chromosome and circularizes. It becomes in appearance a plasmid.
2. The relaxase nicks one strand of the double stranded DNA at the origin of transfer, and binds to it. It forms the transfer DNA (T-DNA).
3. The T-DNA is brought to the T4SS by the intermediate of the type 4 coupling protein.
4. Provided there is a recipient cell nearby, the T4SS will transfer the T-DNA through it.
5. Once in the recipient cell, the T-DNA will be circularized as a single stranded DNA (ssDNA) by the relaxase.
6. In both the recipient and donor cells, the single stranded circular DNA is replicated in a rolling circle manner, the relaxase acting as an initiator of the replication.



7. If the element encodes an integrase, it can integrate into the recipient's chromosome.

For all these steps, the main components can act in *trans*, meaning they can be encoded elsewhere on the chromosome or on another component. This defines different types of element, among conjugative (or self-transmissible), mobilizable and non-mobilizable (Figure 1.2). ICEs and CPs belong naturally to the conjugation type, and their mobilizable counterpart are named accordingly: Integrative and mobilizable elements and mobilizable plasmids. They can be mobilized as soon as they contain a part of the conjugative system, the smallest element required is the *oriT* (Figure 1.2). The class of non-mobilizable element denotes any mobile genetic element whose mobilization mechanism by conjugation is lacking.

### 1.1.4 Objectives

Conjugative elements have been known for 70 years [183]. Most of the work have been carried on model elements and on conjugative plasmids, which provided a detailed comprehension of the different conjugation machineries. This knowledge led to the creation of a method to identify conjugative systems at large scale [90]. This method put forward the evolutionary importance of ICEs with regard to CPs by stressing their abundance in bacterial genomes, unexpectedly outnumbering CPs. Following recent studies suggesting possible interactions between ICEs and CPs with the discovery of plasmid-like functions in ICEs [121, 29], the dichotomy between ICE and CP started to be questioned [28]. Resolving the questions of the potential ubiquity of plasmid-like functions in ICEs and the understanding of ICE and CP dissemination in bacterial populations has been precluded by the lack of tool or method to delimit ICEs in bacterial genomes.

The objective of the work presented in this chapter is to provide an efficient method to delimit ICEs to have a broader view of their composition and distribution in bacterial genomes. The data generated by this method allows, in addition, to describe the relationships between ICEs and CPs at a global scale and to understand what drives the evolutionary success of a given form over the other. This betters our understanding of the benefits and limitations of integrative and extrachromosomal elements in general. It is particularly important to comprehend how these elements are mobilized, as there is a growing concern of the dissemination of elements associated with antibiotic resistance. It has also potential applications in biotechnologies or bioremediation where a strict control on the mobility of these elements is imperative [74].

First, I present the method to detect conjugative systems in bacterial genomes, developed previously [90, 1] and enriched during this work. The second study focuses on ICEs, for which most of the current knowledge is limited to few ICE models. It uses extensively the method to delimit ICEs, described in the first article. The delimitation of ICEs allowed the first large scale analysis of this type element. The delimitation represents a key step before comparing

ICEs with CPs. Finally, the previous works provided tools, methods and data allowing to make the first comparison of ICEs and CPs at large scale. This third paper tries to tackle a broader question: why are there mobile genetic elements that are extrachromosomal while other are integrative?

## 1.2 Methods

### 1.2.1 Article 1: Identifying conjugative plasmids and integrative conjugative elements with CONJScan

The following paper was written following an invitation to contribute a chapter for an edition on *Methods in Horizontal Gene Transfer*, to be published in the lab protocol series *Methods in Molecular Biology*. The paper presents the full methodology to detect conjugative systems in bacterial genomes, along with the method to delimitate ICEs, which is the only one existing so far. Although this method involves manual steps, it allows the delimitations of dozens of elements per days for a trained user. This work has also led to the creation of a tutorial to reproduce step by step, the delimitation of ICEs, when starting from genome identifiers (NC numbers). This tutorial is available at the following address: [https://github.com/gem-pasteur/Macsyfinder\\_models/blob/master/Data/Conjugation/Tutorial\\_ICE.ipynb](https://github.com/gem-pasteur/Macsyfinder_models/blob/master/Data/Conjugation/Tutorial_ICE.ipynb).

## Identifying conjugative plasmids and integrative conjugative elements with CONJscan

5 Jean Cury<sup>1,2</sup>, Sophie Abby<sup>3</sup>, Olivia Doppelt-Azeroual<sup>4,5</sup>, Bertrand Néron<sup>4,5</sup>, Eduardo P. C. Rocha<sup>1,2</sup>

<sup>1</sup> Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France

<sup>2</sup> CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France

10 <sup>3</sup>Univ. Grenoble Alpes, Laboratoire Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications, Grenoble (TIMC-IMAG), F-38000 Grenoble, France; Centre National de la Recherche Scientifique (CNRS), TIMC-IMAG, F-38000 Grenoble, France.

<sup>4</sup> Bioinformatics and Biostatistics Hub – C3BI, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France

15 <sup>5</sup>CNRS, USR 3756, 28, rue Dr Roux, Paris, 75015, France

Running head: Identification of conjugative elements.

20 Key Words: MacSyFinder, conjugation, integrative conjugative element, plasmid, protein profiles, comparative genomics, integrase, genomic islands.

## 1 Summary

We present a computational method to identify conjugative systems in plasmids and  
25 chromosomes using the CONJscan module of MacSyFinder. The method relies on the  
identification of the protein components of the system using hidden markov model profiles  
and then checking that the composition and genetic organization of the system is consistent  
with that expected from a conjugative system. The method can be assessed online using the  
Galaxy workflow or locally using a standalone software. The latter version allows to modify  
30 the models of the module, i.e., to change the expected components, their number, and their  
organization.

CONJscan identifies conjugative systems, but when the mobile genetic element is  
integrative (ICE), one often also wants to delimit it from the chromosome. We present a  
method, with a script, to use the results of CONJscan and comparative genomics to delimit  
35 ICE in chromosomes. The method provides a visual representation of the ICE location.  
Together, these methods facilitate the identification of conjugative elements in bacterial  
genomes.

## 2 Introduction

Conjugative elements transfer large amounts of genetic information between cells, having  
40 an important evolutionary role in bacterial evolution. There are several different types of  
conjugation. For example, some *Actinobacteria* are able to conjugate dsDNA, while others  
make distributive conjugation [1]. These alternative mechanisms will not be discussed here.  
Instead, we will focus on the most common mechanism of conjugation: the transfer of  
ssDNA through a type IV secretion system, or an analogous structure (reviewed in [2,3]).  
45 This mechanism involves three key molecular systems: the relaxosome, the type 4 coupling  
protein (T4CP), and the type IV secretion system (T4SS). The relaxase, often associated with  
other proteins in the relaxosome, interacts with the mobile genetic element (MGE) at the  
origin of transfer, produces a single-stranded cut, and becomes covalently linked with it. The  
T4CP is an ATPase that couples the nucleoprotein filament including the ssDNA and the  
50 relaxase with the T4SS. Finally, the T4SS is a large protein complex that spans both cell  
membranes (in diderms) and is able to deliver the nucleoprotein filament into another cell.  
The identification of a conjugative system requires the identification of these key

components, or at least the most conserved ones: the relaxase, the T4CP and VirB4, the only ubiquitous ATPase of the T4SS. Both relaxases and VirB4 can be divided in a number of families that have been used to type plasmids and to establish the resemblance between T4SS [4,5]. It is important to note that different combination of these three components can have very different functions. MGEs with a relaxase and a T4CP are mobilizable by conjugative systems but are not a conjugative system. Replicons with a T4SS and a T4CP, but devoid of relaxases, may be protein secretion systems or other co-options of the conjugation machinery that are not involved in conjugation.

The analysis of large-scale genome data requires reliable and flexible computational tools to identify and class conjugative elements. Ideally, these tools should allow the user to modify the number and type of components and their genetic organization. MacSyFinder was created with this goal in mind [6]. It can be used with predefined modules (set of protein profiles and definitions), but it can also be customized to meet the researcher needs. The program uses information on the presence and absence of a number of components and their genetic organization to identify systems matching these specifications in genome data. This makes it as simple to use as classical approaches based on blast searches, but allows more powerful queries. It is also more sensitive because it uses protein profiles for similarity search, as well as information on the presence and organization of the different components of the system.

The accuracy of CONJscan depends on its ability to identify the components of the system. When these are highly conserved, *e.g.*, the ATPase VirB4, they can be identified with high sensitivity. When they evolve fast (high sequence divergence) or are facultative components, like the lipoprotein VirB7, the task becomes more complicated. To allow for some flexibility, MacSyFinder searches for components that are expected to be almost always present ("mandatory" components) and those that may be either absent or non-identifiable ("accessory"). Some components can also be described as "forbidden" in which case they cannot occur in the conjugative system. This is useful to distinguish conjugative systems from other co-opted molecular systems (Note 5.1). For each type of components one can set up a minimal quorum. Decreasing the minimal quorum allows to search for more degenerate or distant systems, whereas the inverse only identifies systems closer to the prototypical system.

An important variable in MacSyFinder is the distance allowed between components. While  
85 the T4SS is usually encoded in one operon or a set of contiguous operons (contiguity  
formalized by an authorized inter-genic distance of 30 genes), the relaxase is often encoded  
apart (sometimes with the T4CP). Hence, the inter-genic distance between the relaxase and  
any other elements is increased to 60 genes. In total, three types of information facilitate  
the detection of conjugative systems: the identification of the components (and their type),  
90 the completeness of the set of components expected to form a full system, and their  
proximity in the genome (genetic organization). This information can be put together in a  
text file following a certain grammar that constitutes the *model* of the system that is given  
to MacSyFinder, which searches genomes for instances satisfying these descriptions [6].

The first part of this text shows how one can use MacSyFinder to identify conjugation  
95 systems with the pre-defined models of CONJscan. These models have been validated and  
shown to identify the vast majority of known systems [7,8]. Yet, they may be inadequate in  
certain specific situations. For example, the conjugative systems of a number of taxa (like  
Archaea, Actinobacteria, and Firmicutes) lack known relaxases [2,9]. In this case, the models  
can be adapted to identify novel components (or to accept the absence of the component),  
100 and then easily shared with others *via* text files. The second part of this text describes how  
the analysis of CONJscan can be complemented with a comparative genomics method to  
delimit ICEs in genomes. This uses the core-genome (*i.e.*, the list of gene families present in  
all the genomes available for the species), the genome annotations, the CONJscan results,  
and a script that we provide to plot all this together.

## 105 3 Materials

MacSyFinder reads a *model* and works in two steps: it uses Hidden Markov model (HMM)  
protein profiles to search for the system's components and then checks if their organization  
and quorum respects the specifications of the model (Figure 1).

### 3.1 Sequence data

110 **Proteome data.** MacSyFinder analyzes protein sequences stored in one single file in Fasta  
format. The search for conjugative elements usually requires a completely assembled  
replicon (or a known order of contigs, see Note 5.2). Hence, the file for the analysis should  
contain all the proteins encoded in the genome (or the replicon of interest) in the order of

their genomic position. The corresponding option for this file type is "--ordered\_replicon".

115 The analysis of multiple genomes in one single batch is possible using the type "--gembase", which is similar to the "--ordered\_replicon" but requires special sequence identifiers (see MacSyFinder's documentation).

**Protein profiles.** The protein profiles used by CONJscan are included in the package ([https://github.com/gem-pasteur/Macsyfinder\\_models](https://github.com/gem-pasteur/Macsyfinder_models)). They are described in [7]. The  
120 protein profiles for integrases can be retrieved from PFAM For the Tyrosine recombinase: PF00589 (one single profile). For the Serine recombinase: PF07508 and PF00239 (the protein should hit both profiles to be regarded as an integrase)).

### 3.2 Pre-defined models available in CONJscan

125 CONJscan is a MacSyFinder module that includes a set of pre-defined models and profiles to detect the eight types of ssDNA conjugation systems. These models are used as examples throughout the following sections. The standalone version of MacSyFinder expects to find the files of CONJScan (HMM profiles and model files) at a recognizable location (a folder for the HMM profiles and a folder for the models). Currently, only the standalone version  
130 allows the modification of the models and the introduction of novel protein profiles.

### 3.3 Software and availability

We listed resources of interest for this protocol in Table 2 (see Note 5.3 for issues related with installing the programs).

To run MacSyFinder one needs to install the NCBI/BLAST tools (in particular makeblastdb  
135 version  $\geq 2.8$ , or formatdb), HMMER, and MacSyFinder [10,11,6]. The latter requires a Python interpreter (version 2.7) that must be installed beforehand. See MacSyFinder's online documentation for more details:

<http://macsyfinder.readthedocs.org/en/latest/installation.html>.

To build HMM protein profiles, one also needs a program to make multiple sequence  
140 alignments (*e.g.*, MAFFT [12]), an alignment editor (*e.g.*, Seaview [13]), and a program to cluster proteins by sequence similarity (*e.g.*, Silix or usearch [14,15]). These programs are also required to build the pan-genomes.



CONJScan can be downloaded for local use with MacSyFinder ([https://github.com/gem-pasteur/Macsyfinder\\_models](https://github.com/gem-pasteur/Macsyfinder_models)) or it can be used online (<http://galaxy.pasteur.fr/>, search  
145 CONJScan or the direct link  
[https://galaxy.pasteur.fr/tool\\_runner?tool\\_id=toolshed.pasteur.fr%2Frepos%2Fodoppelt%2Fconjscan%2FConjScan%2F1.0.2](https://galaxy.pasteur.fr/tool_runner?tool_id=toolshed.pasteur.fr%2Frepos%2Fodoppelt%2Fconjscan%2FConjScan%2F1.0.2)). Alternatively, one can query a database of conjugative systems already detected (<http://conjdb.web.pasteur.fr>).

The program MacSyView can be used to visualize the results of MacSyFinder [6]. It can be  
150 used locally (<https://github.com/gem-pasteur/macsyview>) or online  
(<http://macsyview.web.pasteur.fr>).

The script to plot the spots (region between two consecutive core genes) with ICEs can be downloaded for local use ([https://gitlab.pasteur.fr/gem/spot\\_ICE](https://gitlab.pasteur.fr/gem/spot_ICE)).

## 155 4 Methods

The procedure to annotate conjugative elements contains two main steps. In the first step we show how to identify conjugative systems using CONJscan. In the second step we show how to delimit the conjugative element. For the use of the standalone version, we assume some familiarity with a Unix environment (Linux or Mac OS X).

### 160 4.1 Identifying conjugative systems

The identification of conjugative systems in a replicon relies on the CONJScan module for MacSyFinder.

#### 4.1.1 Preparing the data

The protein file should be in multi fasta format (a succession of fasta entries in a text file)  
165 and must be in a directory where the user has permissions to write.

The models and the protein profiles must be in two different folders, typically called "DEF" and "HMM", respectively. The files with the protein profiles must have the same extension. The easiest is to download (or clone) the CONJScan module from the link provided above; where a folder called "Conjugation" has two other folders for the definitions of the models  
170 and for the profiles.

## 4.1.2 Running MacSyFinder

The standalone version of MacSyFinder requires a unix-like terminal. In the terminal, MacSyFinder can be started with a command line. For example, to detect a conjugative system of type F using the default model, one should type:

```
175     macsyfinder typeF \  
        --db-type ordered_replicon \  
        -d Conjugation/DEF \  
        -p Conjugation/HMM \  
180     --profile-suffix .HMM \  
        --sequence-db my_sequence.prt \  
        -o my_sequence_typeF
```

Here, all the paths of the filenames are relative, meaning that MacSyFinder will look for the presence of the files starting from the folder where the command is executed. This can be changed by providing absolute paths. The meaning of the options is the following:

```
185     --db-type is a mandatory parameter that specifies whether the proteins in the  
        multi-fasta file are sorted as they appear along the replicon (for drafts or  
        metagenomes see Note 5.2).  
        -d the path to the definitions of the model.  
        -p the path to the set of protein profiles (--profile-suffix is the suffix of  
190     these files).  
        --sequence-db sets the fasta file with the protein sequences.  
        -o option specifies the name of the output folder. The default name contains the  
        date and time of the command execution. It is advisable to provide meaningful  
        names for these folders [16]. The standard output of the program is saved  
195     automatically in the file macsyfinder.out, in the output folder  
        my_sequence_typeF.
```

In the previous example, we searched for only one of the eight types of MPF available.

Figure 2 describes the different systems in terms of components and genetic organization. If the user wishes to run all the models, we advise to make a loop over each definition with  
200 the previous command line (see Note 5.4 for other possibilities). A bash command to do this follows:

```
        for conj_type in typeF typeB typeC typeFATA typeFA typeG typeI  
        typeT; do  
205     macsyfinder "$conj_type" \  
        --db-type ordered_replicon \  
        -d Conjugation/DEF \  
        -p Conjugation/HMM \  
        --profile-suffix .HMM \  
        --sequence-db my_sequence.prt \  
        -o my_sequence_typeF
```

```

210         -p Conjugation/HMM \
           --profile-suffix .HMM \
           --sequence-db Data/plasmid_seq.prt \
           -o plasmid\_seq\_"$conj_type"
done

```

#### 4.1.3 Analyzing the results

Table 1 presents the different output files with their description. The main output file is located in the folder `my_sequence_typeF/` in the previous example. It is named

215 `macsyfinder.report`. It is a tabular file where each line corresponds to a protein identified as a component of a conjugative system, with its annotation and some results of the detection (including the `hmmer` `i-value` and the alignment coverage with the profile). It contains the predicted system (here `typeF`) (see Note 5.5 for how to class systems). This file is empty when there is no occurrence of a conjugation system that satisfies the model. If in

220 spite of the negative result, one wishes to analyze the presence of proteins that might be components of a conjugation system, which might reveal an atypical or degraded system, this information can be found in the `macsyfinder.out` file. More specifically, this file contains the information on proteins that have matched certain protein profiles of the model and whether they formed a complete system (in which case this is reported in the

225 `macsyfinder.report` file). If the analysis of these results suggests the existence of an atypical yet relevant system, the user can modify the model to account for such cases and re-run MacSyFinder with the novel model (see Note 5.6).

#### 4.1.4 Running MacSyFinder with Galaxy

CONJScan was integrated on the public Galaxy@pasteur instance available at

230 "<https://galaxy.pasteur.fr>". It is classified in the "genome annotation" category. Any user can connect to Galaxy (anonymously or with an account) and launch an execution of CONJScan. The functioning, input, and output of CONJScan in the Galaxy@pasteur instance is similar to the standalone version. The only differences between the two instances concern expert options and the ability to change the models, which are only available for the standalone version.

235 Before selecting a genome to scan, one must upload the data by opening the dialog box (highlighted in green on Figure 3.A). Then this dataset can be selected in the first parameter of the form. The option "Type of dataset to deal with" must be set as in the standalone, typically "ordered replicon" or "gembase" to analyze completely assembled replicons. The option

“Conjugative element to detect” allows to select one model from a precompiled set of models  
240 used by MacSyFinder. When one is not sure of which model to use, one can run the process  
consecutively, changing the model at each time.

Expert users can access and change the hmm search parameters by clicking on the select button  
under the Hmmer code option label. If so, the options, “*Maximal e-value*”, “*Maximal*  
*independent e-value*” and “*Minimal profile coverage*” can be tuned before the execution  
245 (similar options are available in the standalone version). The two former options are specific to  
HMMER (see Table 2) and the latter represents the minimal accepted value for the fraction of  
the profile that is matched in the alignment with the protein sequence.

Once the options are set, the user needs to click on execute to launch the process. The user  
history (the panel on the right) will then be updated with information on the process and the  
250 associated files. New files will appear in grey when the job is waiting to be ran on the Institut  
Pasteur's cluster, in yellow when the process is running (Figure 3.B), and in green when it has  
correctly terminated (clicking the icon circled in red will update the job status).

Among the five outputs of CONJScan (Table 1), the “MacSyView output (CONJScan on data  
1)” allows to visualize the results (Figure 3.C). Clicking on the eye on this file will display the  
255 link “Display in MacSyView” and clicking this link will open the MacSyView web application  
in a new tab of the web browser, automatically filled with the results of CONJScan. The user  
can then browse graphically these results.

The first page of MacSyView displays all occurrences of the systems found on the input data  
(Figure 4). The user can pick an instance to visualize it by clicking on it in the list. The page  
260 displaying the instance is divided in three parts. The first panel shows how the instance fits  
the model in terms of the components of the system. Boxes represent the number of each  
*mandatory*, *accessory*, and *forbidden* components. A tooltip gives the name of the  
component when the mouse hovers a box. The second panel shows the genetic context of  
the system (as transcribed from the input fasta file), with components drawn to scale. When  
265 the mouse hovers a box (circled in red), a tooltip displays information on the corresponding  
component, including the scores of the HMMER hit. This view can be exported as a SVG file  
(tools in blue at panel bottom). The third panel gives detailed information on the  
components of the system.

## 4.2 Delimiting an ICE

270 After the detection of a conjugative system, one is often interested in delimiting the  
 associated mobile genetic element. If the element is a plasmid, then the delimitation is  
 trivial (it's the replicon). However, if the element is integrated in a chromosome (ICE), one  
 needs to delimit the ICE within the replicon. For this, we have developed a method using  
 multiple genomes (usually more than four) of the same species. This is described in the  
 275 following paragraphs.

### 4.2.1 Building core-genomes

The first step of the analysis consists in building the core genome of the species with the  
 ICE. The core genome is the set of genes that are shared by all the genomes of the species,  
 and can be built rapidly using the program Roary [17]. This requires the availability of `gff3`  
 280 files for each genome in the same directory, a file format containing both annotations (list of  
 genes) and the nucleotide sequences. These files can be downloaded from GenBank if the  
 genome sequences are available there. They can also be generated by genome-annotation  
 tools like Prokka [18]. The commands to obtain the core genome are as follow:

```
roary GFF/*.gff
```

285 Roary creates many output files whose description is not in the scope of this chapter. To  
 obtain the information on the core genome one can type:

```
query_pan_genome -g 1493304436/clustered_proteins \  

-a intersection \  

/GFF/*.gff
```

290 Which will create a file called `pan_genome_result` containing the core genes identifiers.  
 The script `query_pan_genome` is installed with Roary. The `-g` option takes the output of  
 the previous command. The `-a intersection` option indicates the script to build the  
 core-genome.

### 4.2.2 Defining the spot

295 Initially, one does not know the location of the ICE, except that its upper boundaries are the  
 two flanking core genes in the genome (we assume here that the ICE is not part of the core  
 genome). We define an interval as the genomic region between the two core genes. We  
 define a spot as the set of intervals flanked by the same two families of core genes across  
 genomes. Since there is only one member of a core gene per genome, the spot has one

300 interval in each genome at most. If there are rearrangements in this region there may not  
be an interval with the same two core gene families in certain genomes. The latter are not  
part of the spot and should be excluded from further analysis. The goal of the method is to  
focus on the interval with the ICE, while accounting for the gene repertoires of the spot, to  
delimit the ICE. We recommend to restrict such analyses to cases with a minimum of four  
305 genomes in the species (the fewer genomes, the weaker the statistical signal).

#### 4.2.3 Delimiting the ICE within the spot

Once the spot with the conjugative system is defined, one needs to build the pan-genome of  
the genes in the spot. The pan-genome is the full set of protein families that are present in a  
given set of genomes (here in the set of proteins in the spot). To identify the pan-genome,  
310 one could re-use the pan-genome build earlier by Roary (built when constructing the core  
genome). However, if that step was skipped because the core-genome was already built  
independently, one can use `usearch` [15], a program that builds protein families rapidly  
using clustering by sequence similarity:

1- Gather all proteins from the spots in one file (`allproteins_spot.prt`).

315 2- Run `usearch`:

```
usearch -quiet -cluster_fast allproteins_spot.prt \  
-id 0.7\  
-uc allproteins_spot-70.uc
```

Where the options are:

320 `-quiet`, to remove the verbose standard output

`-cluster_fast` to use the algorithm `uclust`, a centroid-based clustering  
algorithm.

`-id` sets the percentage of identity for clustering.

`-uc` is a tabular file containing the results of the clustering

325 In the `*.uc` file, a family number is attributed to each protein, and one can easily  
build the protein families.

A visual representation of the spot focused on the genome with the ICE can be done using  
the script `plot_ICE_spot.py` (see [https://gitlab.pasteur.fr/gem/spot\\_ICE](https://gitlab.pasteur.fr/gem/spot_ICE)). This website  
contains also a tutorial on how to delimit ICE. The genes of the ICE are expected to be at  
330 roughly the same frequency in the spot. Hence, their visual representation greatly facilitates

the delimitation of the ICE (for degenerate elements see Note 5.7). The Figure 5 shows an example of the visual representation of the data.

#### 4.2.4 Disentangling tandem elements

335 When two ICEs are inserted into the same spot (in tandem or intermingled), they are both represented on the graph and this can be used to disentangle them. If the elements are intermingled, or if the two ICEs are present in the same set of strains, their discrimination can be difficult (see Note 5.8). When the ICEs are in tandem, the presence of an integrase between the elements can help in this process. Also, tandem ICEs have a similar succession of a tandem of integrases and conjugative systems.

340 When the ICE is in tandem with another mobile genetic element in the same set of strains (such as a prophage or an integrative mobilizable element), their delimitation may be facilitated by the presence of a separating integrase. However, it should be noted that ICEs can excise with neighboring elements, and thus mobilize them [19]. In this case, the difference (and thus the delimitation) between the elements may be questionable.

#### 345 4.2.5 Annotating the elements

The delimited conjugative element can be functionally annotated using a range of computational tools (see Table 2). MGEs such as ICE tend to have many genes with no homologs in the sequence databases (see Note 5.9).

## 5 Notes

### 350 5.1 Co-optations of conjugative systems

Conjugative systems were often co-opted for other functions, notably protein secretion (pT4SS), but also for DNA secretion, and DNA uptake [20,5]. To distinguish between these and conjugative systems it is usually sufficient to identify relaxases in the system, since these are essential for the transfer of DNA, but not for protein secretion. Yet, some systems  
355 may perform both functions. In this case, CONJscan will correctly identify the conjugative system because there is a relaxase. Researchers interested in identifying pT4SS might use the TXSScan module of MacSyFinder [8].

### 5.2 Use of draft genomes and metagenomes

The analysis of draft genomes poses numerous challenges and typically precludes the  
360 identification of conjugative plasmids or ICEs. This is because these elements often carry  
repeats, such as transposable elements, that produce breaks in the assembly process when  
replicons are sequenced using short reads-technology. As a result, MGEs are split in several  
contigs and it is usually difficult to know which contigs belong to which element (except if a  
very similar element is available for comparison). CONJScan can only be used reliably in  
365 draft genomes if one knows the order of the contigs, in which case one can give to the  
program the ordered list of protein sequences, or if the entire conjugative system is a single  
contig (but that is rarely known *a priori*). Note that while T4SS tend to be encoded in one  
single locus, and thus are often in a single contig, the relaxase is often encoded apart and  
may be in another contig. Otherwise, CONJScan may be used to identify the components of  
370 the conjugative machinery in draft genomes.

Metagenomic datasets are even more difficult to analyze because the contigs belong to  
different and unassigned genomes (and are usually very small). In this case, CONJScan can  
be used to identify components of the conjugation system, but the complete systems are  
rarely identifiable.

375 The use of long reads sequencing technologies - that enable easier, longer assemblies, and  
powerful genome binning techniques that efficiently separate contigs between organisms  
using multi-variate analyses - are leveraging these challenges, and promise exciting  
developments concerning the study of conjugative systems in overlooked portions of the  
tree of life [21].

### 380 5.3 Installing issues

The installation of MacSyFinder requires previous installation of makeblastdb (or formatdb)  
and HMMER. It also needs Python 2.7. The detailed procedure for installation can be found  
online (<http://macsyfinder.readthedocs.io/en/latest/installation.html>).

In case of problems with installation or with the use of the programs, we encourage users to  
385 open an “issue” on the corresponding Github's web page (<https://github.com/gem-pasteur/macsyfinder/issues>). Before, users can check if this issue was not already solved  
and described in the sites' “closed issues”.



#### 5.4 Running multiple jobs

One can run MacSyFinder with a number of models and on a number of replicons in several  
390 ways. As a rule, it is better to make a loop in a shell to run the program independently on  
each dataset and on each model (see the main text). The use of the "gembases" format as  
input allows MacSyFinder to analyze a large number of replicons at a time with the same  
model. There is also an option in MacSyFinder to run several models at a time (parameter  
"all", see documentation). Yet, if these models have homologous components or if the  
395 systems are scattered in the replicon, or in tandem, then the program may misidentify  
certain systems. Hence, we advise against the use of this option for non-expert users.

#### 5.5 Classing the systems

Usually the output of CONJScan clearly indicates the MPF type of a system, because of the  
presence of components specific to this MPF. However, some closely related systems,  
400 notably FA and FATA, can sometimes jointly hit the same components. Usually the correct  
system is the one with more components assigned to, and for which the protein profiles  
have better (lower) i-values in the HMMER output. If the situation is unclear, it may be that  
the system is degenerated (few MPF-specific components), intermingled with another, or a  
set of tandem of systems (see Note 5.8).

#### 405 5.6 Modifying the models

The models of MacSyFinder are written in a text file following a specific grammar (see  
MacSyFinder's documentation). It is very simple to change the models' specifications  
regarding the quorum, genetic organization, and role of each component. Most of them can  
be directly altered in the command line, or can be modified permanently in the model text  
410 file by following the predefined syntax. The models can also be improved by adding (or  
removing) novel components. In this case, one must provide the corresponding protein  
profile, which can be retrieved from public databases (like PFAM or TIGRFAM, see Table 2)  
or built specifically for the model (see [22] for a description of this process). They should be  
added to the profiles' directory.

415 5.7 Degenerate elements

Mobile genetic elements may endure inactivating mutations - including large deletions- that result in degenerate elements. These elements may lack a few components of an active conjugation system, but still be regarded as valid systems by CONJscan because they fit the minimal conditions of the model. Degenerate elements may also complicate the

420 identification of the ICEs because the gene families of the element are present at different frequencies in the pan-genome. It is difficult to distinguish mildly degenerate elements from atypical functional ones without experimental data.

5.8 Intermingled ICE

Intermingled components of conjugative elements can occur in a number of cases. Some  
425 MPF types are close and their protein profiles cross-match, leading to a series of components from different MPFs and thus to an apparent intermingled set of conjugative elements. Some cases of true intermingled elements occur when an element integrates inside another element. These cases may be hard to disentangle without experimental evidence.

430 5.9 ORFs with unknown function

Many genes in mobile genetic elements have their function unknown. If the MGE of interest has many such elements after annotation with standard tools (generic tools like Prokka) then specific curation may be necessary, e.g., by using broader databases of protein profiles like EggNOG or using iterative searches with PSI-BLAST or jackhmmer (Table 2). It is not  
435 unusual that several genes remain of unknown function after all these analyses.

## 6 Acknowledgement

This work was supported by the ANR MAGISBAC project, the CNRS and the Institut Pasteur. We thank the collaborators who have worked with us on this topic in the last decade, notably Fernando de la Cruz, Chris Smillie, Marie Touchon, Maria Pilar Garcillan-Barcia, and  
440 Julian Guglielmini.

## 7 References

1. Grohmann E, Muth G, Espinosa M (2003) Conjugative plasmid transfer in gram-positive bacteria. *Microbiol Mol Biol Rev* 67:277-301.
- 445 2. Smillie C, Pilar Garcillan-Barcia M, Victoria Francia M, Rocha EPC, de la Cruz F (2010) Mobility of Plasmids. *Microbiol Mol Biol Rev* 74:434-452.
3. de la Cruz F, Frost LS, Meyer RJ, Zechner E (2010) Conjugative DNA Metabolism in Gram-negative Bacteria. *FEMS Microbiol Rev* 34:18-40.
4. Garcillan-Barcia MP, Francia MV, de la Cruz F (2009) The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev* 33:657-687.
- 450 5. Guglielmini J, de la Cruz F, Rocha EPC (2013) Evolution of Conjugation and Type IV Secretion Systems. *Mol Biol Evol* 30:315-331.
6. Abby SS, Neron B, Menager H, Touchon M, Rocha EP (2014) MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One* 9:e110726.
- 455 7. Guglielmini J, Neron B, Abby SS, Garcillan-Barcia MP, la Cruz FD, Rocha EP (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res* 42:5715-5727.
8. Abby SS, Cury J, Guglielmini J, Neron B, Touchon M, Rocha EP (2016) Identification of protein secretion systems in bacterial genomes. *Sci Rep* 6:23080.
- 460 9. Coluzzi C, Guedon G, Devignes MD, Ambroset C, Loux V, Lacroix T, Payot S, Leblond-Bourget N (2017) A Glimpse into the World of Integrative and Mobilizable Elements in Streptococci Reveals an Unexpected Diversity and Novel Families of Mobilization Proteins. *Front Microbiol* 8:443.
- 465 10. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
11. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7:e1002195.
12. Katoh K, Toh H (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26:1899-1900.
- 470 13. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27:221-224.
14. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116.
15. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.
- 475 16. White EP, Baldrige E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR (2013) Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution* 6:1-10.
17. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691-3693.

- 480 18. Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069.
19. Bellanger X, Morel C, Gonot F, Puymege A, Decaris B, Guedon G (2011) Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. *Mol Microbiol* 81:912-925.
- 485 20. Alvarez-Martinez CE, Christie PJ (2009) Biological diversity of prokaryotic type IV secretion systems. *Microbiol Mol Biol Rev* 73:775-808.
21. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hermsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF (2016) A new view of the tree of life. *Nat Microbiol* 1:16048.
- 490 22. Abby SS, Rocha EPC (2017) Identification of Protein Secretion Systems in Bacterial Genomes Using MacSyFinder. In: Cascales E (ed) *Bacterial protein secretion systems. Methods in Molecular Biology*. Springer, p in press.

**Table 1:** List of MacSyFinder output files and folders.

| <b>Output</b>                   | <b>Description</b>   |
|---------------------------------|--|
| <b>macsyfinder.conf</b>         | Parameters used for the run  |
| <b>macsyfinder.log</b>          | Log information in case of problem                                 |
| <b>macsyfinder.out</b>          | Standard output  |
| <b>macsyfinder.report</b>       | Tabular report with the proteins of conjugation systems            |
| <b>macsyfinder.summary</b>      | Tabular report with the systems and their components               |
| <b>macsyfinder.tab</b>          | Tabular report with the number of systems detected per replicon    |
| <b>results.macsyfinder.json</b> | JSON file summarizing the results for visualization with MacSyView |
| <b>hmmer_results</b>            | Folder containing all the hmmer output files ("raw" and filtered)  |

**Table 2: List of Resources or tools for useful for the detection and annotation of conjugative elements**

| Resource           | Type                    | Description  | Link   |
|--------------------|-------------------------|--|--|
| <b>MacSyFinder</b> | Generic                 | Program to identify molecular systems (of which CONJScan and TXSScan are modules)              | <a href="https://github.com/gem-pasteur/macsyfinder">https://github.com/gem-pasteur/macsyfinder</a>  |
| <b>CONJScan</b>    | Conjugation systems     | MacSyFinder module to identify conjugation systems   | <a href="https://github.com/gem-pasteur/Macsyfinder_models">https://github.com/gem-pasteur/Macsyfinder_models</a><br><a href="https://galaxy.pasteur.fr/">https://galaxy.pasteur.fr/</a> |
| <b>MacSyView</b>   | Generic                 | To visualize MacSyFinder's results   | <a href="https://github.com/gem-pasteur/macsyview">https://github.com/gem-pasteur/macsyview</a>  |
| <b>Blast tools</b> | Sequence analysis       | Rapid sequence similarity searches (incl. blast, psi-blast, makeblastdb)                       | <a href="https://blast.ncbi.nlm.nih.gov">https://blast.ncbi.nlm.nih.gov</a>  |
| <b>HMMER</b>       | HMM analysis            | Allows the use HMM profiles but also build one's own profiles (incl. hmmsearch and jackhmmmer) | <a href="http://hmm.org/">http://hmm.org/</a>  |
| <b>Usearch</b>     | Protein clustering      | Very fast clustering method (uclust) for very similar proteins (>50% identity)                 | <a href="http://drive5.com/usearch/">http://drive5.com/usearch/</a>  |
| <b>MAFFT</b>       | Alignment               | Multiple sequence alignment  | <a href="http://mafft.cbrc.jp/alignment/software/">http://mafft.cbrc.jp/alignment/software/</a>  |
| <b>Seaview</b>     | Alignment Visualization | To visualize and edit multiple alignments  | <a href="http://doua.prabi.fr/software/seaview">http://doua.prabi.fr/software/seaview</a>  |
| <b>Roary</b>       | Pan-genomes             | Construction of core and pan-genomes   | <a href="https://sanger-pathogens.github.io/Roary/">https://sanger-pathogens.github.io/Roary/</a>  |
| <b>PFAM</b>        | Proteins                | Database of HMM profile  | <a href="http://pfam.xfam.org/">http://pfam.xfam.org/</a>  |
| <b>TIGRFAM</b>     | Proteins                | Database of HMM profile  | <a href="http://www.jcvi.org/cgi-bin/tigrfams/index.cgi">http://www.jcvi.org/cgi-bin/tigrfams/index.cgi</a>  |
| <b>EggNOG</b>      | Proteins                | Database of HMM profile for functional annotation  | <a href="http://eggnogdb.embl.de">http://eggnogdb.embl.de</a>  |

|                       |                       |  |   |
|-----------------------|-----------------------|--|---|
| <b>CONJdb</b>         | Conjugation systems   | Database of conjugative systems  | <a href="http://conjdb.web.pasteur.fr/">http://conjdb.web.pasteur.fr/</a>   |
| <b>TXSScan</b>        | Secretion systems     | MacSyFinder module with model for all type of secretion system                   | <a href="https://github.com/gem-pasteur/Macsyfinder_models">https://github.com/gem-pasteur/Macsyfinder_models</a> |
| <b>IntegronFinder</b> | Integrans             | Detects integrons in genomes   | <a href="https://github.com/gem-pasteur/Integron_Finder/">https://github.com/gem-pasteur/Integron_Finder/</a>     |
| <b>RFAM</b>           | RNA                   | Database of Covariance Models to find many types of RNA                          | <a href="http://rfam.xfam.org/">http://rfam.xfam.org/</a>   |
| <b>Infernal</b>       | RNA                   | Searches for RNAs using covariance models, notably for RFAM.                     | <a href="http://eddylab.org/infernal/">http://eddylab.org/infernal/</a>   |
| <b>Silix</b>          | Protein clustering    | Clustering method to build protein families from "blast all against all" results | <a href="http://lbbe.univ-lyon1.fr/-SiLiX-.html?lang=en">http://lbbe.univ-lyon1.fr/-SiLiX-.html?lang=en</a>       |
| <b>Prokka</b>         | Annotation            | Rapid annotation of bacterial genomes  | <a href="https://github.com/tseemann/prokka">https://github.com/tseemann/prokka</a>                               |
| <b>ResFams</b>        | Antibiotic resistance | Database of HMM profiles specific of antibiotic resistance                       | <a href="http://www.dantaslab.org/resfams">http://www.dantaslab.org/resfams</a>                                   |
| <b>CARD</b>           | Antibiotic resistance | Database of antibiotic resistance genes  | <a href="https://card.mcmaster.ca/">https://card.mcmaster.ca/</a>   |
| <b>Victors</b>        | Virulence factor      | Database of Virulent factor  | <a href="http://www.phidias.us/victors/">http://www.phidias.us/victors/</a>                                       |

505

**Figure 1.** Screening genomes for conjugation systems using CONJscan with MacSyFinder.

The components of a conjugation system are an ATPase, a coupling protein ("coupling p.", T4CP), and a relaxase, plus MPF-type specific genes which are found in a particular genetic organization, described in the models for T4SS (see Fig. 2). The CONJscan module turns

510

MacSyFinder into a search engine for conjugation systems in genomes. First, the selected models of conjugation systems are read and the corresponding components are searched by sequence similarity in the genome (multi-fasta file) using HMMER with the HMM profiles of CONJscan. Then the genetic organization of the hits for the components is analyzed to identify sets of hits compatible with the models. Clusters of hits fulfilling the requirements

515

are used to fill up occurrences of the systems. In the end, if the pre-defined number of components is found in the expected genetic organization, the presence of a conjugation system is predicted, for example here, a MPF of type T. Whether fruitful or not, the results of the search are stored in the output files (see Table 1) that can be used to guide the design of customized models.

520

**Figure 2.** The models of CONJscan for each MPF type. Each line is a graphical representation of a model, as defined in the main text. On the left-hand side, in grey, there are the three mandatory proteins, virb4, t4cp and MOB, common for all models. At the bottom, the box "exchangeable" indicates that any of the relaxase profiles can be used for each MPF type.

525

On the right-hand side, colored by type, there are the specific genes of each MPF type. They are coined "accessory" because they are not always identified in the locus for a number of reasons (missing, unidentifiable, etc). Hence, we set a quorum as the minimum number of components in a valid system (in parenthesis, in front of each line). In the model file, one can modify the quorum for the mandatory profiles and the quorum for the total number of profiles (mandatory + accessory).

530

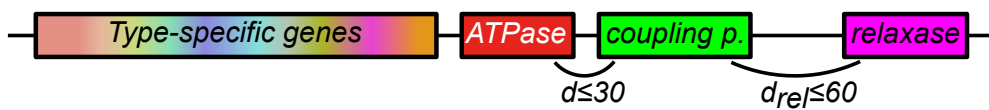
**Figure 3.** Screenshots of the Galaxy interface for CONJscan. See main text for explanations.



**Figure 4.** Screenshot of the Galaxy interface for the visualization of the results of CONJscan  
535 using MacSyViewer. See main text for explanations.

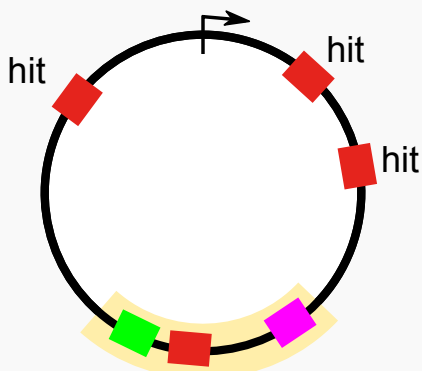
**Figure 5.** Example of a visualization plot. The figure shows two ICEs that are in the same  
spot. Boxes represent genes (width proportional to its length), which are encoded on the  
direct strand (above the line), or the complementary one (below the line). The focal genome  
540 (containing the ICE) is in the middle of its subplot (focal ICEs are turquoise and orange,  
respectively). Each box above a gene in the focal genome belongs to the same pan-genome  
family. On the extremities, the core-genes are represented (these genes families have  
representatives in all genomes, thus the piling of boxes of all colors above them). The two  
ICEs have homologous conjugative systems (with hatches) but not the totality of the ICE.  
545 The grey line represents the GC% along the focal genome. The window on the right-hand  
side represents the number of genes in the interval divided by the number of genes in the  
interval with the largest number of genes of the spot. Here we see that there are few genes  
in the intervals lacking ICEs (this is not necessarily always the case). One can click on the  
genes to see their annotations. A right-click will select a gene as one of the limits (red-boxed  
550 genes near the core-genes). When the two limits are set, the intervening elements will be  
exported in a tabular file when closing this window.

**T4SS model**



**CONJscan models:**

- gene content
- gene organization



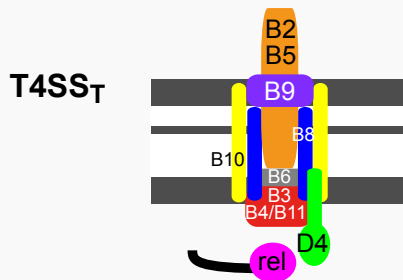
cluster of hits: check

**Proteic multi-fasta**

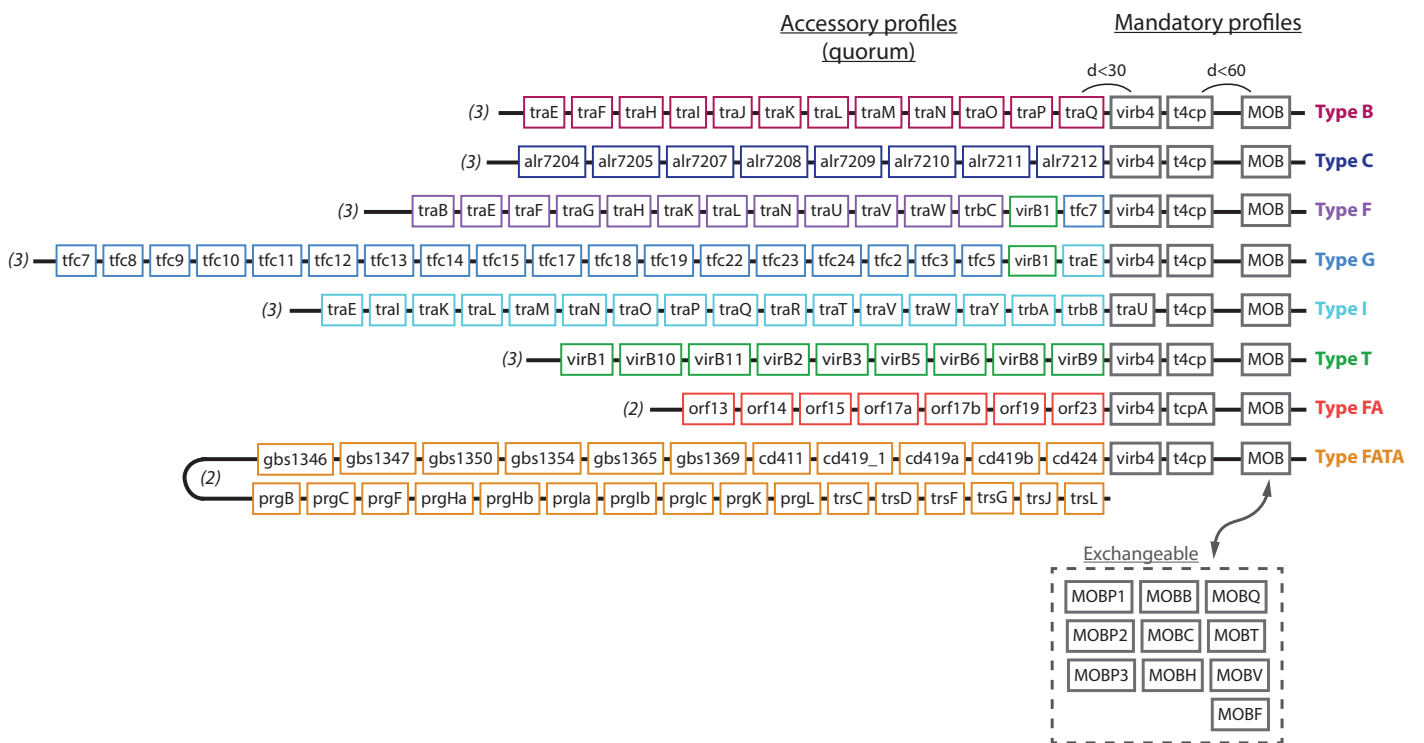
```
>seq1
...
>seq2
...
>seq3
...
```

**Genome screen:**

- similarity search of components
- check organization and content



Predicted systems, output files



**ConjScan : MacSyFinder-based detection of Conjugative elements using systems modelling and similarity search (Galaxy Version 1.0.2)**

**Genome to scan**  
No fasta dataset available.

**The type of dataset to deal with**  
unordered replicon

**Conjugative element to detect**  
typeB

**Tune or leave default values to Hmmer options**  
defaults

**Requirements:** a multifasta file with the protein sequence to be analysed.

1 job has been successfully added to the queue - resulting in the following datasets:

- 2: MaccyView output, ConjScan on data 1
- 3: summary output, ConjScan on data 1
- 4: report output, ConjScan on data 1
- 5: output, ConjScan on data 1
- 6: hmmer results archive, ConjScan on data 1

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

**History**  
Using 10%

6 shown  
104.37 MB

- 6: hmmer results archive, ConjScan on data 1
- 5: output, ConjScan on data 1
- 4: report output, ConjScan on data 1
- 3: summary output, ConjScan on data 1
- 2: MaccyView output, ConjScan on data 1
- 1: Bacteroidetes\_Prok1113a.prot

**2: MaccyView output, ConjScan on data 1**

JavaScript Object Notation (JSON)  
format: maccyview, database: ?

MacSyFinder's results will be stored in conj\_output\_dir  
Analysis launched on /pasteur/projects/policy01/galaxy-prod/galaxy-dist/database/files/000/175/dataset\_175651.dat for system(s):  
- typeB

\*\*\*\*\*  
Analyzing clusters for

```
[{"name": "typeB", "replicon": {"len": 264, "id": "ALF1001c01", "begin_match": 35, "system": "CONJ", "ALF1001c01a_001930", "position": 30.36531365313653136, "sequence_length": 264}}
```

MacSyView - Mozilla Firefox (Private Browsing)

Galaxy MacSyView

macsyview.web.pasteur.fr/#detail:1494576575117:BAFR002c01a\_typeB\_1 67%

Most Visited Getting Started

MacSyView Back to systems list

### Components for system BAFR002c01a\_typeB\_1 (single\_locus)

Replicon BAFR002c01a (length: 4176, topology: circular)

Repertoire of components for detected typeB system

hover boxes to display information

mandatory  
1 1 1

accessory  
1 1 1 1 1 1 1 1 1 1 0 1

forbidden 17

Genomic context for detected typeB system

hover boxes to display information

0 1000 aa

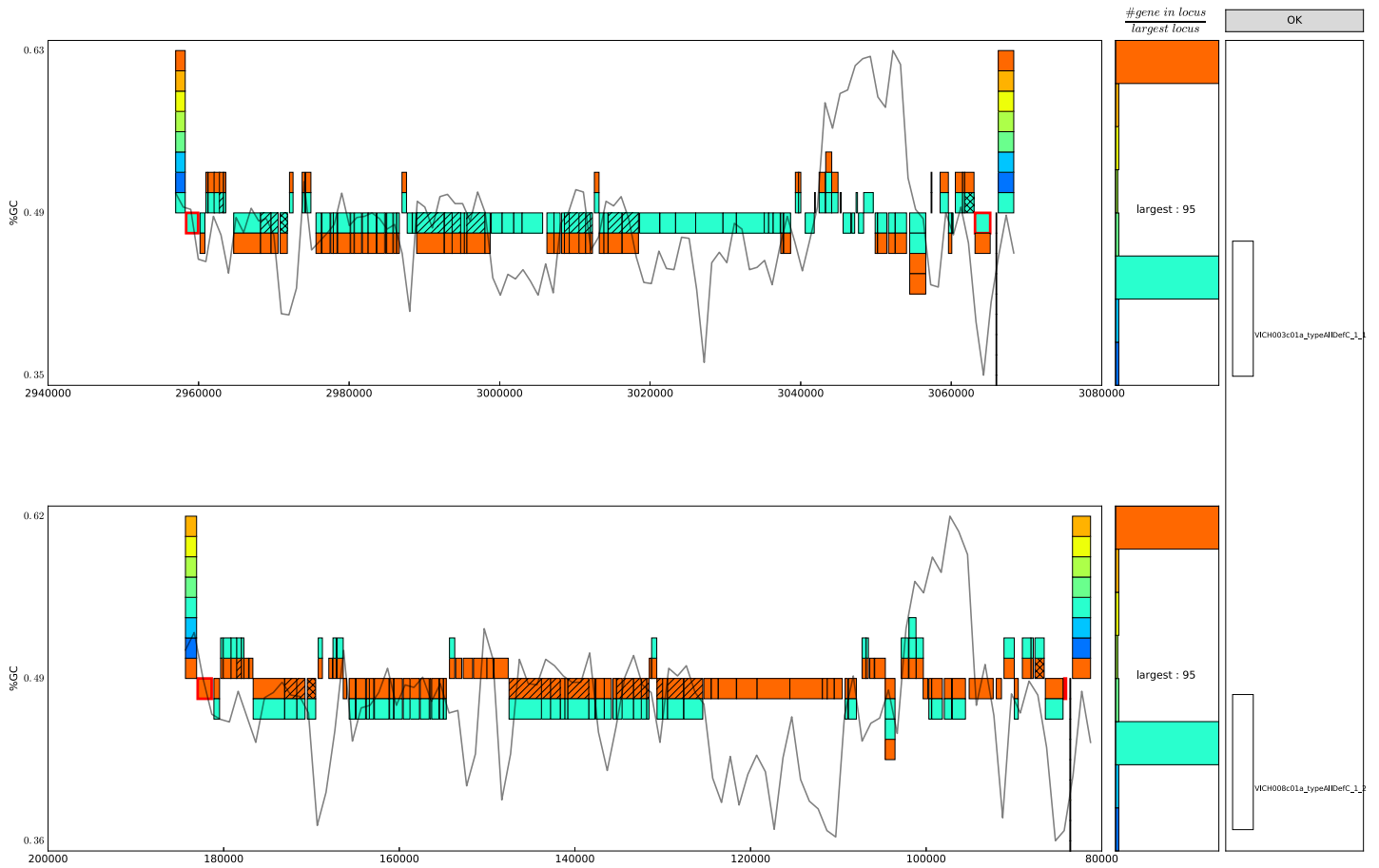
Gene Information

- id: BAFR002c01a\_011710
- length: 302 aa
- match: B\_traN
- i-value: 2.7e-153
- profile coverage: 1.00

Hits summary for detected typeB system

| Color | Sequence Id        | Position | Profile Match | Gene Function | Gene status | System | Protein length (aa) | Score | i-value  | Profile coverage | Sequence coverage | Begin match | End match |
|-------|--------------------|----------|---------------|---------------|-------------|--------|---------------------|-------|----------|------------------|-------------------|-------------|-----------|
|       | BAFR002c01a_011640 | 16510    |               |               |             |        | 71                  |       |          |                  |                   |             |           |
|       | BAFR002c01a_011650 | 16511    |               |               |             |        | 159                 |       |          |                  |                   |             |           |
|       | BAFR002c01a_011660 | 16512    |               |               |             |        | 61                  |       |          |                  |                   |             |           |
|       | BAFR002c01a_011670 | 16513    |               |               |             |        | 498                 |       |          |                  |                   |             |           |
|       | BAFR002c01a_011680 | 16514    |               |               |             |        | 105                 |       |          |                  |                   |             |           |
|       | BAFR002c01a_011690 | 16515    | B_traO        |               | accessory   | typeB  | 149                 | 243.1 | 1e-71    | 1.00             | 0.93              | 2           | 140       |
|       | BAFR002c01a_011700 | 16516    | B_traO        |               | accessory   | typeB  | 192                 | 342   | 1.1e-101 | 1.00             | 1.00              | 1           | 192       |
|       | BAFR002c01a_011710 | 16517    | B_traN        |               | accessory   | typeB  | 302                 | 813.2 | 2.7e-153 | 1.00             | 1.00              | 1           | 302       |

# Conjugation



## 1.3 Results

### 1.3.1 Article 2: Integrative and conjugative elements and their hosts: composition, distribution, and organization

This work was made possible by the development of the previously mentioned method to delimitate ICEs. It is the first large scale description and analysis of numerous ICEs in bacterial genomes. It provides quantitative data on the biology of ICEs which were missing to the field.

# Integrative and conjugative elements and their hosts: composition, distribution and organization

Jean Cury<sup>1,2,\*</sup>, Marie Touchon<sup>1,2</sup> and Eduardo P. C. Rocha<sup>1,2</sup>

<sup>1</sup>Microbial Evolutionary Genomics, Institut Pasteur, 28, rue du Dr Roux, Paris 75015, France and <sup>2</sup>CNRS, UMR3525, 28, rue Dr Roux, Paris 75015, France

Received April 20, 2017; Revised June 30, 2017; Editorial Decision July 04, 2017; Accepted July 04, 2017

## ABSTRACT

**Conjugation of single-stranded DNA drives horizontal gene transfer between bacteria and was widely studied in conjugative plasmids. The organization and function of integrative and conjugative elements (ICE), even if they are more abundant, was only studied in a few model systems. Comparative genomics of ICE has been precluded by the difficulty in finding and delimiting these elements. Here, we present the results of a method that circumvents these problems by requiring only the identification of the conjugation genes and the species' pan-genome. We delimited 200 ICEs and this allowed the first large-scale characterization of these elements. We quantified the presence in ICEs of a wide set of functions associated with the biology of mobile genetic elements, including some that are typically associated with plasmids, such as partition and replication. Protein sequence similarity networks and phylogenetic analyses revealed that ICEs are structured in functional modules. Integrases and conjugation systems have different evolutionary histories, even if the gene repertoires of ICEs can be grouped in function of conjugation types. Our characterization of the composition and organization of ICEs paves the way for future functional and evolutionary analyses of their cargo genes, composed of a majority of unknown function genes.**

## INTRODUCTION

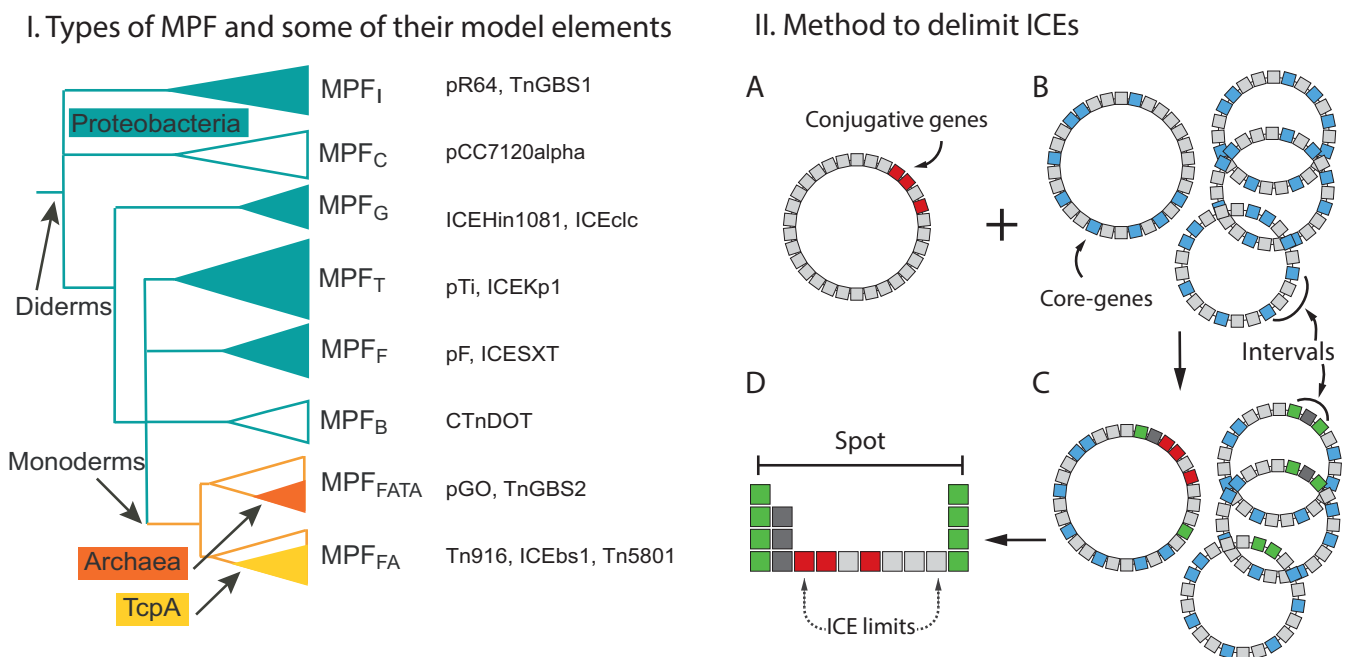
Bacterial diversification occurs rapidly by the constant influx of exogenous DNA by horizontal gene transfer (HGT) (1–3). As a consequence, the diversity of genes found in the strains of a species, its pangenome, is usually much higher than the number of genes found in a single bacterial genome at a given time (4). The pangenome represents a huge reservoir of potentially adaptive genes, whose potential has become evident in the rapid spread of antibiotic

resistance in the last decades (5), and in the emergence of novel pathogens (6). Mobile genetic elements (MGE) drive the spread of genes in populations using a variety of mechanisms, often encoded by the elements themselves.

Conjugative MGEs carry between a few dozens to many hundreds of genes (7). They can be extra-chromosomal (plasmids) or integrative (ICEs). Conjugation requires an initial step of cell-to-cell contact during mating pair formation (MPF). The mechanism is the same for plasmids and ICEs, apart from the initial and final steps of chromosomal excision and integration (8,9). The mechanism of transfer proceeds in three steps. Initially, the relaxase (MOB) nicks the DNA at the origin of transfer (*oriT*), and binds covalently to one of the DNA strands. The nucleoprotein filament is then coupled to the type 4 secretion system (T4SS) and transferred to the recipient cell. Finally, the element is replicated in the original and novel hosts leading to double stranded DNA molecules in each cell. One should note that some integrative elements are transferred using another mechanism, also called conjugation, relying on double-stranded DNA. They are restricted to certain Actinobacteria, have been recently described in detail (10,11), and will not be mentioned in this work. Integration of ICEs is usually mediated by a Tyrosine recombinase (integrase), but some ICEs use Serine or DDE recombinases instead (9,12–14). There are eight types of MPF (15,16), each based on a model system as described in Figure 1. Among those, six MPF types (B, C, F, G, I, T) are specific to diderms, i.e. bacteria with an outer membrane (typically gram negative), while two others (FA, FATA) are specific to monoderms, i.e. bacteria lacking an outer membrane (typically gram positive). Phylogenetic analyses suggest that ss-DNA conjugation evolved initially in diderms and was then transferred to monoderms (15). The identification of a valid conjugative system requires the presence of a relaxase, of the VirB4 ATPase, a coupling ATPase (T4CP), and several other proteins that may differ between types (although they are sometimes distant homologs or structural analogs) (16). Both ICEs and plasmids can be of any MPF type, but some preferential associations have been observed: there are few plasmids of MPF<sub>G</sub> and few ICEs of MPF<sub>I</sub> and MPF<sub>F</sub> (17).

\*To whom correspondence should be addressed. Tel: +33 1 40 61 36 37; Email: jean.cury@normalesup.org





**Figure 1.** Mating Pair Formation (MPF) types and procedure for ICE delimitation. **(I)** The phylogenetic tree displays the evolutionary relationships between MPF types as given by the VirB4 phylogeny. Most lineages are from diderms (green branches), the systems from monoderms (yellow branches, including Firmicutes, Actinobacteria, Archaea, and Tenericutes) being derived from these.  $MPF_B$  (Bacteroides) and  $MPF_C$  (Cyanobacteria) were absent from our data because not enough genomes were sequenced in those clades. The full green clades indicate systems that are typically found in Proteobacteria. The label *TcpA* indicates a clade that uses this protein as T4CP (an homologous ATPase from the typical T4CP – VirD4). In front of each tip of the tree, we indicate a non-exhaustive list of well-known conjugative elements (ICE (starting with ‘ICE’, or ‘(C)Tn’) or plasmid (starting with ‘p’)) for each MPF type. The phylogenetic tree was adapted from (15). **(II)** Scheme of the method. Boxes represent genes, circles represent chromosomes. **(A)** Genes encoding conjugative systems (Red) were detected in bacterial genomes using MacSyFinder. At this stage, this indicates the presence of an ICE that remains to be delimited. **(B)** We restricted the dataset of ICEs to those present in the 37 species for which we had at least four genomes (and a chromosomal conjugative system). We built the core genome (core-genes are represented in blue) of each species. The regions between two consecutive core-genes are defined as an interval. **(C)** The information on the conjugative system and the core-genes is used to delimit the chromosomal interval harboring the ICE. Hence, two core genes flank the ICE (in green). They define an upper bound for its limits. **(D)** Representation of the spot. The two families of core genes (green) define intervals in several genomes of the species (typically in all of them). The set of such intervals is called a spot and is here represented from the point of view of the interval that contains an ICE. We built the spot pan-genome, i.e. we identified the gene families present in the spot, and mapped this information on the interval with the conjugative system. Hence, the bottom layer of genes represents the genes of the interval with the ICE. The upper layers represent other genomes (each layer represents one genome), and the boxes correspond to genes that are orthologs of the genes in the interval with the ICE (genes lacking orthologs are omitted to simplify the representation). Finally, the manual delimitation is based on a visual representation of the spot including this information and the G+C content (see Supplementary Figure S1 and Materials and Methods).

ICEs are more numerous than conjugative plasmids among sequenced genomes (17), but their study is still in the infancy. Beyond the fact that they were discovered more recently, the extremities of ICEs are difficult to delimit precisely in genomic data. Hence, most data available on the biology of ICEs comes from a small number of experimental models, such as the ICE SXT of the  $MPF_F$  type, ICEclc ( $MPF_G$ ),  $MISym^{R7A}$  ( $MPF_T$ ), Tn916 and ICEBs1 ( $MPF_{FA}$ ), CTnDOT ( $MPF_B$ ) (18–21). Several of these elements encode traits associated with pathogenicity, mutualism, or the spread of antibiotic resistance, which spurred the initial interest on ICEs. It has been suggested that this has biased the study of ICE biology and that analyses with fewer *a priori* are needed to appreciate the evolutionary relevance of these elements (9). Some ICEs encode mechanisms typically found in plasmids, such as replication and partition (22–25), and phylogenetic studies showed that interconversions between ICE and conjugative plasmids were frequent in the evolutionary history of conjugation (17). Together, these results suggest that ICE and CP are more simi-

lar than previously thought (26). Interestingly, ICEs also encode functions typically associated with other MGEs, such as phage-related recombinases (27), and transposable elements (28).

The unbiased study of the gene repertoires and structural traits of ICEs is important to improve the current knowledge on these elements. Here, we developed a method based on the comparison of multiple genomes in a species to identify, class, and study ICEs in bacteria. The method allowed us to delimit ICEs, analyze their gene content, study the resemblance between elements and their internal organization, and to study their distribution in chromosomes. Our methodology does not use any *a priori* knowledge on the organization of previously known ICEs, apart from requiring the presence of a conjugative system. Hence, it should not be affected by ascertainment biases caused by the use of model systems to identify novel ICEs. We show that it provides a broad view of the diversity of ICEs among bacterial genomes.

## MATERIALS AND METHODS

### Data

The main dataset used in this study concerns 2484 complete genomes of Bacteria that were downloaded from NCBI RefSeq (<http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/>), in November 2013. A post-hoc validation dataset was obtained from the same database in November 2016 and included a total of genomes. We used the classification of replicons in plasmids and chromosomes as provided in the GenBank files. We searched for conjugation systems in all replicons of all genomes of the two datasets. Yet, the delimitation of ICEs was restricted to species having at least one chromosomally encoded conjugative system and at least four genomes completely sequenced (37 species, 506 genomes) in the main dataset. The validation dataset was used to delimit novel ICEs of the MPF<sub>T</sub> type. These were used for post-hoc validation only. Sequences from experimentally validated ICEs were retrieved from the ICEberg database version 1.0 (<http://db-mm1.sjtu.edu.cn/ICEberg/>).

### Detection of conjugative systems

Conjugative systems were found with the CONJscan module of MacSyFinder (29), using protein profiles and definitions following a previous work (16) (File S1). Protein profiles are probabilistic models built from the information contained in proteins alignments. They allow more sensitive identification of distant homologs than classical pairwise sequence-search approaches (30). MacSyFinder uses the protein profiles and a set of rules (defined in *models*) about their presence in a given MPF type and their genetic organization. For the latter, we used definitions from previous works from our laboratory: two components of the conjugation genes must be separated by less than 31 genes, an exception being granted for relaxases that can be distant by as much as 60 genes. An element was considered as conjugative when it contained the following components of the conjugative system: VirB4/TraU, a relaxase, a T4CP, and a minimum number of MPF type-specific genes: two for types MPF<sub>FA</sub> and MPF<sub>FATA</sub>, or three for the others. MacSyFinder was run independently for each given MPF type with default parameters (hmmer *e*-value < 0.001, protein profile coverage in the alignment higher than 50%). Conjugative elements of some taxa lack known relaxases, this is the case of some Tenericutes and some Archaea. Since T4SS can be mistaken by protein secretion systems in the absence of relaxases, such systems were excluded from the analysis. The models in CONJscan can be modified by the user. The CONJscan module for MacSyFinder (downloadable at [https://github.com/gem-pasteur/Macsfinder\\_models](https://github.com/gem-pasteur/Macsfinder_models)) can be used with command lines in a unix-like terminal, or in a webserver (<https://galaxy.pasteur.fr>, see availability section).

### Identification of gene families, core and pan-genomes, spots and intervals

The identification of an ICE at the locus of the conjugative system uses information from comparative genomics. It re-

quires the definition within each species of a core genome, a set of intervals, and a set of spots.

The *core genome* is the set of families of orthologous proteins present in all genomes of the species. We computed the core genome of each species as in (31). Briefly, orthologous genes were identified as the bi-directional best hits (BBH, using global end-gap-free alignments with more than 80% of protein similarity, <20% of difference in length, and having at least four other pairs of BBH hits within a neighborhood of ten genes), and the core genome was defined as the intersection of the pairwise lists of orthologs between genomes using the reference strain as a pivot.

We defined *intervals* as the loci between two consecutive core genes in a genome (Figure 1). If these core genes have no intervening gene, then the interval is empty; otherwise it contains a number of *accessory* genes (i.e. genes not present in the core genome). We defined a spot as the set of intervals flanked by members of the same pair of core gene families (Figure 1). Consider two families of core genes *X* and *Y* that are consecutive in all *N* genomes of a species (i.e. no core gene is between them). Each genome has thus an interval (*I<sub>i</sub>*) at this location that is flanked by the members of *X* (*X<sub>i</sub>*) and *Y* (*Y<sub>i</sub>*) in the genome (*G<sub>i</sub>*). The spot *i* is the set of the intervals *I<sub>i</sub>* in the species. Note that by definition there cannot be more than one interval per spot in a genome. If the region has not endured chromosomal rearrangements (the most typical situation), then the spot will contain as many intervals as the number of genomes. We identified the gene families of the spots that encode at least one ICE (spot pan-genome). For this, we searched for sequence similarity between all proteins in the spot using blastp (version 2.2.15, default parameters). The output was then clustered to identify protein families using Silix (version 1.2.8) (32). Proteins whose alignments had more than 80% identity and at least 80% of coverage were grouped in the same family. The members of the spot pan-genome that were not part of the core genome constituted the accessory genome.

### Delimitation of ICEs

We analyzed conjugative systems encoded in chromosomes, and delimited the corresponding ICEs using comparative genomics. There is usually a high turnover of ICEs at the species level (i.e. most elements are present in only a few strains), implicating that a few genomes are usually sufficient to delimit the element by analyzing the patterns of gene presence and absence. We restricted our analysis to species with at least four genomes completely sequenced and assembled (without gaps). ICEs were delimited in two steps. First, we identified the spots encoding conjugative systems (see definition above). The core genes flanking these spots provide upper bounds for the limits of the ICE. Second, we analyzed the interval with the ICE and identified the limits of the element by overlaying the information on the presence of genes of the conjugation system, on G+C content, and on the frequency of accessory genes in the spot pan-genome. The genes of the ICE are expected to be present in the spot pan-genome at similar frequencies (some differences may be caused by mutations, deletions, transposable elements, and annotation errors) and this information is usually sufficient to delimit the ICE. We produced a

visual representation of this data in the context of the spot, and used it to precisely delimit the ICE at the gene-level (see Figure 1 and Supplementary Figure S1).

### Specific functional analyses

Antibiotic resistance genes were annotated with the Resfam profiles (core version, v1.1) (33) using HMMER 3.1b1 (34), with the option `-cut_ga`. The cellular localization was determined with PsortB (version 3.0) (35), using the default parameters for diderms and monoderms separately. Genes encoding stable RNAs were annotated using Infernal (36) and Rfam covariance models (37) (hits were regarded as significant when  $e$ -value  $\leq 10^{-5}$ ). Integrons were detected using IntegronFinder v1.5 with the `-local_max` option (38). DDE transposons were annotated with MacSyFinder (29) following the procedure described in (31). Integrases were detected with the PFAM profile PF00589 for tyrosine recombinases and the pair PF00239 and PF07508 for Serine recombinases (<http://pfam.xfam.org/>) (39). HMMER hits were regarded as significant when their  $e$ -value was smaller than  $10^{-3}$  and their alignment covered at least 50% of the protein profile.

Specific HMM protein profiles were built with HMMER v3.1b1 for partition systems (40,41), replication proteins (42), and entry exclusion systems (43). In the general case, we started from a few proteins with experimental evidence of the given function, curated by experts, or reported in published databases. Since these sets were usually small and present in a small number of species, we used a two-step procedure (described below): we started by building preliminary profiles, used them to scan the complete genome database, and then used these results to make the final profiles. First, the proteins of experimental model systems were aligned with mafft v7.154b (with `-auto` parameter) (44), and manually trimmed at the N- and C-terminal ends with SeaView v4.4.1 (45). The alignments were used to make preliminary HMM profiles using hmmbuild from HMMER v3.1b1. A first round of searches with these profiles using hmmsearch ( $e$ -value  $< 10^{-3}$  and coverage  $> 50\%$ ) returned hits that were clustered with usearch (`-cluster_fast` at 90% identity) (46). We took only the longest protein of each cluster, to remove redundant sequences, and searched for sequence identity between all pairs of these representative hits using blastp v2.2.15 (with the `-F F` parameter to not filter query sequences). The output was then clustered to identify protein families using Silix (version 1.2.8, 40% identity and 80% of coverage). We made multiple alignments of the resulting families and used them to build a novel set of HMM protein profiles (alignment and trimming as above). The detailed procedure used for building the protein profiles of each function is given in the supplementary material.

### Functional annotation

We used HMMER v3.1b1 ( $e$ -value  $< 0.001$  and coverage of 50%) to search ICEs for hits against the EggNOG Database of hmm profiles (Version 4.5, bactNOG). These results were used to class genes in broad functional categories. We added a class 'Unknown' for genes lacking hits when queried with EggNOG. We tested the over-representation of given func-

tional categories in ICE using binomial tests where the successes were given by the number of times that the relative frequency of a given functional category was higher in the ICE than in the rest of the host chromosome. Under the null hypothesis, any given category is as frequent in the ICE as in the rest of the chromosome (relative to the number of genes). Formally:

$$H_0 : \forall x \in \text{Cat}_{\text{EggNOG}} : N(f(x_{\text{ICE}}) > f(x_{\text{HOST}})) \sim \mathcal{B}(N_{\text{ICE}}, 0.5).$$

With  $N(f(x_{\text{ICE}}) > f(x_{\text{HOST}}))$  being the number of times the EggNOG category  $x$  had a higher relative frequency in the ICE than in the rest of the host chromosome, and with  $\mathcal{B}(N_{\text{ICE}}, 0.5)$  being a binomial distribution with  $N_{\text{ICE}}$  trials (199 typed ICEs) and an *a priori* probability of 50%.

### Networks of homology

We searched for sequence similarity between all proteins of all ICE elements using blastp v2.2.15 (default parameters) and kept all bi-directional best hits with an  $e$ -value lower than  $10^{-5}$ . We used the results to compute a score of gene repertoire relatedness for each pair of ICEs weighted by sequence identity:

$$wGR R_{A,B} = \sum_i \frac{id(A_i, B_i)}{\min(A, B)} \xleftrightarrow[\text{iff}]{} \text{evaluate}(A_i, B_i) < 10^{-5}$$

where  $(A_i, B_i)$  is the pair  $i$  of homologous proteins in ICEs  $A$  and  $B$ ,  $id(A_i, B_i)$  is the sequence identity of their alignment,  $\min(A, B)$  is the number of proteins of the element with fewest proteins ( $A$  or  $B$ ). The wGRR varies between zero and one, it represents the sum of the identity for all pairs of orthologs of the two ICEs, divided by the number of proteins of the smallest ICE. It is zero if there are no orthologs between the elements and one if all genes of the smaller element have an ortholog 100% identical in the other element. A similar score was previously used to compare prophages (47). We kept pairs of ICEs with a wGRR higher than a certain threshold (5% for the general analysis and 30% for the supplementary analysis). A wGRR of 5% represents, for instance, the occurrence of one homologous gene with 100% identity between two ICEs, where the smallest encodes 20 proteins. We computed two versions of this score for each pair of ICEs, one with all proteins and another after excluding the proteins from the conjugative system.

### Phylogenetic tree

The phylogenetic analysis used the 134 integrases from ICEs containing exactly one tyrosine recombinase. We excluded the other elements because additional recombinases might be involved in other functions (dimmer resolution, DNA inversions) and could confound the results. We added to this dataset: 60 phage integrases randomly selected from a database of 296 non-redundant ( $< 90\%$  identity) phage integrases from RefSeq; 11 integrases from pathogenicity islands (Supplementary Table S2 from (48)); 25 experimentally studied XerC and XerD from (49), XerS from (50), XerH from (51,52); seven integrases representing the diversity of integron integrases (38). The final dataset was composed of 237 integrases. The initial alignment was made

with mafft (with parameters  $-\text{maxiterate } 1000 -\text{genafpair}$ ), and 100 alternative guide-trees were built. We then removed non-informative positions in the multiple alignment (columns) when they had less than 50% of confidence score, as calculated by GUIDANCE 2 (53). We used Phylobayes MPI (version 1.7, model CAT+GTR+Gamma, two chains for 30 000 iterations) to build the phylogeny from that alignment (54). Following the guidelines of Phylobayes, we considered that chains had converged enough to give a good picture of the posterior consensus when the maximum difference across all bipartitions was lower or equal than 0.3 (we obtained 0.24). The tree was represented with Figtree v1.4.2.

### Identification of origin and terminus of replication

The predicted origin (*ori*) and terminus (*ter*) of replication were taken from DoriC (55). When one predicted replicore was more than 20% larger than the other, the genome was removed from the corresponding analysis. The leading strand was defined as the one showing a positive GC skew (56).

## RESULTS AND DISCUSSION

### Identification and delimitation of ICEs by comparative genomics

We developed a procedure to identify and delimit ICEs in two stages (see Materials and Methods). First, we identified conjugative systems in bacterial chromosomes using the CONJscan module of MacSyFinder, a methodology that we have previously shown to be highly accurate (17,57). We then concentrated our attention on the conjugative systems of species for which at least four complete genomes were available. The comparative genomics data provides information on the maximal size of the ICE, given by the flanking core genes, and on the frequency of the genes in the locus. Hence, we defined for each species: the core-genome, the intervals (locations between consecutive core genes), the spots (sets of intervals flanked by the same families of core genes, see Methods), and the pan-genomes of each spot. ICEs were always in a single interval, ICEs were never part of the core genome, and could be preliminarily delimited at their flanking core genes. We then analyzed the frequency of gene families in the spot pan-genome, the position of conjugative systems, and the G+C content (Figure 1 and Supplementary Figure S1). The integration of this information allows to identify the set of genes that are part of the ICE. Of note, we were not able to obtain a general method to identify accurately the attachment sites (*attL* and *attR*) delimiting ICEs. Hence, we placed the elements' borders at the edges of their flanking genes. The customizable standalone and online tools to identify ICEs and a tutorial for their use are here made freely available (see Availability section).

We identified five pairs of ICEs in tandem. They were relatively easy to identify with our procedure because the corresponding intervals had two copies of the conjugation apparatus (and typically two integrases). The tandem ICEs corresponded to different MPF types in four out of five pairs. This fits previous observations that tandems of identical ICEs are very rare in *recA*<sup>+</sup> backgrounds (all the ICEs

analyzed in this work are in these circumstances), presumably because they are rapidly deleted by homologous recombination (58). Some elements encoded two or more integrases or relaxases and only one conjugation system. They may be composite elements resulting from the independent integration of different elements, or they may correspond to ICE encoding additional tyrosine recombinases with other functions (than that of being an integrase). When faced with such elements, and when the frequency of the genes was homogeneous, we included them in one single ICE. This choice was based on the published works showing that ICEs can mobilize composite elements, including genomic islands or IMEs (59–61), and on the observation that conjugative plasmids sometimes also include other integrative elements and multiple relaxases (7).

The curation process started with 601 conjugative systems detected among 2484 complete genomes. Among these, we selected 207 (~35%) elements that were present in species with more than four complete sequenced genomes, and we ultimately were able to delimit 200 ICEs within 37 species. A total of 41 had several ICEs, typically in different genomes, accounting for a total of 118 elements. Hence, integration of different ICEs in the same locus is common. To assess if these ICEs are different elements or the result of single ancestral integrations, we computed the weighted gene repertoire relatedness (wGRR), which represents the proportion of homologous genes between two ICEs weighted by their sequence identity (see Methods, Supplementary Figure S2). Most ICEs had very low wGRR when compared with any other element. Yet, there were 61 ICEs very similar, i.e. with wGRR > 90% (70 ICEs with wGRR > 80%), to at least one of the ICEs in the same spot. If one had kept only one ICE per family with this threshold of wGRR > 90%, this would have reduced the number of ICEs from 200 to 160 (Supplementary Figure S2). Yet, these elements are not necessarily derived from the same ancestral event of integration since we found 29 ICEs with similarly high wGRR values in different species and 24 in different genera. These latter elements are most likely the result of multiple independent integrations in the same locus, not of a single ancestral integration, since ICEs are never part of the species core genome. Hence, instead of arbitrarily selecting one ICE per spot, we opted to maintain all elements in the subsequent analyses and control for this effect, when necessary.

We used information from the ICEberg database to validate the delimitation procedure. We could not use the whole database because there was no complete conjugative system in 51% of the elements of ICEberg (184 out of 358), and 40% of them even lacked the essential protein VirB4/TraU. Hence, we restricted our comparisons to the 19 ICEs in ICEberg that were derived from experimental data or predicted from literature and that were in a genome available in our dataset (Supplementary Table S1). Among these elements, 16 ICEs had their start and end positions within less than 2.5 kb of ours (typically less than 500bp, Supplementary Figure S3). We analyzed in detail the three cases showing discordance between the two datasets. Two of them corresponded to a tandem of ICEs that we split in our procedure because they had two different complete conjugative systems. They were identified as a single ICE in ICEberg.

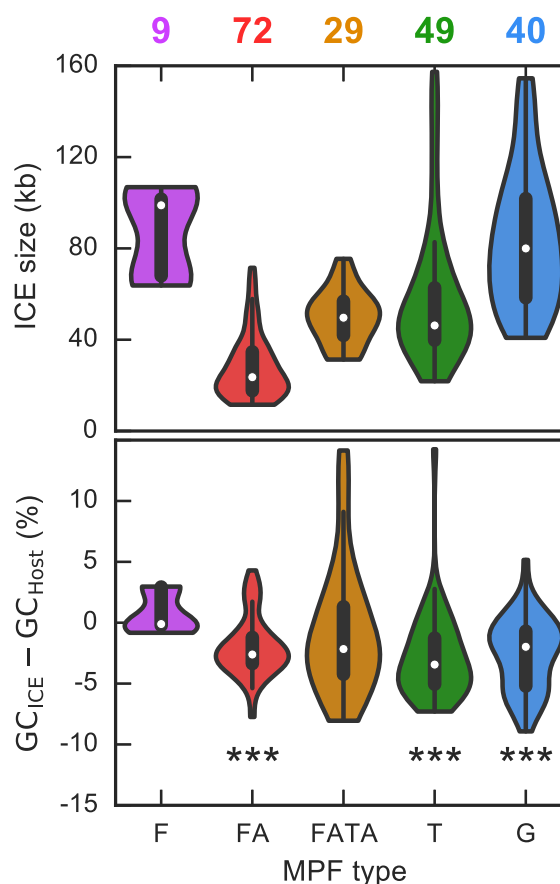
The third discordance arose because the ICEberg annotation included an MPF<sub>FATA</sub> ICE interrupted by an MPF<sub>FA</sub> ICE. Our procedure spotted the complete MPF<sub>FA</sub> ICE. These results suggest that our delimitation is more accurate. It is important to note that our procedure does not use any information on the experimental models of ICE; it is only based on the identification of conjugation proteins, and the frequency of accessory genes. Hence, the accuracy of our method is expected to remain high even when studying poorly known ICE families. It is also expected to improve upon inclusion of more genomes within a species.

By the end of this study there was an influx of novel complete genomes in the public databases. We took advantage of this data to make a *post hoc* validation of our results. We concentrated our attention on MPF<sub>T</sub> elements because they are the most abundant and the best studied. We identified 1181 conjugative systems in the chromosomes of the novel genomes. We delimited 124 novel MPF<sub>T</sub> ICEs in the novel genomes (out of 498 detected) and compared them with the ICEs of the same type in the main dataset, according to five measures. The results showed no significant difference between the two sets (Supplementary Tables S7 and S8), suggesting that our results are robust to sampling effects, *i.e.*, adding novel data will not significantly affect the main conclusions of our work.

### General features of ICEs

We grouped the 200 ICEs (delimited in the main dataset) on the basis of the MPF and relaxase (MOB) types. We identified five of the eight previously defined MPF types among these elements: type F (9), FA (72), FATA (29), T (49) and G (40). Three types were absent from our data because they are strictly associated with clades for which not enough complete genomes were available (MPF<sub>B</sub> for Bacteroides and MPF<sub>C</sub> for cyanobacteria) or corresponded to systems that were previously shown to be extremely rare among ICE (MPF<sub>I</sub>) (17). A preliminary analysis of the *post hoc* validation dataset mentioned above revealed no ICEs for MPF<sub>I</sub> and MPF<sub>C</sub>, and very few MPF<sub>B</sub> (because they are in poorly sampled species). Study of these elements will have to wait before more data becomes available. As a result, the distribution of MPF types differed from that of the 601 conjugative systems identified in all chromosomes of the main dataset ( $\chi^2$ ,  $P$ -value < 0.001, Supplementary Figure S4), and included mostly Firmicutes (for MPF<sub>FA</sub> and MPF<sub>FATA</sub>) and Proteobacteria (for the others). One ICE had an undetermined MPF type (albeit it included a VirB4, a MOB and a coupling protein), and was excluded from the remaining analyses (except from the wGRR network, see below). Expectedly, the types of relaxases identified in the ICEs corresponded to those frequently found in Proteobacteria and Firmicutes (17) (Supplementary Figure S5).

The size of ICEs described in the scientific literature varies between ~13kb (ICE<sub>Sal</sub> in *Staphylococcus aureus* (62)) and ~500kb (ICE<sub>M1SymR71</sub> in *Mesorhizobium loti* R7A (63)). In our dataset, the smallest ICE was also identified in *S. aureus* (strain USA300-FPR3757) and was only 11.5 kb long. The largest ICEs were found in *Rhodopseudomonas palustris* BisB18 (MPF<sub>T</sub>) and *Pseudomonas putida* S16 (MPF<sub>G</sub>) and were around 155 kb long. The distribution



**Figure 2.** ICE statistics as a function of the MPF type. **Top.** Distribution of the size of ICEs (in kb). The numbers above each violin plot represent the number of elements in each category. **Bottom.** Distribution of pairwise differences between the GC content of the ICE and that of its host. The violin plots represent the kernel density estimation of the underlying values. Here the violin plots are limited by the minimum and maximum values. \*\*\* $P$ -value < 0.001, Wilcoxon signed-rank test (rejecting the null hypothesis that the difference is equal to zero).

of ICE size per MPF type showed that type F (median size of 99 kb) and G (median 80 kb) were the largest, whereas those of type FA were the smallest (median 23.5 kb) (Figure 2). The distribution of the size of ICE in our dataset is close to that of ICEs reported in the literature, even if we could not identify any ICE of a size comparable to ICE<sub>M1SymR71</sub>. Such large ICEs may be rare or specific of taxa not sampled in our study. The majority of ICEs were found to be AT-rich compared to their host's chromosome (Figure 2). This is, with some exceptions, a general trend for mobile genetic elements (64). The distribution of sizes of these 160 families of ICEs is similar to the one for the entire dataset (Supplementary Figure S6), validating our decision to keep all ICEs in our analysis.

It was known that the largest plasmids are typically in the largest genomes (7). We observed a positive correlation between the size of ICEs and that of their host chromosomes ( $\rho = 0.47$ ,  $P$ -value < 0.0001, after discounting the ICE size from chromosome size), even if this is partly because the smaller types of elements are associated with the

clades with the smallest genomes (Supplementary Figure S7). Larger ICEs may be disfavored in smaller genomes because they are harder to accommodate in terms of genome organization (65), thus leading to higher fitness cost, or because smaller genomes endure lower rates of HGT (66). Interestingly, the average size of ICEs (52.4 kb) could be a general feature of integrative elements, since it comes near to that of temperate phages (~50 kb) of enterobacteria (67), and of known pathogenicity islands (ranging between 10 kb and 100 kb (68)).

### The families of ICEs

We analyzed the network of protein sequence similarity between ICEs in relation to some of the best-known experimental models (SXT, ICEclc, Tn916, ICEBs1, TnGBS2, ICEKp1). To this end, we used the abovementioned weighted gene repertoire relatedness (wGRR) scores between every pair of ICEs. The network of wGRR-based relationships between ICEs was represented as an undirected graph, where nodes represent ICEs and edges were weighted according to the wGRR (if wGRR > 5%) (Figure 3). This graph showed that all ICEs were connected in a single component (all nodes can be accessed from any others) with the exception of a group of ICEs from *H. pylori*. ICEs grouped predominantly according to their MPF types, which were sometimes split in several clusters. Interestingly, ICEs in Firmicutes (FA and FATA) and Proteobacteria (T, G and F) formed two main groups, and their sub-groups typically included different species. This fits the observation that FA and FATA are sister-clades in the phylogeny of VirB4 (15). Interestingly, a similar graph where MPF proteins were excluded from the calculation of wGRR produced very similar results, suggesting that the effect of the host taxa may be preponderant in the split into two groups according to the major phyla (Supplementary Figure S8). To detail the similarities between ICEs, we also clustered them using a more restrictive threshold (wGRR > 30%, Supplementary Figure S9). This graph is composed of many connected components of single MPF types, highlighting the huge diversity of ICEs.

### Independent integrase acquisitions by ICE

Integrases allow the integration of ICEs in the chromosome and are one of their most distinctive features (relative to conjugative plasmids). Around 70% of the ICEs encoded a single tyrosine recombinase, nine encoded a Serine recombinase, 21 had at least two integrases, among which six had both Serine and Tyrosine recombinases. A total of 37 ICEs lacked integrases, among which six encoded one or more DDE recombinases (in two cases the genes are at the edge of the ICE) and six encoded pseudogenized integrases. DDE recombinase-mediated ICE integration was previously described for ICE TnGBs2 in *Streptococcus agalactiae* (13) and for ICEA in *Mycoplasma agalactiae* (14). Experimental work will be necessary to test if some of the six ICEs with only DDE recombinases use them to replace integrases. Of note, we actually found more transposases in ICEs with Serine or Tyrosine recombinases than in those lacking them ( $\chi^2$  on a contingency table,  $P$ -value < 0.01). Recently, it has

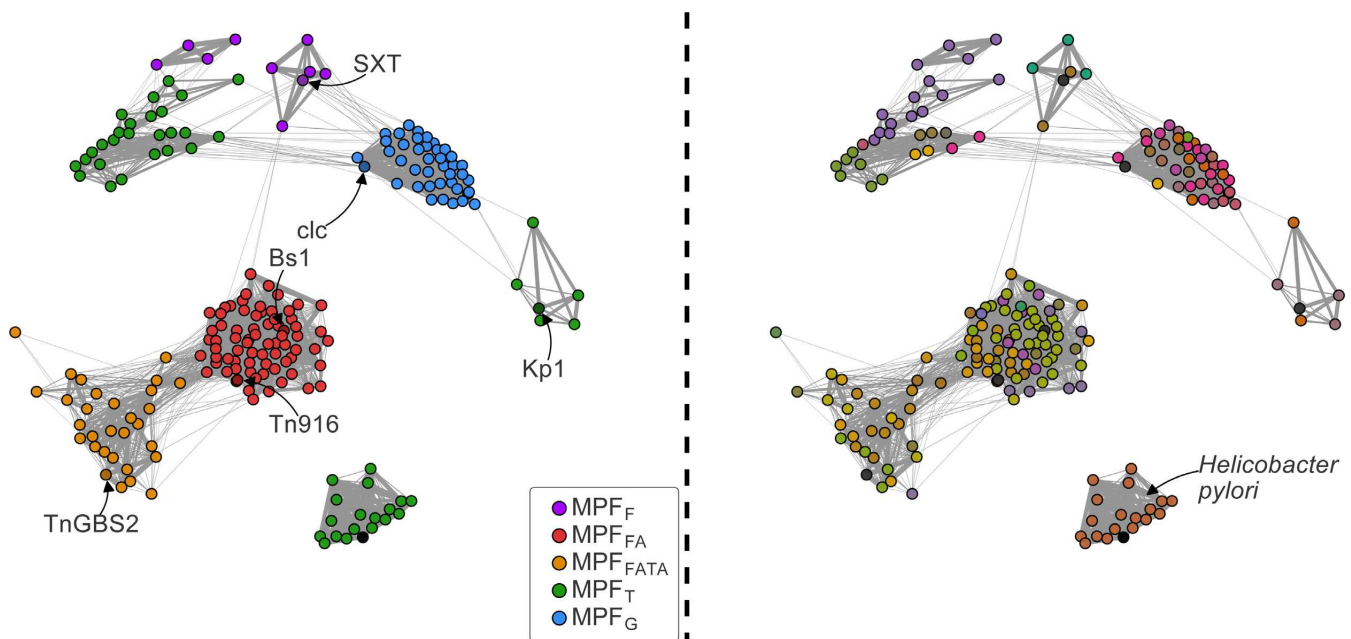
been shown that some ICEs may use relaxases instead of integrases to integrate the chromosome when there is an *oriT* in the genome that can be recognized by the relaxase (69).

A previous analysis using integrases from the tyrosine recombinase family of genomic islands, phages, and six ICEs (of which four of the SXT family), showed that ICE integrases clustered separately (70). However, doubts have been casted on this analysis because of the small number of ICEs that had been used (71). We have thus made a phylogenetic tree of tyrosine recombinases from the ICEs encoding one single integrase (Figure 4). We analyzed a total of 237 tyrosine recombinases from ICEs and other elements including integrons, four different types of recombinases involved in chromosome dimer resolution (XerCD, XerS, XerH), pathogenicity islands, and phages (see Methods, Supplementary Table S2). This tree showed that the Xer recombinases and the integron integrases were all monophyletic. In contrast, genomic islands and ICEs were scattered in the tree. Even ICEs of similar MPF types are systematically paraphyletic. A particularly striking example is provided by a clade in the tree (arc in Figure 4) that contains integrases from several phages, and two types of ICE (Type G and type T) from different species (*K. pneumoniae* and *P. fluorescens*), as well as three different groups of pathogenicity islands. The clear paraphyly of the integrases at the level of MPF types suggests that conjugative elements often exchange the key genes allowing chromosomal integration with otherwise unrelated mobile genetic elements.

### The functional repertoires of ICEs

We investigated the functional classification of the genes in ICEs in relation to those in the rest of the host chromosome using the EggNOG database (see Methods). Unknown or unannotated functions accounted for 61% of all genes in ICEs, showing how much remains to be known about the functions carried by these elements. We observed three functional categories that were systematically more frequent in ICEs ( $P$ -value < 0.01 with Bonferroni correction for multiple test, Figure 5), including typical ICE functions: secretion (genes related to conjugation), replication/recombination/repair (integrases, relaxases, and transposable elements), and to a lesser extent cell cycle control/cell division/chromosome partitioning. The category associated with transcription (gene expression regulation) showed similar frequencies in the ICE and in the host chromosome. Most functions were systematically less frequent in ICEs. Removing from the analysis the proteins implicated in conjugation did not reveal novel families over-represented in ICEs (Supplementary Figure S10).

Since the functional analysis of ICEs pinpointed an over-representation of functions typical of plasmids, we developed more specific approaches to characterize them (see Materials and Methods). We identified 23 partition systems, 13 from type Ia, five of type Ib, and five of type II (no type III). The presence of partition systems suggests the existence of replication systems (even if relaxases can themselves be implicated in ICE replication (25,72)). We found 16 ICEs with proteins predicted to be associated with theta replication (15 in MPF<sub>T</sub> and one in MPF<sub>G</sub>) and two associated with rolling circle replication (all in MPF<sub>FATA</sub>). In-



**Figure 3.** Representation of the wGRR-based network of ICEs. The nodes represent the ICEs and the edges link pairs of ICEs with wGRR score  $>5\%$  (the thickness of the edge is proportional to the score). **Left.** Nodes are colored according to the MPF type. Darker nodes represent ICEs commonly used as experimental models, and are indicated by an arrow. **Right.** Nodes are colored according to the species of the host to highlight the distribution of the 37 species. The information of the species and type are in Supplementary Table S4. The position of the point has been determined by the Fruchterman-Reingold force-directed algorithm, as implemented in the NetworkX python library (spring layout).

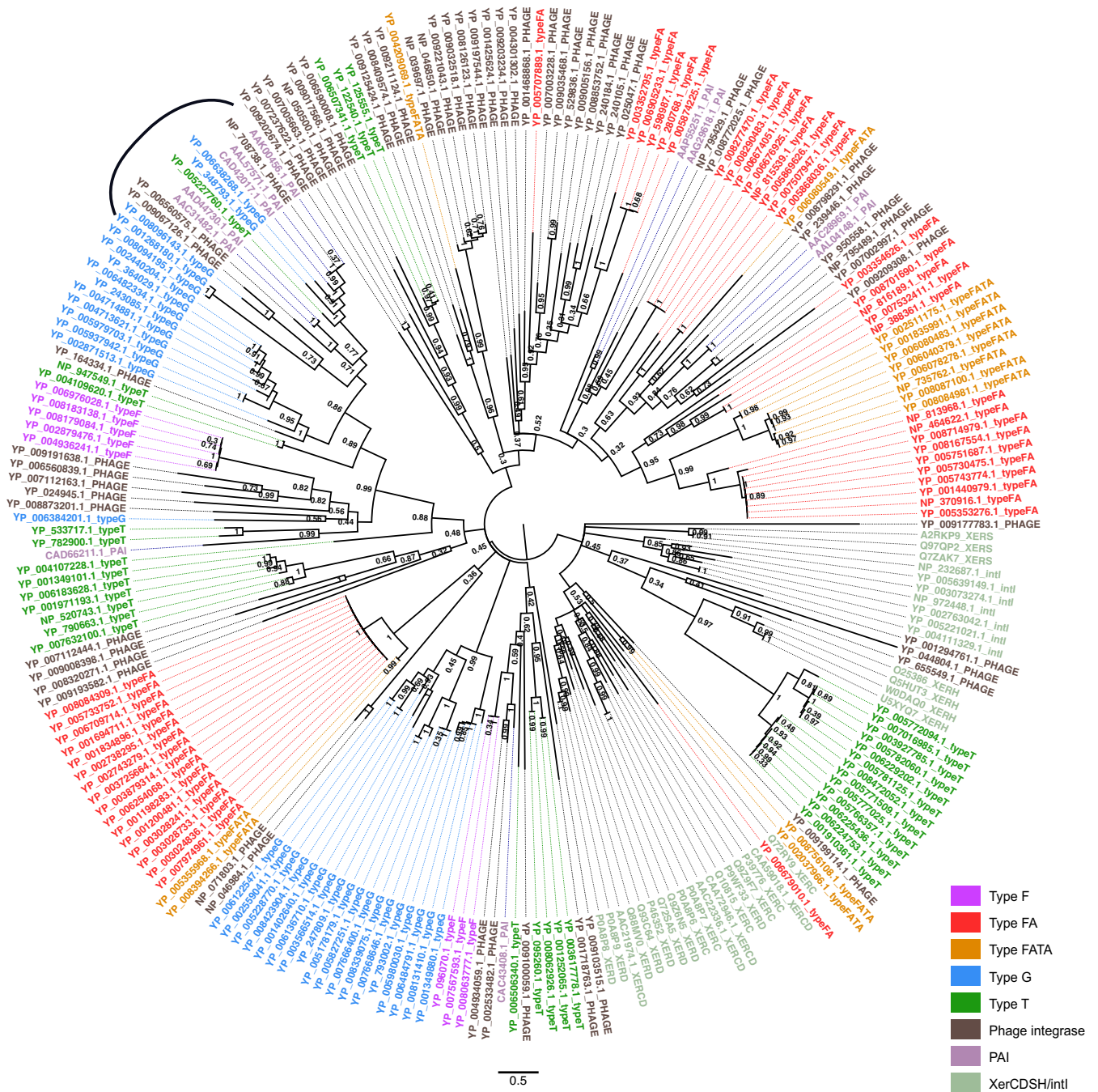
terestingly, all six  $MPF_T$  ICEs with a partition system also encoded a replication protein. Apart from these traits (MPF type, partition and replication systems), these six ICEs are very different (average wGRR score of 34%). To the best of our knowledge ICE replication has not been reported in this family (the most numerous one among complete genomes). Naturally, if some of the relaxases act as replicases, then the actual number of replicases in ICE could be much larger. This might explain why few or no replication systems were found in type F and G although they encode partition systems.

Several ICEs encode accessory functions typical of mobile genetic elements, such as integrons (73), restriction-modification (R-M, (74)), toxin-antitoxin (TA, (75)) and exclusion (43) systems. We identified two integrons in ICEs of the SXT family ( $MPF_F$ ) (as first shown in (76)) and one array of *attC* sites lacking the integron-integrase (CALIN elements (38)) in another ICE. We identified 23 ICEs with at least one R-M system (all four types of R-M systems could be identified). The frequency of R-M systems in ICE (12% of all elements encoded at least one complete system) is similar to that observed in plasmids (10.5%) and higher than in phages (1%). Interestingly, as it was the case in these MGEs (74), the frequency of solitary methylases (25%) was higher than that of the complete systems. These solitary methylases might provide a broad protection from the host R-M systems. Most RNA genes identified in ICE corresponded to intron group II associated RNAs (36% of all detected RNA, excluding those from type G), but  $MPF_G$  ICEs encoded many *radC* and STAXI RNA (representing 80% of RNA in  $MPF_G$ ). Both genes are associated with anti-

restriction functions (77). They might defend the element from R-M systems, thus explaining the relative rarity of the latter in this family of ICEs (7.5% versus 11%). Few ICEs encoded recognizable entry exclusion systems (6%, mostly in  $MPF_T$ ). One ICE contained a type II CRISPR-Cas system in *Legionella pneumophila* str. Paris. Overall, genes encoding many molecular systems associated with plasmid biology could be identified in ICEs, even if some were relatively rare.

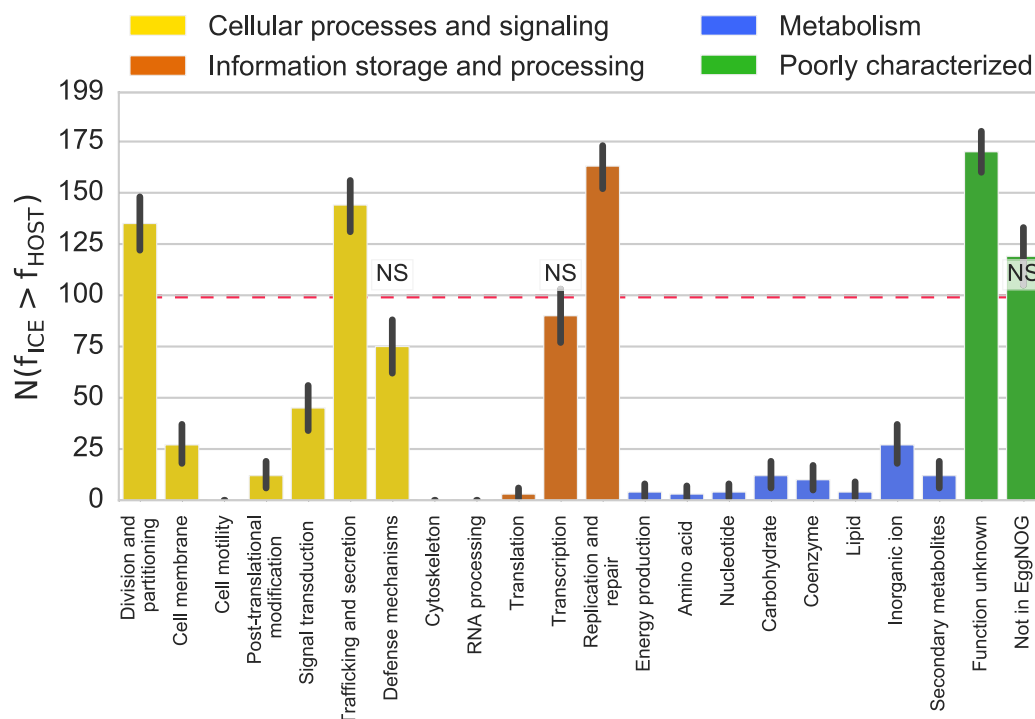
### The organization of ICE

We grouped ICEs by their MPF types and analyzed their genetic organization (Figure 6 and Supplementary Figure S11). We restricted our attention to ICEs with one single integrase of the Serine or Tyrosine recombinase families (141/199), to avoid the inclusion of recombinases with functions unrelated to the integration of the element and to facilitate the representation of the ICE organization. We represented ICEs in such a way that the integrase was located in the first half of the element. Actually, almost 90% of the Tyrosine and Serine recombinases were within the five first percent of the ICE, as expected given their role in the integration of the element. DDE transposases were randomly distributed within the ICE (Kolmogorov-Smirnov test,  $P$ -value = 0.27), which suggests that most of them are not involved in the integration of the ICE. The transposases in the inner parts of the element may be involved in accretion and deletion of parts of the element, and can lead to its occasional integration in spots that would not be targeted otherwise.



**Figure 4.** Phylogenetic tree of tyrosine recombinases. The phylogenetic tree was built using 60 prophage integrases (labelled as ‘...PHAGE’, brown), 11 integrases from pathogenicity islands (‘...PAI’, mauve), 25 XerC,D,S or H (‘...XER...’, greenish grey), 7 integron-integrases (‘...intl’, greenish grey), and 134 integrases from ICEs (colored after the MPF type). The tree was built using Phylobayes and the values represent posterior probabilities support of the partition, with a cut off equal to 0.3, below which the nodes were collapsed because there is insufficient resolution (see Materials and Methods). The black arc denotes a clade with good support, which contains integrases from prophages, PAI and different MPF types, that is explicitly cited in the text.





**Figure 5.** Representation of EggNOG functional categories in ICEs relative to the host chromosome. The bars represent the number of times a given category is found more frequently in an ICE than in its host chromosome ( $N(f_{ICE} > f_{HOST})$ ). The red dotted line represents the expected value under the null hypothesis, where a category is in similar proportion in ICE and its host's chromosome. Bars marked as NS represent a lack of significant difference ( $P > 0.05$ , Binomial test with 199 trials and expected value of 0.5), whereas the others are all significantly different ( $P < 0.05$ , same test). Note that there are 199 trials because one of the 200 ICEs could not be typed and was thus excluded, see text. Error bars represent 95% confidence interval computed with 1000 bootstraps. 'Not in EggNOG' represents the class of genes that didn't match any EggNOG profile.

To facilitate the representation of the organization of ICE, we oriented the genes relative to *virB4*, which was placed on the top strand. Expectedly, given that they are often part of the same operon, the remaining components of the T4SS were usually found in a single locus and almost always (>96%) on the same strand as *virB4* (Supplementary Figure S11A). The T4SS locus spanned, on average, 26% of the ICEs (Figure 6 and Supplementary Figure S12). As observed in plasmids (7), the relaxase (MOB) gene was sometimes encoded close to the T4SS genes (MPF<sub>F</sub>, MPF<sub>T</sub>, MPF<sub>FA</sub>) and sometimes apart (MPF<sub>G</sub>, MPF<sub>FATA</sub>). Interestingly, the relaxase and *virB4* were encoded in the same strand in most cases (86%).

Most accessory functions were encoded apart from the T4SS genes, with the known exception of the entry exclusion systems (43) (Supplementary Figure S11C). We showed above that partition and replication functions co-occurred in ICEs. Here, we show that they co-localize within the element. In MPF<sub>G</sub>, they are often found in the edge opposite to the integrase, whereas they are close to the integrase in MPF<sub>F</sub>. The genes classed as 'Metabolism' were also encoded away from the region of the T4SS, and were typically found in discrete modules. Intriguingly, some regions were particularly rich in genes of unknown function, e.g. integrase-proximal regions, and also at the opposite end of the elements in MPF<sub>G</sub> and MPF<sub>F</sub> ICEs (Figure 6). The MPF<sub>T</sub> ICEs were an exception to most of these trends, since their genes were almost uniformly distributed along the ele-

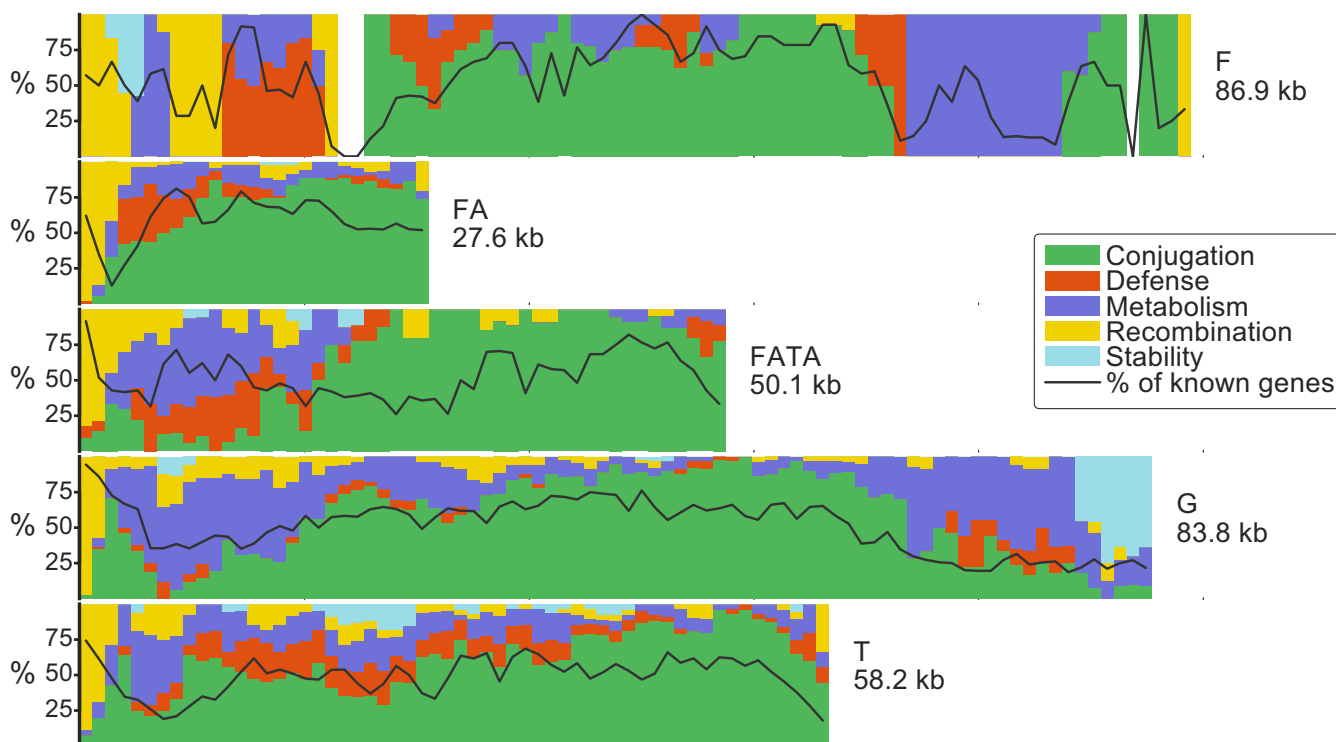
ments. This group may be genetically more diverse than the others, which would explain these results and the scattering of these ICEs in the homology network.

Interestingly, almost all genes in the ICEs were encoded in the strand of *virB4* (>80%), including the RNA genes (Supplementary Figure S11F). Furthermore, genes were predominantly encoded in the leading strand for all types of ICEs but MPF<sub>G</sub> (Supplementary Figure S13). Overall, these results show a certain level of modularity in the organization of ICEs, as previously described in phages and plasmids (78,79), and frequent co-orientation of genes, as identified in different ICE families (SXT (80), Tn916 (81)) and in lambdoid phages (82).

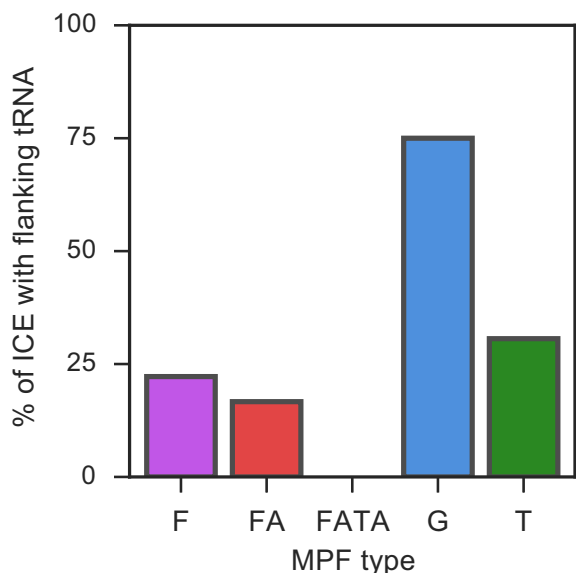
### The chromosomal context of ICE

We analyzed the chromosomal context of ICEs to characterize their integration patterns. The analysis of the two chromosomal genes bordering the elements showed that ICEs are often integrated near hypothetical proteins (52%) or tRNAs (30%). These tRNAs decoded 12 different amino acids, in most cases Leucine, Lysine, and Glycine. The tropism towards integration near a tRNA varied with type of ICE, it was high for MPF<sub>G</sub> (75%) and null for MPF<sub>FATA</sub> (0%) (Figure 7).

We then analyzed the distribution of ICEs in the larger context of the bacterial chromosome. Based on the analysis of 15 ICEs, it had recently been suggested that ICEs



**Figure 6.** Average organization of ICEs. Each row represents an MPF type and has a length proportional to the mean size of the ICEs of the corresponding type. Colors represent different classes of functions. The black line represents the proportion of genes with a predicted function per bin. The classes of functions correspond to those described in detail in Supplementary Figure S9A-B-C-F. More precisely, *Conjugation* includes MPF-associated genes, and the relaxase. *Defense* includes antibiotic resistance genes, restriction modification, solitary methylases. *Metabolism* includes genes annotated by EggNOG as such. *Recombination* includes tyrosine and serine recombinases, and DDE transposases. *Stability* includes replication, partition and entry exclusion systems. Bar heights are proportional to the proportion of genes of a given function at that position among the genes with a predicted function. The width of each bin is 1kb.



**Figure 7.** Proportion of ICEs with flanking tRNAs on either side of the ICE.

would be more frequent close to the origin of replication because they target essential highly conserved genes (26),

which are more frequent near the origin of replication in fast growing bacteria (83). However, we could not find a significant correlation between the frequency of ICEs and their position in the origin-to-terminus axis of replication (Supplementary Figure S13). When we analyzed the chromosomal distribution of ICEs across MPF types, MPF<sub>FATA</sub> were over-abundant in the terminus region ( $\chi^2$ ,  $P$ -value < 0.003), whereas the others didn't show significant trends ( $P$ -value > 0.1, same test). Neither the strand location ( $\chi^2$ ,  $P$ -value = 0.39) nor the size of the ICE (Spearman- $\rho$ ,  $P$ -value = 0.52) were associated with its distance to the origin of replication.

We identified the origins and terminus of replication of genomes and inferred the leading and lagging strand of each gene in ICEs (see Materials and Methods). Most genes were oriented in the same direction as the replication fork (leading strand,  $\chi^2$ ,  $P$ -value =  $10^{-5}$ ), with the exception of MPF<sub>G</sub> ICEs that showed an opposite trend (Supplementary Figure S14). The high frequency of leading strand MPF<sub>FATA</sub> ICEs may be associated with the hosts' genome organization, in this case they are all Firmicutes, because genomes from this phyla show high frequency of genes in the leading strand (84).

**CONCLUSION**

Our work shows that one can identify and delimit ICEs from genome data using comparative genomics. The pre-

cise identification of integration sites and the validation of the functions of these ICEs will require further experimental work by experts on a large number of different species and conjugation systems. In this respect, a current limitation of our approach is the reliance on a set of genomes for a given species. This means that ICEs from poorly covered taxa could not be studied. We have restricted our study to complete genomes to avoid using poor quality data. This is not a restriction of the method: draft genomes can also be analyzed with our method, as long as the ICE is in the same contig as the flanking core genes. Unfortunately, our experience is that the presence of repeats in ICEs, like transposases, often splits these elements in several contigs when genomes are sequenced using short-read technologies. The identification ICEs split in several contigs requires the availability of a very similar ICE for reference, which may bias the study of these elements. The increasing use of long-read technologies to sequence bacterial genomes will soon solve this limitation.

Even if our dataset is representative of the diversity of experimentally studied ICEs, we lacked ICEs from types B (Bacteroidetes), of which there are experimental systems (85), and C (Cyanobacteria), for which there is no experimental system. Further data will be necessary to study these elements. Another limitation of this work is the assumption that the presence of a certain number of components of the T4SS system and a relaxase are necessary and sufficient to define an ICE. While our previous studies have shown that we are able to identify known conjugative systems accurately, we cannot exclude the possibility that some of the identified ICE are defective for transfer. This may explain the existence of some elements that lack identifiable integrases. In spite of these limitations, the availability for the first time of large dataset of ICEs having undergone a systematic expert curation allowed to quantify many traits associated with ICEs and confirm, and sometimes infirm, observations made from the small number of well-known ICE models. It also allowed to characterize their genetic organization, and identify common traits.

Integration and conjugation are the only functions that we found to be present in most ICEs. Interestingly, even these functions had very different phylogenetic histories, as revealed by the scattered distribution of ICEs per MPF type in the phylogenetic tree of the tyrosine recombinases. A number of other functions were often identified in some types of ICE, notably defense systems, partition, and replication. Collectively, they reinforce the suggestions of a thin line separating ICEs from conjugative plasmids (26). The analysis of the genetic organization of ICEs suggests that they are organized in functional modules. Together, these results suggest that ICEs are highly modular, which may contribute to the evolution of their gene repertoires by genetic exchange between elements, as previously observed in temperate phages (78). If so, our data suggests that either ICEs tend to recombine more with elements of the same MPF type, or that the fitness of the products of recombination tends to be higher when recombination takes place within ICE of the same type (e.g. for functional reasons).

## AVAILABILITY

The program to identify conjugative systems is available on [https://github.com/gem-pasteur/Macsfinder\\_models](https://github.com/gem-pasteur/Macsfinder_models)

The webserver is hosted on: <https://galaxy.pasteur.fr/> > Search > CONJScan

The program and data to make representations like those of Supplementary Figure S1 is available at [https://gitlab.pasteur.fr/gem/spot\\_ICE](https://gitlab.pasteur.fr/gem/spot_ICE).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

J.C. is a member of the « École Doctorale Frontière du Vivant (FdV) – Programme Bettencourt ». We thank Christine Citti for insightful comments on a previous version of this manuscript, Fernando de la Cruz for insights and providing the list of rep proteins of PLACNET, Bertrand Néron and Olivia Doppelt-Azeroual for setting up the webserver version of CONJscan.

*Author contributions:* J.C. and E.P.C.R. designed the study. J.C. and M.T. produced the data. J.C. made the analysis. J.C. and E.P.C.R. drafted the manuscript. All authors contributed to the final text of the manuscript.

## FUNDING

European Research Council [EVOMOBILOME, 281605 to E.P.C.R.]. Funding for open access charge: ERC EVOMOBILOME.

*Conflict of interest statement.* None declared.

## REFERENCES

- Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
- Popa, O. and Dagan, T. (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.*, **14**, 615–623.
- Polz, M.F., Alm, E.J. and Hanage, W.P. (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.*, **29**, 170–175.
- Medini, D., Donati, C., Tettelin, H., Massignani, V. and Rappuoli, R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
- Davies, J. and Davies, D. (2010) Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.: MMBR*, **74**, 417–433.
- Boyd, E.F. and Brussow, H. (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.*, **10**, 521–529.
- Smillie, C., Pinar Garcillan-Barcia, M., Victoria Francia, M., Rocha, E.P.C. and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
- Burrus, V., Pavlovic, G., Decaris, B. and Guedon, G. (2002) Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.*, **46**, 601–610.
- Johnson, C.M. and Grossman, A.D. (2015) Integrative and conjugative elements (ICEs): what they do and how they work. *Annu. Rev. Genet.*, **49**, 577–601.
- Ghinet, M.G., Bordeleau, E., Beaudin, J., Brzezinski, R., Roy, S. and Burrus, V. (2011) Uncovering the prevalence and diversity of integrating conjugative elements in Actinobacteria. *PLoS ONE*, **6**, e27846.
- Goessweiner-Mohr, N., Arends, K., Keller, W. and Grohmann, E. (2013) Conjugative type IV secretion systems in Gram-positive bacteria. *Plasmid*, **70**, 289–302.

12. Wang, H. and Mullany, P. (2000) The large resolvase TndX is required and sufficient for integration and excision of derivatives of the novel conjugative transposon Tn5397. *J. Bacteriol.*, **182**, 6577–6583.
13. Brochet, M., Da Cunha, V., Couve, E., Rusniok, C., Trieu-Cuot, P. and Glaser, P. (2009) Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. *Mol. Microbiol.*, **71**, 948–959.
14. Dordet Frisoni, E., Marena, M.S., Sagne, E., Nouvel, L.X., Guerillot, R., Glaser, P., Blanchard, A., Tardy, F., Sirand-Pugnet, P., Baranowski, E. et al. (2013) ICEA of *Mycoplasma agalactiae*: a new family of self-transmissible integrative elements that confers conjugative properties to the recipient strain. *Mol. Microbiol.*, **89**, 1226–1239.
15. Guglielmini, J., de la Cruz, F. and Rocha, E.P.C. (2013) Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.*, **30**, 315–331.
16. Guglielmini, J., Neron, B., Abby, S.S., Garcillan-Barcia, M.P., la Cruz, F.D. and Rocha, E.P. (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.*, **42**, 5715–5727.
17. Guglielmini, J., Quintais, L., Pilar Garcillan-Barcia, M., de la Cruz, F. and Rocha, E.P.C. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
18. Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W. and Crook, D.W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.*, **33**, 376–393.
19. Roberts, A.P. and Mullany, P. (2011) Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol. Rev.*, **35**, 856–871.
20. Carraro, N. and Burrus, V. (2014) Biology of Three ICE Families: SXT/R391, ICEBs1, and ICES1/ICES3. *Microbiol. Spectr.*, **2**, doi:10.1128/microbiolspec.MDNA3-0008-2014.
21. Auchtung, J.M., Aleksanyan, N., Bulku, A. and Berkmen, M.B. (2016) Biology of ICEBs1, an integrative and conjugative element in *Bacillus subtilis*. *Plasmid*, **86**, 14–25.
22. Wang, J., Wang, G.R., Shoemaker, N.B. and Salyers, A.A. (2001) Production of two proteins encoded by the *Bacteroides mobilizable* transposon NBU1 correlates with time-dependent accumulation of the excised NBU1 circular form. *J. Bacteriol.*, **183**, 6335–6343.
23. Lee, C.A., Babic, A. and Grossman, A.D. (2010) Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.*, **75**, 268–279.
24. Grohmann, E. (2010) Autonomous plasmid-like replication of *Bacillus* ICEBs1: a general feature of integrative conjugative elements? *Mol. Microbiol.*, **75**, 261–263.
25. Carraro, N., Poulin, D. and Burrus, V. (2015) 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. *PLoS Genet.*, **11**, e1005298.
26. Carraro, N. and Burrus, V. (2015) The dualistic nature of integrative and conjugative elements. *Mobile Genet. Elem.*, **5**, 98–102.
27. Garriss, G., Waldor, M.K. and Burrus, V. (2009) Mobile antibiotic resistance encoding elements promote their own diversity. *PLoS Genet.*, **5**, e1000775.
28. Guerillot, R., Siguier, P., Gourbeyre, E., Chandler, M. and Glaser, P. (2014) The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. *Genome Biol. Evol.*, **6**, 260–272.
29. Abby, S.S., Neron, B., Menager, H., Touchon, M. and Rocha, E.P. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
30. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
31. Touchon, M., Cury, J., Yoon, E.-J., Krizova, L., Cerqueira, G.C., Murphy, C., Feldgarden, M., Wortman, J., Clermont, D., Lambert, T. et al. (2014) The genomic diversification of the whole *Acinetobacter* genus: origins, mechanisms, and consequences. *Genome Biol. Evol.*, **6**, 2866–2882.
32. Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
33. Gibson, M.K., Forsberg, K.J. and Dantas, G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
34. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
35. Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J. et al. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
36. Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
37. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J. et al. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
38. Cury, J., Jove, T., Touchon, M., Neron, B. and Rocha, E.P. (2016) Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.*, **44**, 4539–4550.
39. Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A. et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
40. Gerdes, K., Moller-Jensen, J. and Bugge Jensen, R. (2000) Plasmid and chromosome partitioning: surprises from phylogeny. *Mol. Microbiol.*, **37**, 455–466.
41. Larsen, R.A., Cusumano, C., Fujioka, A., Lim-Fong, G., Patterson, P. and Pogliano, J. (2007) Treadmilling of a prokaryotic tubulin-like protein, TubZ, required for plasmid stability in *Bacillus thuringiensis*. *Genes Dev.*, **21**, 1340–1352.
42. Lanza, V.F., de Toro, M., Garcillan-Barcia, M.P., Mora, A., Blanco, J., Coque, T.M. and de la Cruz, F. (2014) Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.*, **10**, e1004766.
43. Garcillan-Barcia, M.P. and de la Cruz, F. (2008) Why is entry exclusion an essential feature of conjugative plasmids? *Plasmid*, **60**, 1–18.
44. 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.
45. Gouy, M., Guindon, S. and Gascuel, O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
46. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
47. Bobay, L.-M., Rocha, E.P.C. and Touchon, M. (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.*, **30**, 737–751.
48. Yoon, S.H., Hur, C.G., Kang, H.Y., Kim, Y.H., Oh, T.K. and Kim, J.F. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics*, **6**, 184.
49. Nunes-Duby, S.E., Kwon, H.J., Tirumalai, R.S., Ellenberger, T. and Landy, A. (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res.*, **26**, 391–406.
50. Le Bourgeois, P., Bugarel, M., Campo, N., Daveran-Mingot, M.L., Labonte, J., Lanfranchi, D., Lautier, T., Pages, C. and Ritzenthaler, P. (2007) The unconventional Xer recombination machinery of *Streptococci/Lactococci*. *PLoS Genet.*, **3**, e117.
51. Debowski, A.W., Carnoy, C., Verbrugghe, P., Nilsson, H.O., Gauntlett, J.C., Fulurija, A., Camilleri, T., Berg, D.E., Marshall, B.J. and Benghezal, M. (2012) Xer recombinase and genome integrity in *Helicobacter pylori*, a pathogen without topoisomerase IV. *PLoS One*, **7**, e33310.
52. Leroux, M., Rezoug, Z. and Szatmari, G. (2013) The Xer/dif site-specific recombination system of *Campylobacter jejuni*. *Mol. Genet. Genomics*, **288**, 495–502.
53. Sela, I., Ashkenazy, H., Katoh, K. and Pupko, T. (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.*, **43**, W7–W14.
54. Lartillot, N., Rodrigue, N., Stubbs, D. and Richer, J. (2013) PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, **62**, 611–615.

55. Gao, F., Luo, H. and Zhang, C.T. (2013) Doric 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.*, **41**, D90–D93.
56. Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.
57. Abby, S.S., Cury, J., Guglielmini, J., Neron, B., Touchon, M. and Rocha, E.P. (2016) Identification of protein secretion systems in bacterial genomes. *Scientific Rep.*, **6**, 23080.
58. Burrus, V. and Waldor, M.K. (2004) Formation of SXT tandem arrays and SXT-R391 hybrids. *J. Bacteriol.*, **186**, 2636–2645.
59. Paauw, A., Leverstein-van Hall, M.A., Verhoef, J. and Fluit, A.C. (2010) Evolution in quantum leaps: multiple combinatorial transfers of HPI and other genetic modules in Enterobacteriaceae. *PLoS One*, **5**, e8662.
60. Lechner, M., Schmitt, K., Bauer, S., Hot, D., Hubans, C., Levillain, E., Loch, C., Lemoine, Y. and Gross, R. (2009) Genomic island excisions in *Bordetella petrii*. *BMC Microbiol.*, **9**, 141.
61. Bellanger, X., Morel, C., Gonot, F., Puymège, A., Decaris, B. and Guedon, G. (2011) Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. *Mol. Microbiol.*, **81**, 912–925.
62. Schijffelen, M.J., Boel, C.H., van Strijp, J.A. and Fluit, A.C. (2010) Whole genome analysis of a livestock-associated methicillin-resistant *Staphylococcus aureus* ST398 isolate from a case of human endocarditis. *BMC Genomics*, **11**, 376.
63. Ramsay, J.P., Sullivan, J.T., Stuart, G.S., Lamont, I.L. and Ronson, C.W. (2006) Excision and transfer of the *Mesorhizobium loti* R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. *Mol. Microbiol.*, **62**, 723–734.
64. Rocha, E. and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**, 291–294.
65. Touchon, M., Bobay, L.M. and Rocha, E.P. (2014) The chromosomal accommodation and domestication of mobile genetic elements. *Curr. Opin. Microbiol.*, **22**, 22–29.
66. Oliveira, P.H., Touchon, M. and Rocha, E.P. (2016) Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5658–5663.
67. Bobay, L.M., Touchon, M. and Rocha, E.P. (2014) Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 12127–12132.
68. Gal-Mor, O. and Finlay, B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol.*, **8**, 1707–1719.
69. Agundez, L., Gonzalez-Prieto, C., Machon, C. and Llosa, M. (2012) Site-specific integration of foreign DNA into minimal bacterial and human target sequences mediated by a conjugative relaxase. *PLoS One*, **7**, e31047.
70. Boyd, E.F., Almagro-Moreno, S. and Parent, M.A. (2009) Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.*, **17**, 47–53.
71. Bellanger, X., Payot, S., Leblond-Bourget, N. and Guedon, G. (2014) Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.*, **38**, 720–760.
72. Thomas, J., Lee, C.A. and Grossman, A.D. (2013) A conserved helicase processivity factor is needed for conjugation and replication of an integrative and conjugative element. *PLoS Genet.*, **9**, e1003198.
73. Iwanaga, M., Toma, C., Miyazato, T., Insiengmay, S., Nakasone, N. and Ehara, M. (2004) Antibiotic resistance conferred by a class I integron and SXT constin in *Vibrio cholerae* O1 strains isolated in Laos. *Antimicrob. Agents Chemother.*, **48**, 2364–2369.
74. Oliveira, P.H., Touchon, M. and Rocha, E.P. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 10618–10631.
75. Wozniak, R.A. and Waldor, M.K. (2009) A toxin-antitoxin system promotes the maintenance of an integrative conjugative element. *PLoS Genet.*, **5**, e1000439.
76. Hochhut, B., Lotfi, Y., Mazel, D., Faruque, S.M., Woodgate, R. and Waldor, M.K. (2001) Molecular analysis of antibiotic resistance gene clusters in *Vibrio cholerae* O139 and O1 SXT constins. *Antimicrob. Agents Chemother.*, **45**, 2991–3000.
77. Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H. and Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, R31.
78. Botstein, D. (1980) A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.*, **354**, 484–490.
79. Toussaint, A. and Merlin, C. (2002) Mobile elements as a combination of functional modules. *Plasmid*, **47**, 26–35.
80. Beaber, J.W., Hochhut, B. and Waldor, M.K. (2002) Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *J. Bacteriol.*, **184**, 4259–4269.
81. Clewell, D.B., Flannagan, S.E. and Jaworski, D.D. (1995) Unconstrained bacterial promiscuity: the Tn916-Tn1545 family of conjugative transposons. *Trends Microbiol.*, **3**, 229–236.
82. Campbell, A.M. (2002) Preferential orientation of natural lambdaoid prophages and bacterial chromosome organization. *Theor. Popul. Biol.*, **61**, 503–507.
83. Couturier, E. and Rocha, E. (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.*, **59**, 1506–1518.
84. Rocha, E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.*, **10**, 393–395.
85. Bonheyo, G.T., Hund, B.D., Shoemaker, N.B. and Salyers, A.A. (2001) Transfer region of a *Bacteroides* conjugative transposon contains regulatory as well as structural genes. *Plasmid*, **46**, 202–209.

### **1.3.2 Article 3: Host range expansion and genetic plasticity drive the trade-off between integrative and extra-chromosomal mobile genetic elements**

This study aims at answering the broader ecological question on why two forms of mobile genetic elements exist and persist. This is particularly important as mobile genetic elements are major driver of bacterial evolution.

## Host range expansion and genetic plasticity drive the trade-off between integrative and extrachromosomal mobile genetic elements

### Abstract

Self-transmissible mobile genetic elements drive horizontal gene transfer between prokaryotes. Some of these elements integrate a replicon within the cell, typically the chromosome, whereas other elements replicate autonomously as plasmids. Recent works showed the existence of few differences, and occasional interconversion, between these types of conjugative elements. Here, we enquired on why evolutionary processes have maintained the two types of mobile genetic elements. For this, we compared integrative (ICE) with extrachromosomal (plasmid) conjugative elements of the highly abundant MPF<sub>T</sub> type. We hypothesized that plasmids might be more plastic, because their transfer in the genome does not disrupt chromosome structure and organization, and indeed they show much higher variations in size and inter-MGE gene exchange. We hypothesized that ICEs would be able to stably transfer between more distant clades because plasmid replication is the most limiting factor of their host range. Accordingly, transfers between distant taxa occurred more frequently in ICEs. Interestingly, when an ICE and a plasmid were very similar but present in very distant hosts, the ICE was the one most often transferred to the distant host. Hence, stabilization of a plasmid in a novel host may be facilitated by its integration in the chromosome as an ICE. We found differences in the types of accessory genes carried by the two types of elements and speculate that certain functions may favor certain types of elements in function of their plasticity and transmissibility.

### Introduction

Genomes are composed of chromosome(s) and extrachromosomal elements. In prokaryotes, the latter are usually the result of horizontal gene transfer and encode non-essential but ecologically important traits such as those necessary for interactions with eukaryotes (1, 2). Conjugative plasmids and extrachromosomal prophages replicate autonomously in the cell using replicases to recruit the bacterial DNA replication machinery (or using their own). These mobile elements are able to transfer themselves between cells, plasmids through conjugation and prophages by encapsidating copies of their genomes in virions. These mobile elements have developed ways of increasing their stability in cellular lineages after transfer, by encoding partition systems that ensure proper segregation during bacterial replication (3), resolution systems that prevent accumulation of multimers (4) and poison-antidote systems that lead to post-segregation killing of their hosts (5). Additionally, some horizontally transferred mobile genetic elements integrate the chromosome. This is the case of the vast majority of known prophages, of most conjugative elements, and of many elements with poorly characterized mechanisms of genetic mobility (*e.g.*, most pathogenicity islands)(6–8). Integrative elements are replicated with the host chromosome and require an additional step of excision before being transferred between cells. This raises a question that has not been addressed in the literature: what are the relative benefits and disadvantages of the integrated and extrachromosomal forms?

To address this question, we decided to study the differences and similarities between integrative and extrachromosomal conjugative elements. We focused on these elements because both forms are frequently found in bacteria, they can be easily detected in genomes, and the mechanism of conjugation is well known. Conjugative elements are important contributors for horizontal gene transfer and have a crucial role in spreading antibiotic resistance and virulence genes among bacterial pathogens (9–12). Recently, several works suggested that the line separating integrative conjugative elements (ICE) and conjugative plasmids (CP) could be thinner than anticipated (13), because some ICEs encode plasmid-associated functions like replication (14) or partition systems (15), some plasmids encode integrases (16), and the phylogenetic tree of conjugative systems shows intermingled distributions of ICE and CPs (17). Finally, both forms – ICE and CP - are found



throughout the bacterial kingdom, but their relative frequency depends on the taxa and on the mechanisms of conjugation (7). These differences suggest that conjugative elements endure diverse selective pressures for being integrative or extrachromosomal depending on unknown environmental, genetic, or physiological variables.

We thought that the key differences in the biology of integrative and extrachromosomal elements might illuminate their specific advantages. ICEs require the extra step of integration/excision during transfer, which may slow-down the process of transfer and require extra genetic regulation. Their integration in the chromosome may affect the latter's organization and structure, and these deleterious effects might increase with the size of the element. On the other hand, ICEs replicate with chromosomes and could thus be lost from the cell at lower rates than plasmids. Furthermore, plasmids must recruit the host replication machinery, which may pose problems of compatibility between elements and is regarded as a major constrain in terms of host range: many plasmids are able to conjugate into distantly related hosts, but unable to replicate there (18, 19). We thus hypothesize that ICEs might be favored when transfers occur between distant hosts, whereas plasmids might provide more genetic plasticity because their size is not constrained by chromosomal organization.

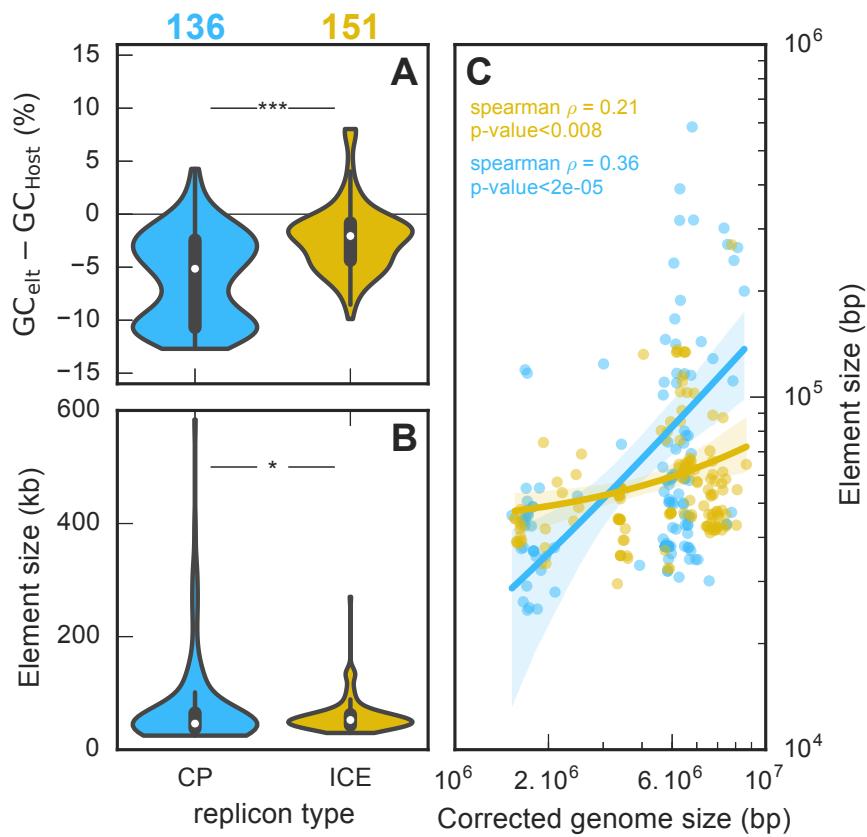
Here, we present a study of conjugative elements of the type  $MPF_T$ . This is the most frequent and best-studied type of conjugative systems in bacteria, and the only one for which we can identify hundreds of elements of each of the forms (ICEs and CPs). We restricted our analysis to genera with both CPs and ICEs, to avoid, as much as possible, taxonomical biases. We will first describe the content of both types of elements and highlight their differences and similarities. Next, we will illustrate the relations of genetic similarity and the genetic exchanges between them. Finally, we will use this data to test if chromosomal integration facilitates colonization of novel taxa by mobile elements.

### Results

#### Functional and genetic differences between ICEs and CPs

We analyzed a dataset of 151 ICEs and 136 conjugative plasmids of the same genera and of type MPF<sub>T</sub>, most of which were from Proteobacteria (96.9%). Both ICEs and CPs were found to be AT-richer than their host chromosomes, which is common in mobile genetic elements and horizontally transferred genes (40), but this difference was three times smaller for ICEs (Figure 1A). Conjugative plasmids are slightly larger than ICEs in average (75kb vs 59kb), and slightly smaller in median (46kb vs 52kb). In spite of this, CPs have much more diverse sizes than ICEs (Figure 1B), showing a coefficient of variation twice as large (1.05 vs 0.49). The size of the conjugative elements depends on the size of the bacterial genome (after discounting the size of their conjugative elements), with much sharper increase on plasmids than on ICEs (Figure 1C). Hence, conjugative plasmids are A+T richer, account for a majority of the smallest and largest conjugative elements, and their size co-varies more steeply with the host genome size.

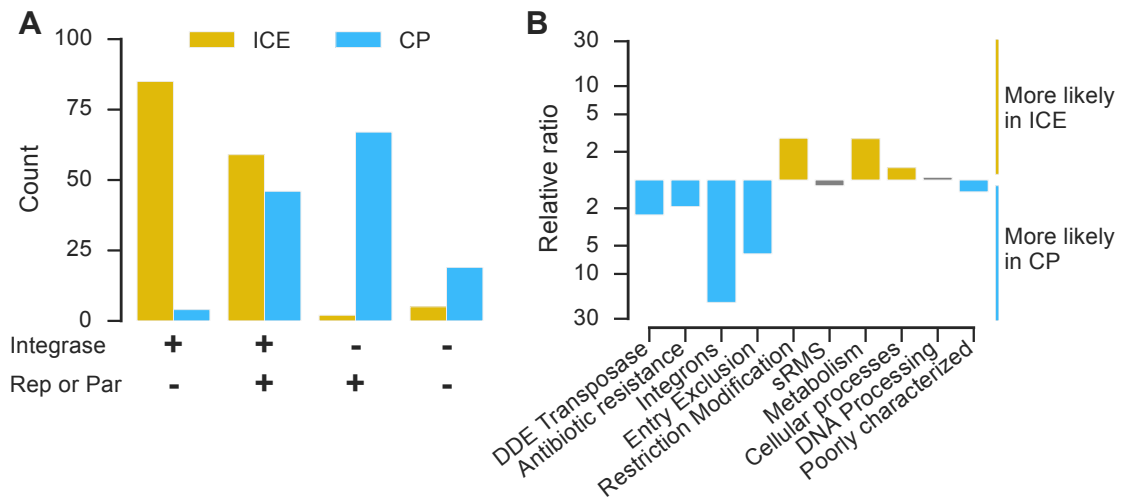
The previous results suggest that conjugative plasmids endure more frequent gene gain and loss than ICEs. It has been suggested that HGT concentrates in few locations – hotspots – in bacterial chromosomes, to minimize disruption in their organization. We used HTg50, a measure of the concentration of HGT in chromosomes that corresponds to the minimal number of spots required to account for 50% of the horizontally transferred genes (HTg) (23), to test if the higher plasticity of plasmids facilitates HGT in genomes. Lower HTg50 indicates higher concentration of transferred genes in a few locations and thus suggests the presence of stronger organizational constraints. Indeed, there was a negative association between the number of plasmids weighted by their size and the chromosomes HTg50 (Spearman  $\rho=-0.35$ ,  $p$ -value=0.0016, Figure S1).



**Figure 1:** Comparisons between ICEs (yellow) and CPs (blue) compositions and sizes. **A.** Distribution of the GC% difference between the element and the host for CPs and ICEs. Elements are AT-rich below the horizontal line. The distributions are significantly different (Wilcoxon rank sum test,  $p\text{-value} < 10^{-3}$ ); ICEs are AT-rich relative to CPs. **B.** Distribution of the size of the elements. The distributions are significantly different (Wilcoxon rank sum test,  $p\text{-value} < 0.05$ ); ICEs are more often larger than CPs (median of ICE (52.5 kb) > median of CP (46.1 kb)), although the means shows the opposite trend (mean of ICE (59 kb) < mean of CP (74.6 kb)). **C.** Size of the element as a function of the one of its host. The host's genome includes all of its replicons, but the "Corrected genome size" discounts the size of the mobile element in the comparison.

We then searched to quantify the differences between integrative and extrachromosomal conjugative elements in a wide range of functions associated with the biology of the two forms of conjugative systems, with a focus on stabilization functions. Expectedly, ICEs had higher frequency of integrases and CPs had more frequently identifiable partition and replication systems. Yet, some ICEs encoded partition systems (11%) and many encoded a replicase (40%), while 37% of CPs encoded at least one Tyrosine or Serine recombinase (Figure 2A). This illustrates a certain continuum between the two types of elements. In the

extremes, one finds elements having integrases and lacking replication and partition functions (mostly ICEs) or its opposite (plasmids). In between these extremes, about half of the elements (40% and 48%, ICE and CP, respectively) have functions usually associated with the other type of element and may (rarely) lack functions typically associated with its own type (Figure 2B).



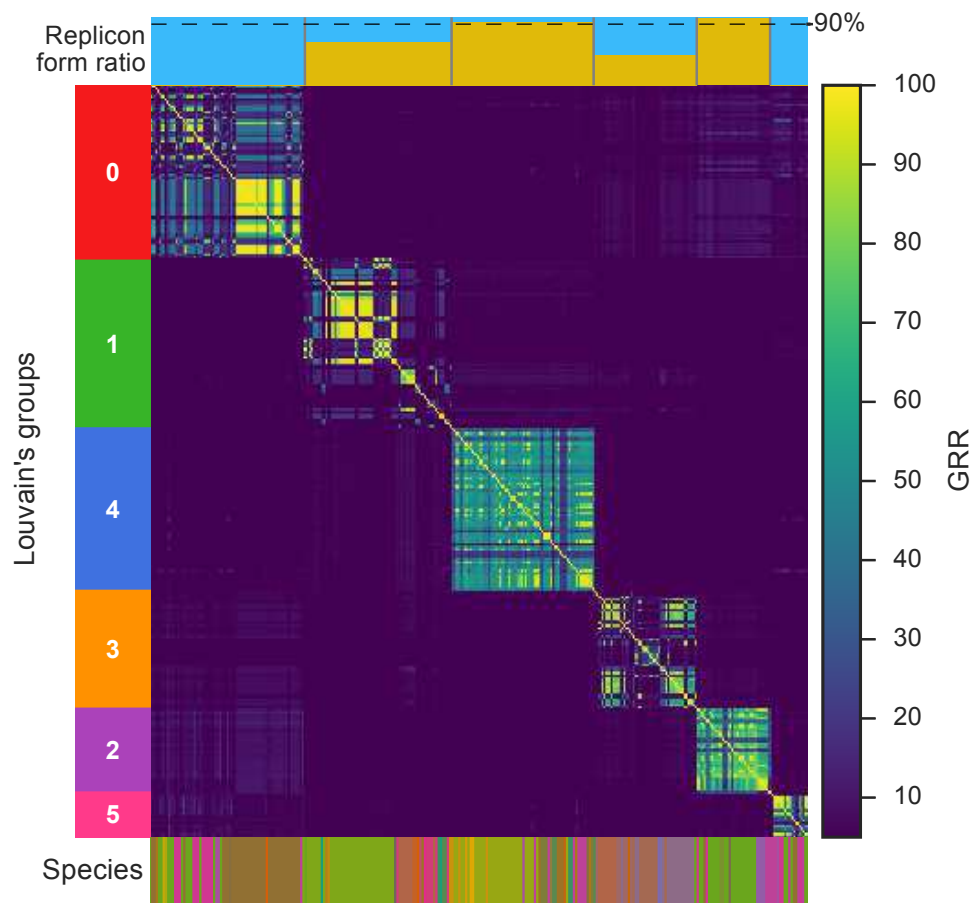
**Figure 2:** Difference and similarities in term of functions ICEs and CP carry **A.** Number of elements encoding (+) or lacking (-) a replication (Rep) or partition (par) systems ("Rep or Par") or an integrase. **B.** Relative ratio of key functions in ICEs relative to the expectation if they were the same as in CPs. Values above the main horizontal line represent functions over-represented in ICE and those below represent functions under-represented in ICE. The bars are colored when the Fisher exact test allowed a significant rejection of the null hypothesis, and grey otherwise (p-value < 0.05 after Bonferroni-Holm correction for multiple tests).

Some incompatibility groups are known to be associated with broad or with narrow plasmid host-range (36). Since some ICEs encode replication proteins, we searched to define the incompatibility groups of both types of elements. IncX CPs were the most frequent (26% of plasmids), and all the other groups were present at frequencies below 10%. Importantly, we could not find an incompatibility type for any of the ICEs, nor for 45% of the CPs. Finally, the ICEs and CP harbored different types of MOB, ( $\chi^2$  on contingency table, p-value <  $10^{-10}$ ), although they both have a majority of MOB P1.

We then made similar analyses for functions usually regarded as accessory or unrelated to the biology of mobile elements (Figure 2B). ICEs were more likely to carry restriction-modification systems (x2.8) than CPs (but not solitary methylases). In contrast, they were significantly less likely to carry antibiotic resistance genes, integrons, or entry-exclusion systems. To have a broader view of the element's functional repertoires, we classed the genes in the four major functional categories of the EggNOG database. We found that ICEs had relatively more genes encoding metabolic and cellular processes, and fewer of unknown or poorly characterized functions. The latter accounted for about 46% of genes in ICEs and a striking 61% of the genes in conjugative plasmids. We have previously shown that genes of unknown function were over-represented in ICEs relative to their host chromosome (22). The present results show that this over-representation is even more pronounced in CPs. Finally, we enquired on the density of repeats that are often associated with MGE. CPs are four times as much denser in repeats than ICEs on average (0.30 vs 0.078 repeats per kb,  $p$ -value  $< 10^{-10}$ , Wilcoxon rank-sum test). Hence, both types of elements have many functions in common, but their relative frequency often differs significantly.

#### Genetic similarities between ICEs and CPs

The results of the previous section, together with previously published studies (see Introduction), suggest that ICEs and conjugative plasmids either share a common history or often exchange genes (or both). We detailed the relationships of homology between ICEs and CPs using a weighted Gene Repertoire Relatedness (wGRR) measure of similarity. The wGRR is the frequency of bi-directional best hits between two elements weighted by their sequence similarity (see Methods). We clustered the matrix of wGRR using the well-known Louvain algorithm (38), and found six well-distinguished groups (Figure 3). Two groups (0 and 5) are only constituted of CPs, two are composed of more than 90% of ICEs (4 and 2) and two have a mix of both types of elements (1 and 3) (Figure 3, top bar). Bacterial species are scattered between groups, showing that they are not the key determinant of the clustering (bottom bar). Instead, groups showed differences in their functional repertoires (Figure 4). Group 3, includes many ICEs and CPs, where all ICEs have integrases while more than half of the CPs lack both replication and partition systems. In contrast, almost all ICEs of groups 1 and 2 encode an integrase and all CPs have partition or replication systems.



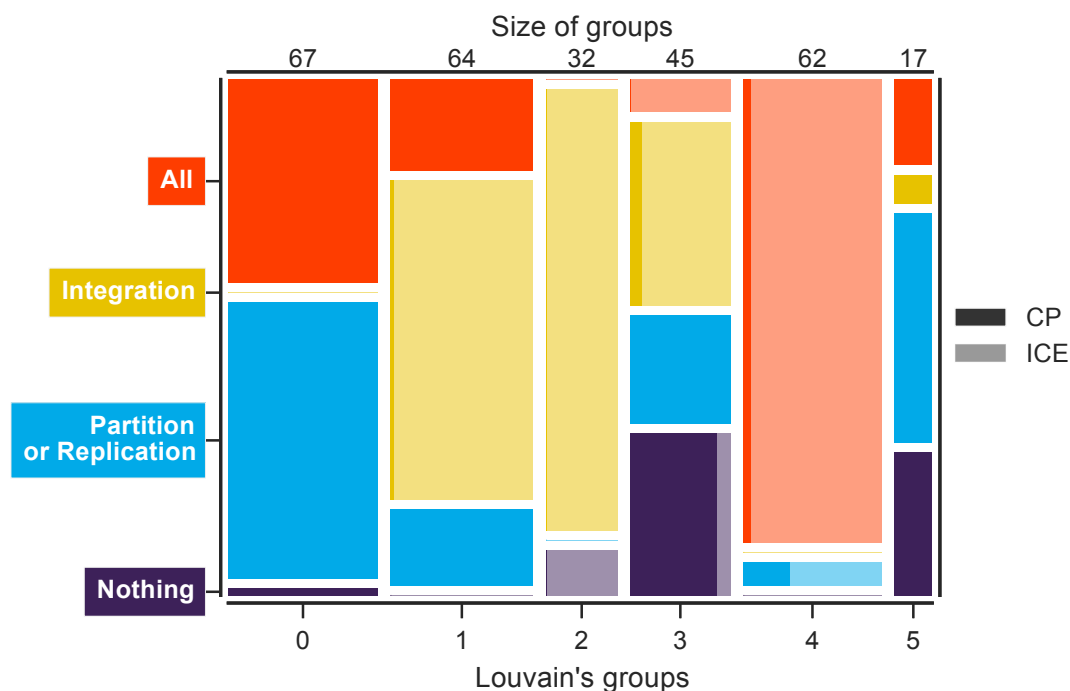
**Figure 3:** Network of weighted gene repertoire relatedness (wGRR) between ICEs and CPs. **A.** Graphical representation of the network, where nodes represent conjugative elements (triangle: CP, circle: ICE), and edges represent the wGRR score between a pair of elements. The thickness of edges is proportional to the wGRR value (edges with a wGRR < 5 % were excluded from the representation). Nodes were colored after the groups determined by the Louvain algorithm (see Methods). **B.** Heatmap of the wGRR scores, ordered after the size of Louvain's group, which are depicted on the left bar. The top bar represents the proportion of ICEs (yellow) and CP (blue) for each group. The bottom bar assigns a color corresponding to the host's species to every element. It illustrates that the host taxonomy has little impact on the Louvain clustering.

We controlled for the effect of the MPF genes in the clustering by re-doing the analysis without these genes (Figure S2). This produced the same number of groups, (N0 to N6) and the elements they contained were often the same, indeed 90% of the elements of the first groups are classed in the same novel groups (Figure S3). The only qualitatively significant difference between the two analyses concerned the group 1 for which 36% of the elements are shared among groups N3 and N5. Overall, these controls confirm that ICEs and CPs can be grouped together, and apart from other elements of the same type. The grouping is not exclusively caused by the sequence similarity between their conjugative systems. Instead, it

probably reflects either within group genetic exchanges between ICEs and CPs, or interconversions of the two types of elements.

#### Genetic exchanges between elements

The results above suggest the existence of genetic transfer between the elements. To address this question, we tested whether very similar genes were found in very different elements. Thus, for each pair of elements, we represented the wGRR as a function of the percentage of identity of the homologous proteins (Figure 5). The region of interest for this analysis thus corresponds to the lower right corner defined by a low wGRR (lower than 30%) and high protein identity (higher than 80%). It shows several genes with close to 100% identity in otherwise very distinct mobile elements, thus confirming recent transfer of a single or a few genes between different elements (Figure 5.A). We showed above that CPs have more variable sizes, suggesting that they could exchange genes frequently. Indeed, comparisons between ICEs show little evidence of such recent transfer, whereas comparisons between plasmids accounted for the vast majority of these exchanges (Figures 5.B-D).



**Figure 4:** Mosaic plot of the contingency table between Louvain's groups and key functions of conjugative elements. The width of the bar is proportional to the number of elements in a given Louvain's group (see the number of elements of each group on the top of the bars). The colors represent the type of function found on a given element. The element has neither integration nor partition or replication functions ("Nothing", grey), the element has only partition or replication functions (blue), the element has only the integration function (yellow), or, the element has both types of functions (red). Each rectangle is proportional to the number of elements in the given category. The proportion of ICE and CP is represented by the tint of the color, darker colors for CPs and lighter color for ICEs.

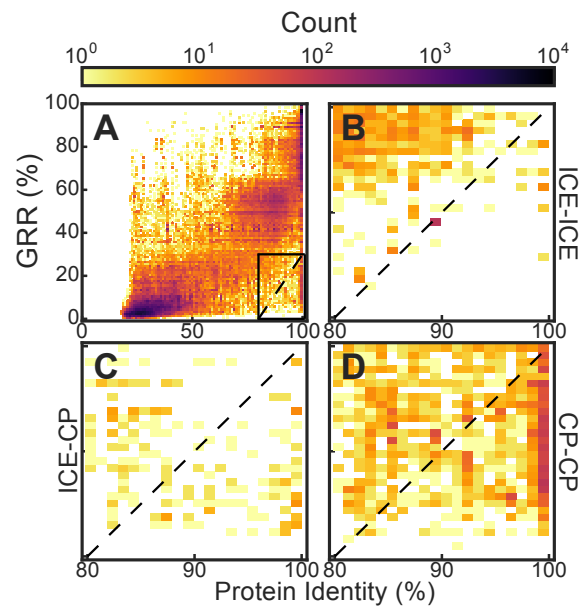
#### CPs turn to ICEs for broader host range

We hypothesized that ICEs could hold an advantage over CPs to colonize novel hosts, because replication is the key determinant of plasmid host range and this does not seem to be an essential function of most ICEs. To test this hypothesis, we analyzed the wGRR score between all pairs of ICEs and all pairs of CPs in function of the phylogenetic distance between their hosts (those with which they were sequenced). This showed similar patterns for the two types of elements, with the notable exception that there are no pairs of highly similar plasmids (wGRR>50%) in distant hosts (more than 0.1 substitutions/position, *e.g.*, the distance between *E. coli* a *P. aeruginosa.*), while a third of all ICEs (n=50) are present in the same region (Figure 6, Figure S4). The same analysis after removing the MPF genes shows wGRR values shifted to lower values for all elements, but qualitatively similar trends (Figure S5). This suggests a major difference in the ability of ICEs and CPs to be stably maintained after their transfer into a very distant host.

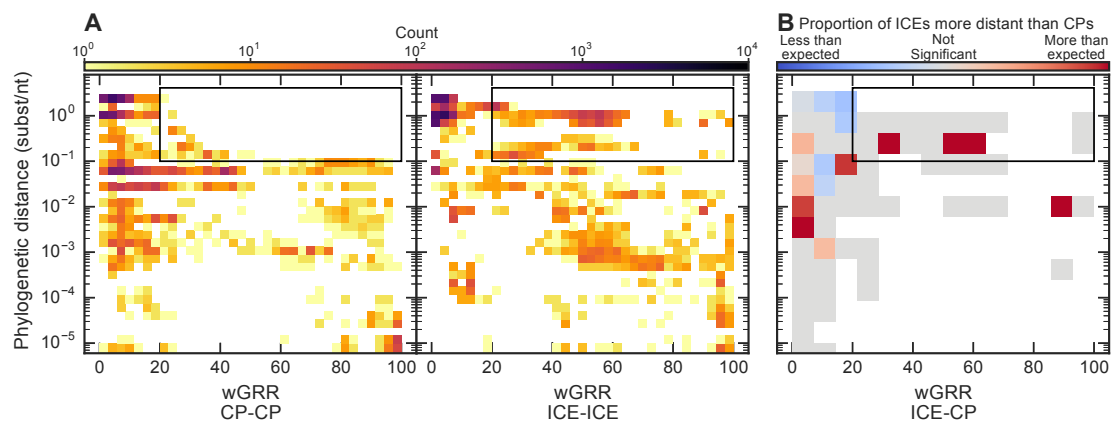
We then analyzed specifically the pairs of ICEs and CPs that were very similar and present in distant hosts (n=38, Figure 6), *i.e.*, those pairs for which one of the elements was transferred to a distant bacterial host. The determination of the element in the pair that made such a transfer was done using a method based on tri-nucleotide composition and designed to evaluate the genetic element's host range (36). In agreement to our hypothesis that ICEs can be transferred more easily across different phyla, we observed that ICEs were often significantly more distant from their host signatures than CPs (Figure 6). Interestingly, there are few pairs of highly similar ICEs and CPs in closely related hosts (bottom right corner in Figure 6, n=8 for wGRR>50% and  $d < 10^{-2}$ ). Yet, one would expect that similar pairs of ICE-CP in distant hosts were once in close hosts. This suggests that those ICEs were in fact CPs that



integrated into their host to be stably inherited. Or, stated in another way, conjugative plasmids that managed to integrate the chromosome of the novel distant host seem to have had a much higher probability of successful stabilization in its lineage.



**Figure 5:** 2D histogram of the GRR score of pairs of conjugative elements as a function of the protein identity of their homologues. **A.** Distribution for the entire dataset, expectedly, high GRR values involve high protein identity. The black rectangle zooms on a region where the pairs of elements are very different ( $GRR < 30\%$ ), yet they encode at least one very similar protein (identity  $> 80\%$ ). The dashed line separates the elements where protein identity is higher than  $wGRR \times 2/3$ . **B.** Zoom in the above-mentioned region when comparing ICEs with ICEs. **C.** Same, but comparing ICEs with CPs. **D.** Same, but comparing CPs with CPs.



**Figure 6:** 2D histogram of the distribution of the wGRR score as a function of the phylogenetic distance. The phylogenetic distance axis is in log scale. The bottom row corresponds to all pairs with distance lower than  $10^{-5}$  (including those in the same host). **A.** The two plots represent the distribution of the values for every pair of CPs and every pair of ICEs. A more intuitive representation of the information contained in the black rectangle is depicted in Figure S4. **B.** The plot represents the same distribution with larger bins to allow for a statistical analysis. The color of the bins represents the outcome of the binomial test on the p-value of the Mahalanobis distance. Bins in red indicate a higher proportion of ICEs that were measured as more distant from the host in terms of tri-nucleotide composition than the CPs of the same bin. Bins in blue indicate the opposite trend. Bins in grey show no statistically significant difference and those in white correspond to regions of the graph where there were no pairs of ICE-CP elements.

## Discussion

### The plasticity of conjugative plasmids and the transferability of ICEs

Integrative and extrachromosomal elements have been known for many decades, but the reasons for the existence of the two types of elements have not been explored. Our study highlighted advantages to each of the forms in the case of conjugative elements. Given their size distribution, conjugative plasmids seem to be more flexible relative to ICEs in terms of the amount of genetic information they can carry and the ability they have to accommodate novel information. Accordingly, we showed that transfers between distantly related plasmids are much more frequent than among ICEs. Mechanistically, the rate of recombination among plasmids may be higher because they are often in higher copy number in cells than ICEs, which are usually single-copy relative to the chromosome, and because plasmids, as shown in this work, encode more repeats, integrons, and transposable elements that will promote exchanges between replicons. ICEs may also be more constrained in terms of size because they integrate the bacterial genome. Accordingly, the size of ICEs varies less with genome size than that of conjugative plasmids. Larger genomes engage more frequently in horizontal gene transfer and recombination (41), and plasmids may thus play a key role in the evolution of the largest genomes. This fits the previous observations of frequent mega-plasmids in the large genomes of alpha- and beta-Proteobacteria (42).

In contrast, our results suggest that ICEs are advantaged concerning host range. Indeed, we could not identify pairs of very similar plasmids in very distant hosts, whereas many ICEs were in these conditions. It is interesting to notice that the family of the first known ICE, Tn916, became notorious due to its ability to spread antibiotic resistance between very distant phyla (43). In complement with this observation, we showed that in pairs of very similar CP-ICE present in distant bacteria, the ICEs are those systematically more distant from the host in terms of sequence composition than the plasmid of the pair. This observation is even more striking because we also show that on average ICEs are usually compositionally closer to the host than conjugative plasmids (presumably because they replicate with the chromosome). We thus propose that plasmids transferred to very distant

novel hosts have a much higher probability of being maintained in the novel lineage if they integrate the genome and become an ICE.

There may be other traits that provide advantages to ICEs or conjugative plasmids in certain circumstances. For example, the copy number of plasmids is usually higher than that of the chromosome, and thus of ICEs. This may be very costly to the cell (44). On the other hand, the ability to modify the copy number of plasmids may accelerate adaptive evolutionary processes, such as the acquisition of antibiotic resistance (45). Another characteristic that may differentiate conjugative plasmids and ICEs concerns their relative stability. Plasmids, by being independently replicated and segregated, might be more easily lost from bacterial lineages than ICEs. This agrees with our observations that exclusion systems and restriction-modification systems are more abundant in CPs, since both may increase the stability of the elements in the cell. Further work will be needed to test these hypotheses.

### The thin line between ICEs and CPs

Several studies in the last years have shown that the classical divide between ICEs and CPs was less pronounced than previously thought (13, 14, 22). Of course, it has been known for a long time, thanks to Hfr strains (46), that plasmids can integrate the bacterial chromosome, and demonstrated a few years ago that some ICEs replicate in bacterial cells (14, 15). However, the quantification of similarities and differences between the two types of elements had never been quantified precisely. Here, we show that numerous plasmids have integrases and that numerous ICEs encode replication and partition functions. Accordingly, the clustering of elements based on gene repertoire relatedness showed groups with both ICEs and CPs, and this was not just caused by similarities in the conjugative systems. There are thus many similarities between ICEs and CPs. Yet, we would argue that there are also some clear differences between them. First, plasmids do have much higher frequencies of genes encoding replicases and partition systems than ICEs, and the inverse applies in relation to integrases. Interestingly, we were only able to attribute incompatibility groups to CPs, suggesting that the replication module is rarely exchanged between ICEs and CPs and this may lead to (or be a consequence of) specialization of its roles in the two types of elements. It should also be noted that Tyrosine recombinases of plasmids may be involved in processes unrelated to integration, such as dimer resolution (47). Second, the

accessory traits encoded by the two elements show significant quantitative differences. Notably, plasmids are more likely to encode antibiotic resistance genes whereas ICEs encode more metabolism-related genes. Finally, as discussed above, the patterns of gene variation and exchange, and the host range are different.

The quantification of the differences between ICEs and conjugative plasmids opens the way to study their interactions and how they shape bacterial evolution. The exchange of genetic information seems to be more frequent between plasmids than between ICEs, but the occasional transfers observed between plasmids and ICE allow the latter to access the larger gene pool of the former. These events of genetic transfer may occasionally lead to elements that have all key traits of ICEs and of conjugative plasmids. This was the case of more than a third of all conjugative elements analyzed in this study. ICEs and CPs with these traits often grouped together, as observed in the group 4, and facilitate the interconversion of one type of element into another type.

#### [Implication and limitation for other MGE](#)

Our results have implications beyond the evolution of conjugative elements. Many integrative of extrachromosomal elements are not conjugative but mobilizable. These elements often encode a relaxase that recognizes the element's origin of transfer and is able to interact with a T4SS from an autonomously conjugative element to transfer to other cells. Many of the disadvantages of conjugative plasmids and ICEs are similar to those of mobilizable plasmids and integrative mobilizable elements, whether they encode a relaxase or not. Notably, the former must be replicated in the extrachromosomal state, and the latter integrate the genome where they must not disrupt genome organization. Patterns observed in conjugative elements are thus likely to be applicable to mobilizable ones.

These results may also be relevant to understand the biology of temperate phages at the onset of lysogeny. The vast majority of known prophages are integrated in the chromosome, and plasmid-replicating phages (like N15 and P1 (48, 49)) were long considered as exceptions. Yet, recent works have found that some extrachromosomal elements replicating like plasmids are actually prophages (50–52). Considering that integrative prophages may be under constraints very similar to those of ICE and that

plasmid-prophages must replicate in the extrachromosomal state, they are likely to be under similar trade-offs as, respectively, ICEs and CPs. However, phages are under additional constraints. Notably, their genome size is much less variable than that of conjugative elements, because it must be packaged into the virion (53). This lower plasticity may render the extrachromosomal prophages less advantageous in terms of accumulating novel genes, and thus explain why conjugative systems are so evenly split between integrative and extrachromosomal elements, whereas most prophages are integrative.

## Material and Methods

### Data

Conjugative systems of type T (MPF<sub>T</sub>) were searched in the set of complete bacterial genomes from NCBI RefSeq (<http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/>, last accessed in November 2016). We analyzed 5562 complete genomes from 2268 species, including 4345 plasmids and 6001 chromosomes. The classification of the replicon in plasmid or chromosome was taken from the information available in the GenBank file. Our method to delimit ICEs is based on comparative genomics of closely related strains. Hence, we restricted our search for conjugative systems to the species for which we had at least five genomes completely sequenced (164 species, 2990 genomes).

### Detection of conjugative systems and delimitation of ICEs

Conjugative systems were detected using the CONJscan module of MacSyFinder (20), with protein profiles and definitions of the MPF type T, published previously (21). ICEs were delimited with the same methodology we developed in a previous work (22). Briefly, we identified the core genomes of the species. The region between two consecutive genes of the core genome defined an interval in each chromosome. We then defined spots as the sets of intervals in the chromosome flanked by genes of the same two families of the core genome (23). We then identified the intervals and the spots with conjugative systems. The information on the sets of gene families of the spots with ICEs (i.e., the spot pan-genome) was used to delimit the elements (script available at [https://gitlab.pasteur.fr/gem/spot\\_ICE](https://gitlab.pasteur.fr/gem/spot_ICE)). This methodology has been shown to be accurate at the level of the genes (precise nucleotide boundaries are not identifiable by this method, see (22)).

### Functional analyses

Partition systems, replication systems, entry-exclusion systems and restriction modification systems were annotated with HMM profiles, as described in our previous work (22, 24). Integrases were annotated with the PFAM profile PF00589 for the Tyrosine recombinases and the combination of PFAM profiles PF00239 and PF07508 for Serine recombinases. DDE Transposases were detected with Macsyfinder (20) with models used previously (25).

Antibiotic resistance genes were detected with ResFams profiles (core version v1.1) (26) using the `--cut_ga` option. We determined the functional categories of genes using their annotation as provided by their best hits to the protein profiles of the EggNOG database for bacteria (version 4.5, bactNOG) (27). Genes not annotated by the EggNOG profiles were classed as “Unknown” and included in the “Poorly characterized” group. The HMM profiles were used to search the genomes with HMMER 3.1b2 (28), and we retrieved the hits with an e-value lower than  $10^{-3}$  and with alignments covering at least 50% of the profile. Integrons were detected using IntegronFinder version 1.5.1 with the `--local_max` option for higher accuracy (29). Repeats were detected with Repseek (version 6.6) (30) using the option `-p 0.001` which set the p-value for determining the minimum seed length.

### Statistics

We tested the over-representation of a given function or group of functions using Fisher's exact tests on contingency tables. For partition, replication and integration, the contingency table was made by splitting replicons in those encoding or not encoding the function and between ICEs and CPs. The use of presence/absence data instead of the absolute counts was made because the presence of at least one occurrence of a system is sufficient to have the function and because the counts were always low. For the other functions, the contingency table was made by splitting the proteins of the element in those annotated for a given function and the remaining ones. This allowed to take into account the differences in the number of genes between elements. The Fisher-exact tests were considered as significant after sequential Holm-Bonferroni correction, with a family-wise error rate of 5% (the probability of making at least one false rejection in the multiple tests, the type I error). From the contingency table, we computed the relative ratio (or relative risk) of having a given function more often in ICEs than in CPs. The relative ratio is computed as follow:  $RR = \frac{ICE_{WF}/N_{ICE}}{CP_{WF}/N_{CP}}$  where  $ICE_{WF}$  is the number of ICE (or proteins in ICEs) with the given function, and  $N_{ICE}$ , the total number of ICE (or proteins in ICEs), and likewise for CP. The term  $ICE_{WF}/N_{ICE}$  is an estimation of the probability of an ICE (or a protein in an ICE) to carry a given a function.



### Phylogenetic tree

We identified the genes present in at least 90% of the 2897 genomes of Proteobacteria larger than 1 Mb genomes available in GenBank RefSeq in November 2016. A list of orthologs was identified as reciprocal best hits using end-gap free global alignment. Hits with less than 37% similarity in amino acid sequence and more than 20% difference in protein length were discarded. We then identified the protein families with relations of orthology in at least 90% of the genomes. They represent 341 protein families. We made multiple alignments of each protein family with MAFFT v.7.205 (with default options) (31) and removed poorly aligned regions with BMGE (with default options) (32). Genes missing in a genome were replaced by stretches of "-" in each multiple alignment, which has been shown to have little impact in phylogeny reconstruction (33). The tree of the concatenate alignment was computed with FastTree version 2.1 under the LG model (34). We chose the LG model because it was the one that minimized the AIC.

### Distance to the host

We used the differences in tri-nucleotide composition to compute the genetic distance between the mobile element and its host chromosome, as previously proposed (35). The analysis of ICEs was done by comparing the element with the chromosome after the removal of its sequence from the latter. Briefly, we computed the trinucleotide relative abundance ( $x_{ijk} \forall i, j, k \in \{A, T, C, G\}$ ) for the chromosomes (in windows of 5 kb) and for the conjugative elements (entire replicon), which is given by:  $x_{ijk} = f_{ijk} / f_i f_j f_k$ , with  $f$  the frequency of a given k-mer in the sequence (36). We first computed the Mahalanobis distance between each window and the host chromosome as follow:

$$D = \sqrt{(w - h)^T H^{-1} (w - h)}$$

with  $w$ , the vector of tri-nucleotide abundances ( $x_{ijk}$ ) in a given window, and  $h$ , the mean of the vector of  $x_{ijk}$  (*i.e.*, the average tri-nucleotide abundance in the chromosome).  $H$  is the covariance matrix of the tri-nucleotide relative abundances. The inverse of the covariance matrix ( $H^{-1}$ ) downweights frequent trinucleotides, like the tri-nucleotides corresponding to start codons, which are common to conjugative elements and chromosome and could bias the distance. We computed the Mahalanobis distance between

conjugative elements and their hosts' chromosomes (same formula as above, but  $w$  is now for a conjugative element instead of a chromosome window). We then computed the probability (p-value) that the measured distance between a conjugative element and the host's chromosome is the same as any fragment of the host's chromosome.

We compared ICEs and CPs in relation to their compositional distance to the host. For this, we made the null hypothesis that the proportion of ICEs having a p-value lower than CPs follows a binomial distribution whose expected proportion is that of the entire dataset (the proportion of ICEs having a p-value lower than CP), precisely:  $H_0 = N(pvalue_{ICE} < pvalue_{CP})/N_{Comparisons}$ , where  $N_{Comparisons}$  is the total number of ICE-CP pairs, *i.e.*  $151 \times 136 = 20536$ .

#### Network analysis of gene repertoire relatedness

We built a network describing the relations of homology between the elements. The nodes in the network are conjugative elements and they are linked if they share a given relation of homology. More precisely, the relationship between two elements was quantified with the weighted Gene Repertoire Relatedness score (wGRR). This score represents the number of homologous proteins between two elements, weighted by their sequence identity, as described in (22). The formula is:

$$wGRR_{A,B} = \sum_i \frac{id(A_i, B_i)}{\min(A, B)} \leftrightarrow_{iff} evaluate(A_i, B_i) < 10^{-5}$$

Where  $(A_i, B_i)$  is the  $i^{th}$  pair of homologous protein between element A and element B,  $id(A_i, B_i)$  is the sequence identity of their alignment,  $\min(A, B)$  is the number of proteins of the element with fewest proteins (A or B). The sequence identity was computed with blastp v.2.2.15 (default parameters)(37) and kept all bi-directional best hits with an e-value lower than  $10^{-5}$ .

The network was built based on the wGRR matrix. Its representation was made using the Fruchterman-Reingold force-directed algorithm as implemented in the NetworkX v1.11 python library. The groups were made using the Louvain algorithm (38). We controlled for the consistency of the heuristic used to assess that the group found are not form a local

optimal. We performed 100 clustering, which led to the same classification in 95% of the time.

#### Incompatibility typing

We determined the incompatibility group of replicons using the method of PlasmidFinder (39). We used BLASTN (37) to search the replicons for sequences matching the set of 116 probes used by PlasmidFinder. We kept the hits with a coverage above 60% and sequence identity above 80%, as recommended by the authors. Several incompatibility types can be attributed to a single element using this method.

## References

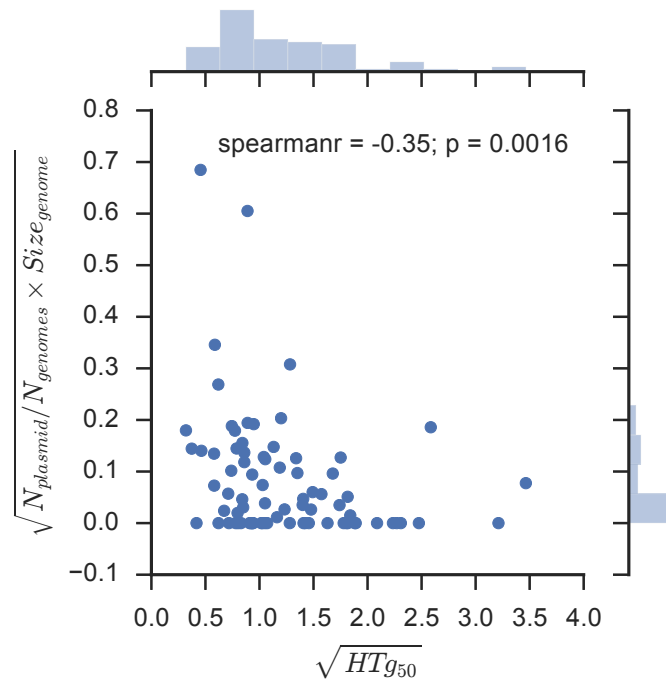
1. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–32.
2. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
3. Ebersbach,G. and Gerdes,K. (2005) Plasmid segregation mechanisms. *Annu. Rev. Genet.*, **39**, 453–79.
4. Summers,D.K. (1991) The kinetics of plasmid loss. *Trends Biotechnol.*, **9**, 273–278.
5. Kobayashi,I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.*, **29**, 3742–3756.
6. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.*, **2**, 414–24.
7. Guglielmini,J., Quintais,L., Garcillán-Barcia,M.P., de la Cruz,F. and Rocha,E.P.C. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
8. Canchaya,C., Proux,C., Fournous,G., Bruttin,A. and Brussow,H. (2003) Prophage Genomics. *Microbiol. Mol. Biol. Rev.*, **67**, 238–276.
9. Johnson,C.M. and Grossman,A.D. (2015) Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu. Rev. Genet.*, **49**, annurev-genet-112414-055018.
10. Bellanger,X., Payot,S., Leblond-bourget,N. and Guédon,G. (2014) Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.*, **38**, 720–760.
11. Carraro,N. and Burrus,V. (2014) Biology of Three ICE Families: SXT/R391, ICEBs1, and ICEst1/ICEst3. *Microbiol. Spectr.*, **2**, 1–20.
12. Delavat,F., Miyazaki,R., Carraro,N., Pradervand,N. and van der Meer,J.R. (2017) The hidden life of integrative and conjugative elements. *FEMS Microbiol. Rev.*, **41**, 512–537.
13. Carraro,N. and Burrus,V. (2015) The dualistic nature of integrative and conjugative elements. *Mob. Genet. Elements*, **5**, 98–102.
14. Lee,C. a, Babic,A. and Grossman,A.D. (2010) Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.*, **75**, 268–79.
15. Carraro,N., Poulin,D. and Burrus,V. (2015) 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. *PLOS Genet.*, **11**, e1005298.
16. Nunes-Düby,S.E., Kwon,H.J., Tirumalai,R.S., Ellenberger,T. and Landy, a (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res.*, **26**, 391–406.
  17. Guglielmini,J., de la Cruz,F. and Rocha,E.P.C. (2013) Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.*, **30**, 315–31.
  18. Guiney,D.G. (1982) Host range of conjugation and replication functions of the Escherichia coli sex plasmid Flac. Comparison with the broad host-range plasmid RK2. *J. Mol. Biol.*, **162**, 699–703.
  19. Zhong,Z., Helinski,D. and Toukdarian,A. (2005) Plasmid host-range: Restrictions to F replication in Pseudomonas. *Plasmid*, **54**, 48–56.
  20. Abby,S.S., Néron,B., Ménager,H., Touchon,M. and Rocha,E.P.C. (2014) MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One*, **9**, e110726.
  21. Guglielmini,J., Néron,B., Abby,S.S., Garcillán-Barcia,M.P., la Cruz,F. de and Rocha,E.P.C. (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.*, 10.1093/nar/gku194.
  22. Cury,J., Touchon,M. and Rocha,E.P.C. (2017) Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.*, 10.1093/nar/gkx607.
  23. Oliveira,P.H., Touchon,M., Cury,J. and Rocha,E.P.C. (2017) The chromosomal organization of horizontal gene transfer in Bacteria. *Nat. Commun.*, **8**, 1–10.
  24. Oliveira,P.H., Touchon,M. and Rocha,E.P.C. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 1–14.
  25. Touchon,M., Cury,J., Yoon,E.-J., Krizova,L., Cerqueira,G.C., Murphy,C., Feldgarden,M., Wortman,J., Clermont,D., Lambert,T., *et al.* (2014) The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. *Genome Biol. Evol.*, **6**, 2866–2882.
  26. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2014) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
  27. Huerta-Cepas,J., Szklarczyk,D., Forslund,K., Cook,H., Heller,D., Walter,M.C., Rattei,T., Mende,D.R., Sunagawa,S., Kuhn,M., *et al.* (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*, **44**, D286-93.

28. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
29. Cury,J., Jové,T., Touchon,M., Néron,B. and Rocha,E.P. (2016) Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.*, **44**, 4539–50.
30. Achaz,G., Boyer,F., Rocha,E.P.C., Viari,A. and Coissac,E. (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics*, **23**, 119–121.
31. 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.
32. Criscuolo,A. and Gribaldo,S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, **10**, 210.
33. 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.
34. Price,M.N., Dehal,P.S. and Arkin,A.P. (2009) Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.
35. Suzuki,H., Sota,M., Brown,C.J. and Top,E.M. (2008) Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res.*, **36**, e147.
36. Suzuki,H., Yano,H., Brown,C.J. and Top,E.M. (2010) Predicting plasmid promiscuity based on genomic signature. *J. Bacteriol.*, **192**, 6045–55.
37. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
38. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech*, 10.1088/1742-5468/2008/10/P10008.
39. Carattoli,A., Zankari,E., García-Fernández,A., Larsen,M.V., Lund,O., Villa,L., Aarestrup,F.M. and Hasman,H. (2014) In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.
40. Rocha,E.P.C. and Danchin,A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**, 291–4.
41. Oliveira,P.H., Touchon,M. and Rocha,E.P.C. (2016) Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 5658–63.
42. Smillie,C., Garcillán-Barcia,M.P., Francia,M.V., Rocha,E.P.C. and de la Cruz,F. (2010)

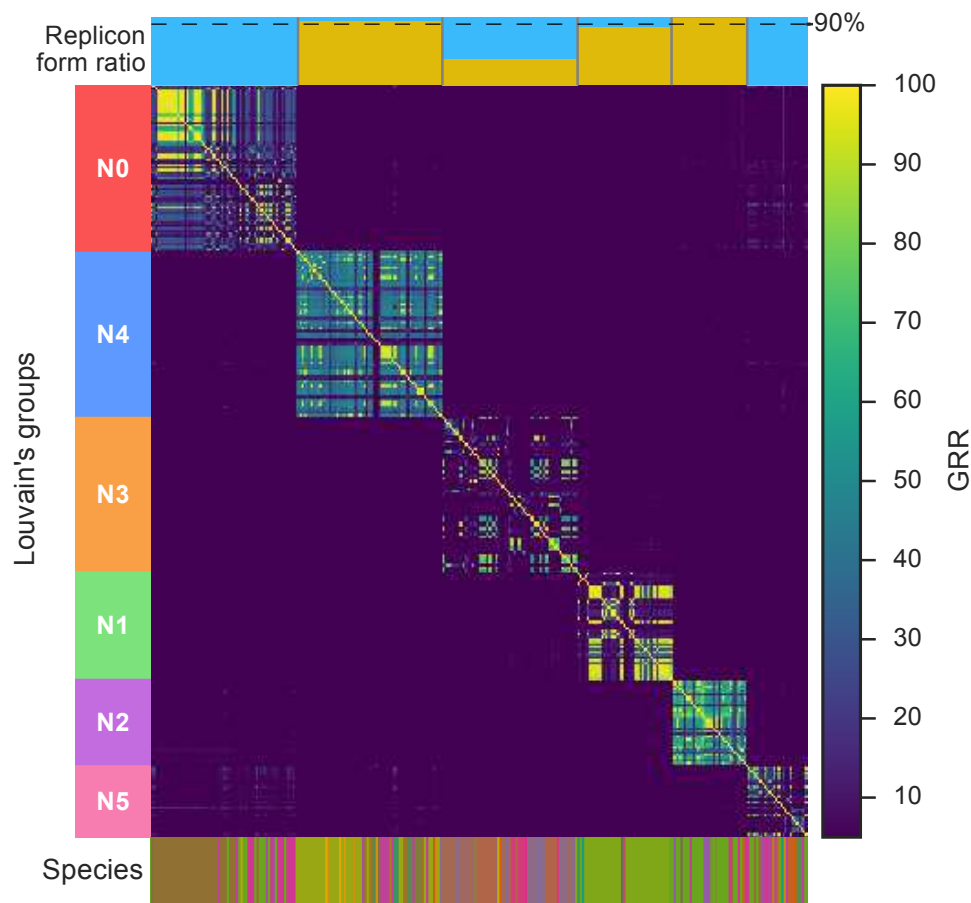
- Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–52.
43. Clewell,D.B., Flannagan,S.E. and Jaworski,D.D. (1995) Unconstrained bacterial promiscuity: the Tn916–Tn1545 family of conjugative transposons. *Trends Microbiol.*, **3**, 229–236.
  44. San Millan,A. and MacLean,R.C. (2017) Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiol. Spectr.*, **5**, 1–12.
  45. San Millan,A., Escudero,J.A., Gifford,D.R., Mazel,D. and Maclean,R.C. (2016) Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.*, **1**, 1–8.
  46. Wollman,E.-L., Jacob,F. and Hayes,W. (1956) Conjugation and Genetic Recombination in Escherichia coli K-12. *Cold Spring Harb. Symp. Quant. Biol.*, **21**, 141–162.
  47. Carnoy,C. and Roten,C.-A. (2009) The dif/Xer recombination systems in proteobacteria. *PLoS One*, **4**, e6531.
  48. Ravin,N. V. (2011) N15: The linear phage-plasmid. *Plasmid*, **65**, 102–109.
  49. Łobocka,M.B., Rose,D.J., Plunkett,G., Rusin,M., Samojedny,A., Lehnerr,H., Yarmolinsky,M.B. and Blattner,F.R. (2004) Genome of Bacteriophage P1. *J. Bacteriol.*, **186**, 7032–7068.
  50. Xue,H., Cordero,O.X., Camas,F.M., Trimble,W., Meyer,F., Guglielmini,J., Rocha,E.P.C. and Polz,M.F. (2015) Eco-Evolutionary Dynamics of Episomes among Ecologically Cohesive Bacterial Populations. *MBio*, **6**, e00552-15.
  51. Billard-Pomares,T., Fouteau,S., Jacquet,M.E., Roche,D., Barbe,V., Castellanos,M., Bouet,J.Y., Cruveiller,S., Médigue,C., Blanco,J., *et al.* (2014) Characterization of a P1-like bacteriophage carrying an SHV-2 extended-spectrum  $\beta$ -lactamase from an Escherichia coli strain. *Antimicrob. Agents Chemother.*, **58**, 6550–6557.
  52. Brolund,A., Franzén,O., Melefors,Ö., Tegmark-Wisell,K. and Sandegren,L. (2013) Plasmidome-Analysis of ESBL-Producing Escherichia coli Using Conventional Typing and High-Throughput Sequencing. *PLoS One*, **8**.
  53. Touchon,M., Moura de Sousa,J.A. and Rocha,E.P. (2017) Embracing the enemy: The diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.*, **38**, 66–73.

Supplementary Figures

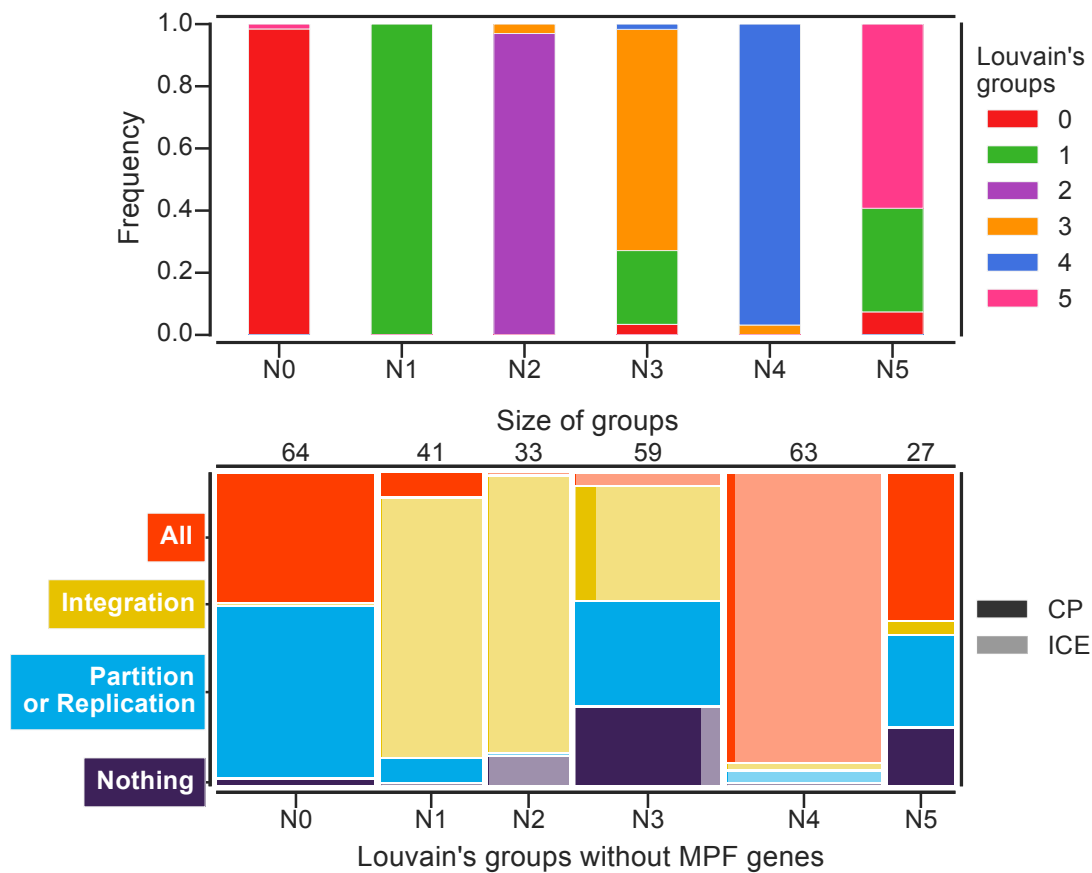


**Figure S1:** Distribution of the frequency of plasmids in genomes of a species as function of the HTg50 in the corresponding species. There is a negative correlation as measured with the Spearman coefficient  $\rho = -0.35$ ;  $p$ -value = 0.0016.

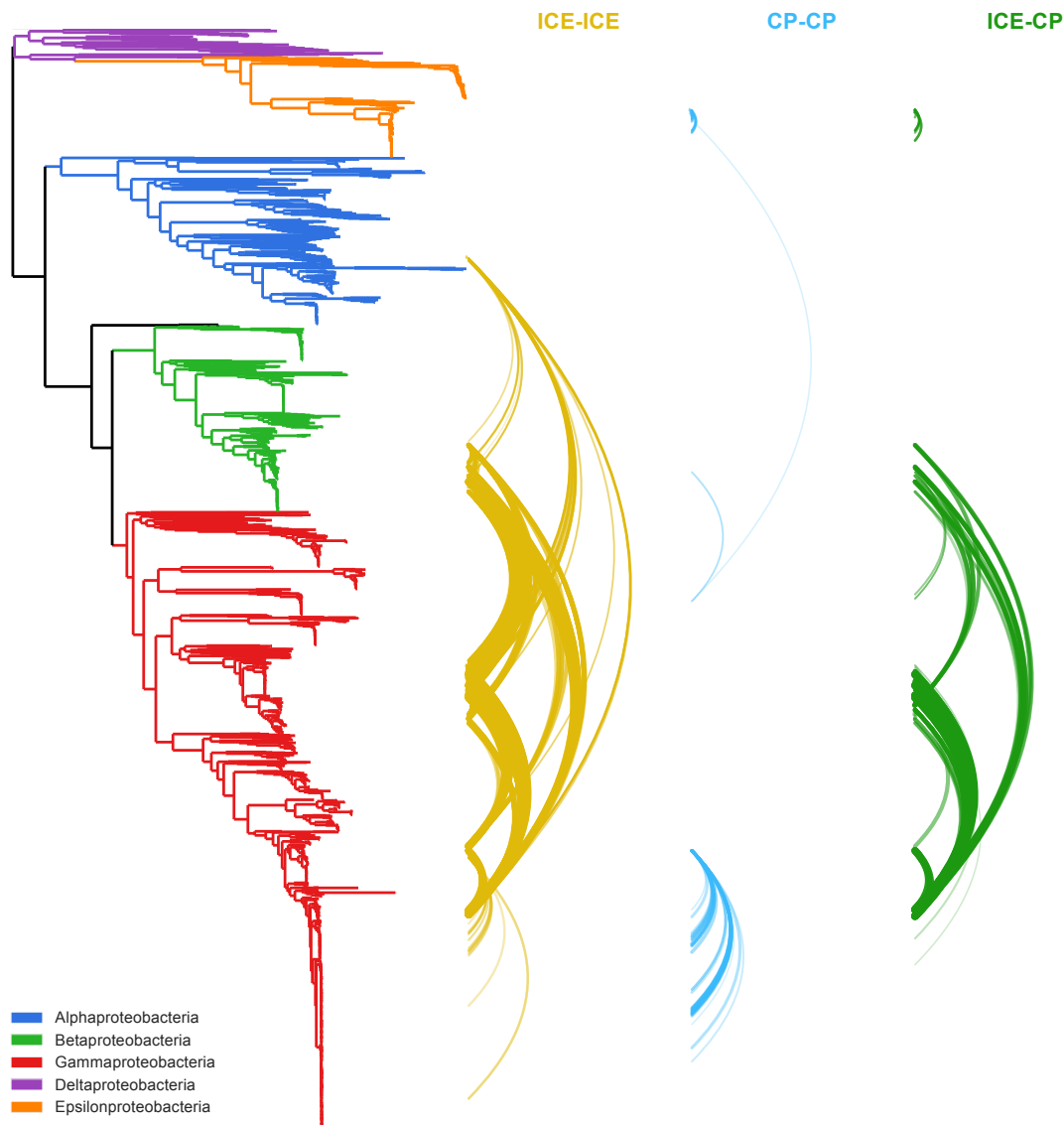




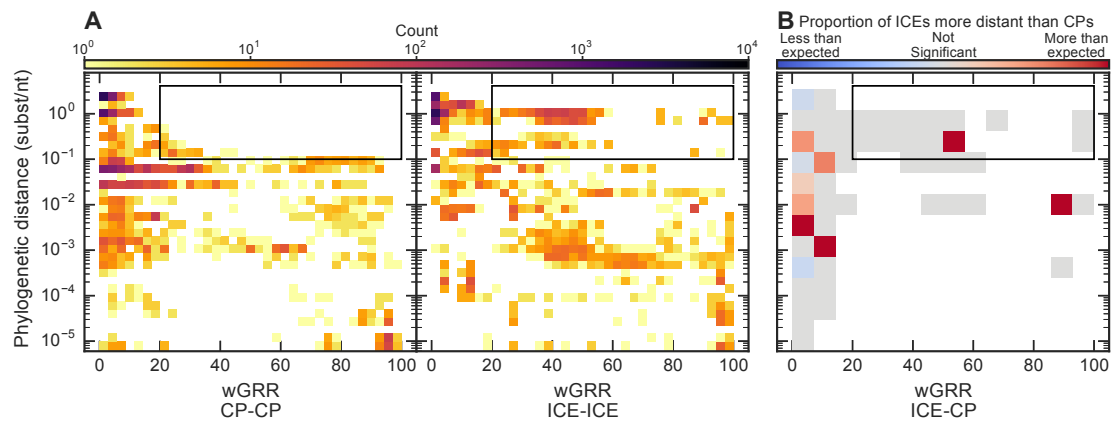
**Figure S2:** Network of weighted gene repertoire relatedness (wGRR) between ICes and CPs without MPF genes. **A.** Graphical representation of the network, where nodes represent conjugative elements (triangle: CP, circle: ICE), and edges represent the wGRR score between a pair of elements. The thickness of edges is proportional to the wGRR value. Nodes were colored after the groups determined by the Louvain algorithm (see Methods). **B.** Heatmap of the wGRR scores, ordered after the size of Louvain's group, which are depicted on the left bar. The top bar represents the proportion of ICes (yellow) and CP (blue) for each group. The bottom bar assigns a color corresponding to the host's species to every element. It illustrates that the host taxonomy has little impact on the Louvain clustering.



**Figure S3: Top.** Distribution of the Louvain’s group in the new groups formed after removing the MPF genes. Each new group correspond to the first groups. **Bottom.** Mosaic plot of the contingency table between Louvain’s groups without MPF and key functions of conjugative elements. The width of the bar is proportional to the number of elements in a given Louvain’s group (see the number of elements of each group on the top of the bars). The colors represent the type of function found on a given element. The element has neither integration nor partition or replication functions (“Nothing”, grey), the element has only partition or replication functions (blue), the element has only the integration function (yellow), or, the element has both types of functions (red). Each rectangle is proportional to the number of elements in the given category. The proportion of ICE and CP is represented by the tint of the color, darker colors for CPs and lighter color for ICEs. The group 1 is the most changed after removing the MPF genes (top), but as the proportion of functions (bottom) does not change in these latter groups, the MPF genes was thus responsible for their grouping in group 1, but the remaining 64% still contain a mixed of ICEs and CPs. Group 2 found its few CPs grouped elsewhere (N0) and leave a group composed exclusively of ICEs



**Figure S4:** Relation between pairs elements and their hosts for every pairs found in the black rectangle of Figure 6. The phylogenetic tree of Proteobacteria are depicted on the left, and major clade are colored, blue for Alphaproteobacteria, green for Betaproteobacteria, red for Gammaproteobacteria, purple for Deltaproteobacteria and orange for Epsilonproteobacteria. The single genome in black is an *Acidithiobacillia*. On the right, a line links two hosts if they contain two elements with a wGRR higher than 20% and a phylogenetic distance higher than 0.1. The thickness of the line is proportional to the wGRR. ICE-ICE relationships are in yellow, CP-CP in blue and ICE-CP in green.



**Figure S5:** 2D histogram of the distribution of the wGRR score computed without MPF genes, as a function of the phylogenetic distance. The phylogenetic distance axis is in log scale. The bottom row corresponds to all pairs with distance lower than  $10^{-5}$  (including those in the same host). **A.** The two plots on the left-hand side, represent the distribution of the values for every pair of CPs and every pair of ICEs. **B.** The plot on the right-hand side represents the same distribution with larger bins to allow for a statistical analysis. The color of the bins represents the outcome of the binomial test on the p-value of the Mahalanobis distance. Bins in red indicate a higher proportion of ICEs that were measured as more distant from the host in terms of tri-nucleotide composition than the CPs of the same bin. Bins in blue indicate the opposite trend. Bins in grey show no statistically significant difference and those in white correspond to regions of the graph where there were no pairs of ICE-CP elements.

## 1.4 Conclusion

In this chapter, we developed a method allowing the first analysis at large scale of conjugative elements. The method allowing the detection of conjugative element is based on two previous works made in the laboratory [90, 1]. Their combination led to an efficient tool for genomic analysis of conjugative elements in bacterial genomes (MacSyFinder with the CONJScan module). The development of a semi-automatic pipeline for the delimitation of ICEs, expanded the possibilities offered by this tool.

Together, the use of these tool and method resulted in the most exhaustive analysis of ICEs so far. This analysis constituted a strong base for further analysis of ICEs and conjugative elements, notably by confirming or not previous observations. For instance, we confirmed the high modularity of ICEs by showing their structured organization. Consequently, phylogenetic histories of the conjugative system and of the integrase differ, and a high proportion of unknown genes exists in cargo regions. The paraphyly of the integrase was contrasting with former results [21]. Importantly, we confirmed the presence of plasmid-like functions, which increases the importance of better understanding the relationship between ICEs and CPs.

To comprehend this relationship, we focused on MPF type T, because both forms are frequent and are well studied. Using other MPF types would have brought phylogenetic bias which should have been corrected and more importantly, a much larger dataset would have been needed. This work reveals that ICEs and CPs had qualitative similarities but quantitative differences. Indeed, both have similar average size, but the variance of CPs' size is much higher. Similarly, although replication and partition systems were found in ICEs, and integrases were found in CPs, these functions were not equally likely to be found in each element. Likewise for other functions, some were more present in ICEs and other more present in CPs. We hypothesized that ICEs should stay longer in a genome, and thus should be more frequent among the strains of a species, whereas CPs have a higher mobility and lower stability, leading to their more frequent loss. The lack of data prevents us to characterize the persistence of a given element in a species. To tackle this question, we needed sufficient species in which both CPs and ICEs were present to measure their relative frequency within each species. Our dataset did not allow us to perform statistical analysis with sufficient power. However, we are investigating the possibility to use draft genomes sequences to increase the size of our dataset to get enough power to perform this comparison. The outcome of this analysis would provide information on another key difference between ICEs and CPs. Overall, the results showed that both ICEs and CPs can be broad host range. ICEs are stably maintained in a distant host whereas CPs cannot, due to incompatibility issues, unless they integrate and become ICEs. CPs, on the other hand, were more plastic than ICE and engaged in more frequent gene exchanges.

This work relies on the hypothesis that because the conserved genes are detectable, the con-

jugative system must be functional. Definitive evidence would require experimental validation to assess the functionality of the conjugative machinery, which is impossible at large scale and will require a community level effort. However, adapted bioinformatics analyses could give clues on whether the conjugative system is likely functional or not. In case of a defective system, one could expect that part of the conjugative machinery is lacking. For instance, it has been shown that prophages endure rapid genetic degradation followed by domestication of the remaining element [17]. Similar exaptation processes could happen in conjugative systems, especially as we know that conjugative systems can be coopted into protein secretion systems or transformation machineries [69, 104]. Mobilizable elements could also result of a genetic degradation process, since inactivation of the T4SS would lead to a *de facto* mobilizable element. However, they could be just independent elements and parasites of conjugative systems, since some relaxases have been found exclusively on mobilizable elements [174]. Hence, analysis of the composition of the conjugative systems and how fast they are degraded would give insight on which systems are functional or not. This work is particularly challenging given that it has been shown that some ICEs could be separated in three different parts of the chromosome, which can assemble before conjugation [98]. A better understanding of mobilizable elements will hopefully shed new lights on the mobility of genomic islands devoid of detectable machinery allowing their transfer.

Another outcome from this work is the proposal that integrative elements can engage in longer evolutionary transfer and be maintained in the recipient, whereas extrachromosomal elements need to integrate to insure their stability in very distant lineages. Again, although this result is based on ICEs and CPs of MPF type T, the extrapolation to other elements remains to be properly demonstrated. A limit to this needed demonstration is the restricted number of systems where both forms are frequently present. For instance type G conjugative systems are almost exclusively found in ICEs. Our method cannot apply to this system and one would have to perform other analyses to assess whether this predominance of integrated form is due to similar selective forces or not. For instance, type G ICEs could combine advantages of both extrachromosomal and integrated forms. Indeed, some ICEs of type G have partition systems providing the stability of CPs if they happen to remain extrachromosomal for a certain period, like it has been shown for certain ICEs [29]. They also have a second cargo region on the opposite side of the integration module, which would allow them to engage in more gene exchanges than ICEs of type T whose cargo region is smaller<sup>1</sup>.

Finally, this work could be used to parametrize a mathematical model of horizontal gene transfer with either integrated or extrachromosomal elements. In this model, integrative elements have higher probability of successful long range transfer compared to extrachromosomal one, while the latter have higher probability of capturing and exchanging genes. This model

---

<sup>1</sup>See Figure 6, page 90

could help predict the distribution of genes across bacterial populations. One might think of a reaction-diffusion framework [195] where ICEs allow global transfer (long phylogenetic distance) at low rate, while CPs allow local transfer at high rate. Other differential equation approaches have been used in the context of mobile genetic element evolution [143]. However, this type of model does not take into account the stochasticity of evolutionary processes and the individual variation (*e.g.* the probability of conjugation depends on local environmental settings). Individual-based models, although more expensive computationally, have been proposed to model microbial interactions [101]. This type of model provides informations on both individual and population levels.





# Chapter 2

## Integrans

### 2.1 Introduction

#### 2.1.1 Background

An integron is a genetic platform allowing the capture, expression and stockpiling of genes. This element is famous for its major role in the spread of antibiotic resistance genes and for making difficult the control of bacterial infections in clinical settings [77]. What is actually famous is the clinical class 1 integron, a subtype of class 1 integron. Its success is due to its association with a Tn402-like transposon, which can integrate in other transposons or plasmids [80]. This has led to its global dissemination in many bacterial species of medical importance [52]. Thus, clinical class 1 integrons are widespread on the planet, such that some consider it as an invasive species polluting every human associated environment [80]. In addition to its high mobility, class 1 integrons have a low fitness cost for the bacteria [115]. It has been estimated that the cost of antibiotic resistance carried by class 1 integrons is up to 7 times less important than if the resistance genes were carried by a plasmid only, and up to 15 times less costly if the resistance were in the chromosome [115]. Thus, integrons can be stably maintained in bacteria while carrying costly genes, which explains the success of this class 1 integron at capturing and spreading antibiotic resistance genes. Integron captured scientists interest in the late 1990s with the discovery of an atypical integron at the time, containing hundreds of cassettes and located on the secondary chromosome of *Vibrio cholerae* [132]. This integron appeared to exist before the antibiotic era, and its cassettes were mobilizable by the already famous class 1 integron. This discovery highlighted the breadth of integron diversity.

#### 2.1.2 Diversity

Although class 1 integrons represent a major part of integron literature, they represent a small proportion of integron diversity. Integrons are found in about 10 to 20% of bacteria in

various bacterial phyla [19]. The number of cassettes can vary greatly from zero [15], to up to 200 cassettes [132]. Integrans are usually classed in two types according to their mobility, the mobile and the sedentary-chromosomal integrans.

Mobile integrans were defined by their association with a mobile genetic element, originally with a transposon or plasmid [26], later with ICE [103]. They often carry a few cassettes (up to eight). Five classes of mobile integrans have been described, and within a class, little sequence diversity exists among the integrases, suggesting their recent emergence from a broader pool of integrans. One should note, however, that the classification of mobile integrans is based on the clustering of the integrase but the threshold of sequence identity to discriminate among classes has never been formally defined [78].

The other type of integran derived from the atypical integran found in the *Vibrio cholerae* chromosome and its name was subject to controversy [94]. It was first called super-integrin due to its large number of cassettes [132]. But integrans with few cassettes were also named super-integrin due to their chromosomal location [55], which provoked controversy [94], leading to the use of chromosomal integrin. However chromosomal integrans can also be mobile. Finally, this other type of integrin is also named sedentary chromosomal integrin, and is more and more used [26, 66]. Unfortunately, the term “chromosomal integrin” is still frequently used. The sedentary chromosomal integrans are thought to be at the origin of mobile integrans that made their way into clinical settings [164]. The functions carried by mobile integrans are almost exclusively associated with antibiotic resistance or other antiseptic resistance [66]. Conversely, the functions carried by sedentary integrans rarely involve antibiotic resistance genes and most of the time encode genes of unknown function. When known, the functions encoded by sedentary integrans are very diverse, they are associated with metabolism, various cellular processes or information storage mechanisms [158].

### 2.1.3 Mechanism

Integrans are formed of two main components: a stable and a mobile one. The former is composed of an integrase called integrin-integrase, or IntI, a promoter for the integrase *P<sub>int</sub>*, a recombination site *attI* (*attachment site for the Integron*) and promoter *P<sub>c</sub>* for the array of cassettes Figure 2.1. This array of cassettes is the mobile part of the integrin, and is usually encoded on the opposite strand of the stable part [19]. A cassette is defined as a DNA sequence (generally a gene) flanked by two recombination sites, called *attC* sites, for *attachement site of the Cassette*. Most of the cassettes are promoter-less, although, some encode a promoter [137]. *attC* sites have the interesting and rather unique property for a recombination site, to be recognized in a single stranded form with a particular secondary structure [20] (Figure 2.1 — B). Consequently, the integrase IntI also has this interesting and rather unique property for a tyrosine recombinase, of being able to recombine folded single stranded DNA [127]. Because

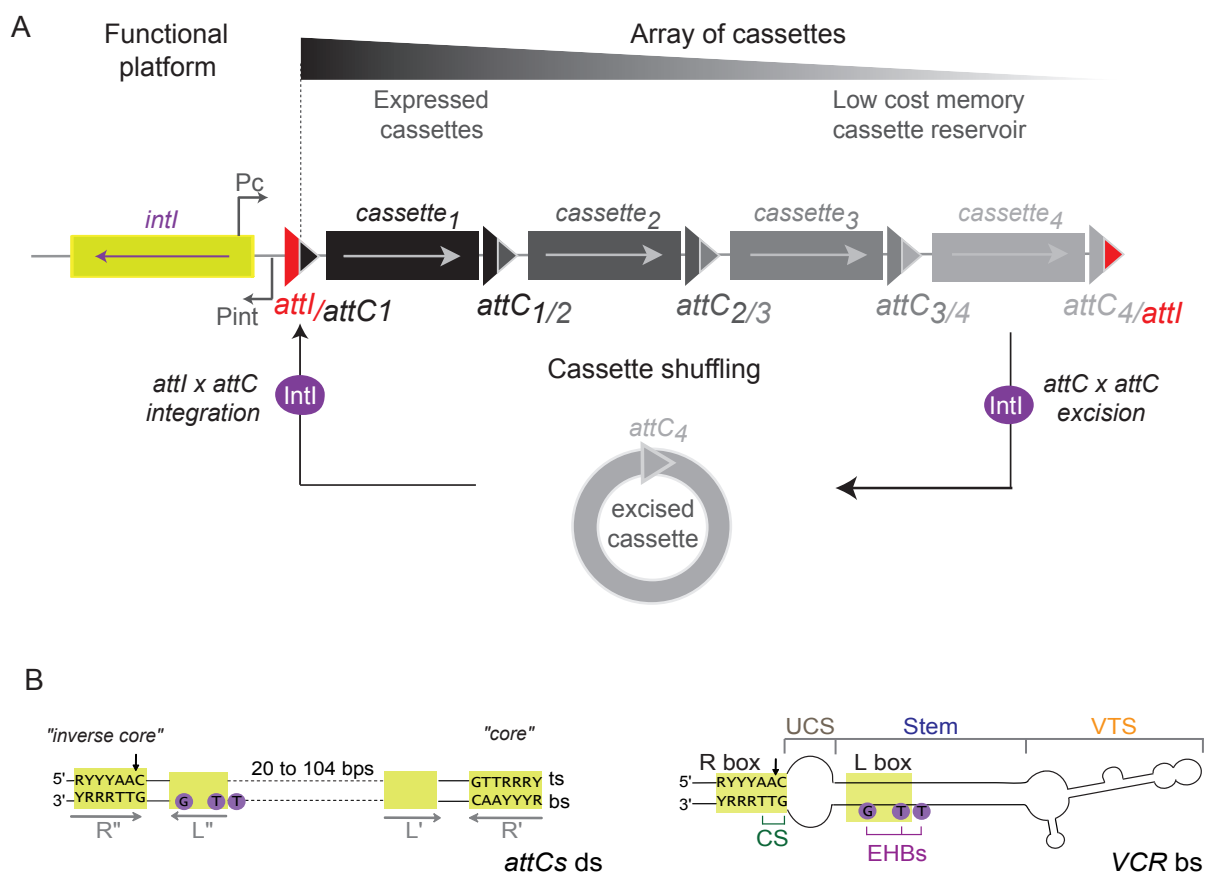


Figure 2.1: Schema of an integron and an *attC* site sequence and secondary structure.

**A.** Organization of the integron, composed of a stable platform and a cluster of cassettes. Each cassette is flanked by two *attC* sites. The recombination of between two *attC* sites leads to the excision of the cassette, while integration is mediated by the recombination between an *attC* site and an *attI* site. Note that the *attC* sites in the array are mixed with the surrounding cassettes. Cassettes far from the *attI* site and hence from the Pc promoter will be less expressed, thus, imposing a low fitness cost to the bacteria. **B.** Left, the conserved features of the *attC* site sequence. Right, the secondary structure of the *attC* site in a single stranded form. Figure adapted from [124].

the *attC* site forms a secondary structure, its primary sequence is poorly conserved and varies in size. Figure 2.1 — **B** depicts the few conserved bases and the size variation of the *attC* site. The recombination mediated by IntI between two *attC* sites will lead to the excision of the cassette. The excised cassette is thought to be single stranded and contains a single *attC* site per cassette excised (two or more contiguous cassette could potentially be excised) [66]. During excision, each flanked *attC* site is split and combined with the other *attC* site, such that it will lead to the creation of two mosaic *attC* sites, one on the excised cassette, and one in place

of the cassette in the integron (see Figure 2.1). Although *attC* site is equally used to refer to sites in the integron or sites in an excised cassette, the “true” site is the one on the excised cassette as it will always be the same, while the integrated *attC* site changes in function of the neighboring cassettes. The recombination between an *attC* site and an *attI* site will lead to the integration of the cassette at the *attI* site. The end of the *attI* site is also split upon integration of the first cassette (Figure 2.1 - **A**). Near the *attI* site, the Pc promoter allows the expression of the cassettes. The current model of integron dynamics suggests that the farther the cassettes are from the *attI* site, the less expressed they are. Therefore, the fitness cost of promoter-less cassettes is diminished, making integrans a low-cost memory tool [66].

### 2.1.4 Objectives

Most of the integrans studied are either from the five classes of mobile integrans, and especially from the class 1 integron, or from sedentary chromosomal integrans in the *Vibrio* genus. The former emerged about a century ago [80], while about a hundred of millions of years ago for the latter [164]. The knowledge-gap between these two opposite types of integron remains large. Metagenomic data revealed the abundance of cassettes, likely associated with other types of integrans [105], and that other types integrans have been described [81], but the global diversity of integrans is probably lacking. The discovery of new integrans in bacterial genomes had been hampered by the lack of tool allowing the detection of integrans. Tools have been developed, but the degenerated nature of the *attC* site restricted their limit of action to integrans related to mobile integrans [111, 194] or to *Vibrio* sedentary chromosomal integrans [163]. The crux resides in the elegant simplicity of the integron (very few components), and in its lack of conserved traits. Indeed, apart from the integrase, the cassettes are usually different, and the only conserved feature of the cassette, the *attC* site, has a very poorly conserved sequence. However, its secondary structure (Figure 2.1 - **B**) is conserved, and it has been shown that even artificial *attC* sites are recombinogenic, if they maintain this secondary structure [14]. This particular feature can be detected using covariance models, which were built to specifically search for DNA palindrome-like sequences, the imprint of secondary structures [140].

This work aims at providing an efficient tool for integron discovery, to improve our understanding of integron evolution, notably by characterizing the hidden diversity of integrans that are not associated with antibiotic resistance. It is of importance to have a full understanding of the integron system to better tackle the problem caused by the class 1 integrans, and possibly anticipate the emergence of a new “class 1 integron”.

In the following section, I present the program, IntegronFinder, that permitted the most exhaustive analysis of integrans in bacterial genomes to date, and revealed interesting characteristics of the integrans. A second study in collaboration with experimentalists, illustrate the use of this tool to better understand the dynamic and evolution of the integron system.

## 2.2 Methods and Results

### 2.2.1 Article 4: Identification and analysis of integrons and cassettes arrays in bacterial genomes

This work presents IntegronFinder, the first program to accurately detect integron in bacterial genomes, and the analysis that followed its first utilization. It sheds new lights on integron evolution and distribution among bacterial genomes and reveals unexpected patterns.

# Identification and analysis of integrons and cassette arrays in bacterial genomes

Jean Cury<sup>1,2,\*</sup>, Thomas Jové<sup>3</sup>, Marie Touchon<sup>1,2</sup>, Bertrand Néron<sup>4</sup> and Eduardo PC Rocha<sup>1,2</sup>

<sup>1</sup>Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France, <sup>2</sup>CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France, <sup>3</sup>Univ. Limoges, INSERM, CHU Limoges, UMR\_S 1092, F-87000 Limoges, France and <sup>4</sup>Centre d'Informatique pour la Biologie, C3BI, Institut Pasteur, Paris, France

Received November 09, 2015; Revised March 08, 2016; Accepted April 13, 2016

## ABSTRACT

**Integrations recombine gene arrays and favor the spread of antibiotic resistance. Their broader roles in bacterial adaptation remain mysterious, partly due to lack of computational tools. We made a program – IntegronFinder – to identify integrations with high accuracy and sensitivity. IntegronFinder is available as a standalone program and as a web application. It searches for *attC* sites using covariance models, for integrion-integrases using HMM profiles, and for other features (promoters, *attI* site) using pattern matching. We searched for integrions, integrion-integrases lacking *attC* sites, and clusters of *attC* sites lacking a neighboring integrion-integrase in bacterial genomes. All these elements are especially frequent in genomes of intermediate size. They are missing in some key phyla, such as  $\alpha$ -Proteobacteria, which might reflect selection against cell lineages that acquire integrions. The similarity between *attC* sites is proportional to the number of cassettes in the integrion, and is particularly low in clusters of *attC* sites lacking integrion-integrases. The latter are unexpectedly abundant in genomes lacking integrion-integrases or their remains, and have a large novel pool of cassettes lacking homologs in the databases. They might represent an evolutionary step between the acquisition of genes within integrions and their stabilization in the new genome.**

## INTRODUCTION

Integrions are gene-capturing platforms playing a major role in the spread of antibiotic resistance genes (reviewed in (1–4)). They have two main components (Figure 1). The first is made of the integrion-integrase gene (*intI*) and its promoter ( $P_{intI}$ ), an integration site named *attI* (attachment site of the integrion), and a constitutive promoter ( $P_C$ ) for the gene cassettes integrated at the *attI* site (5). The second com-

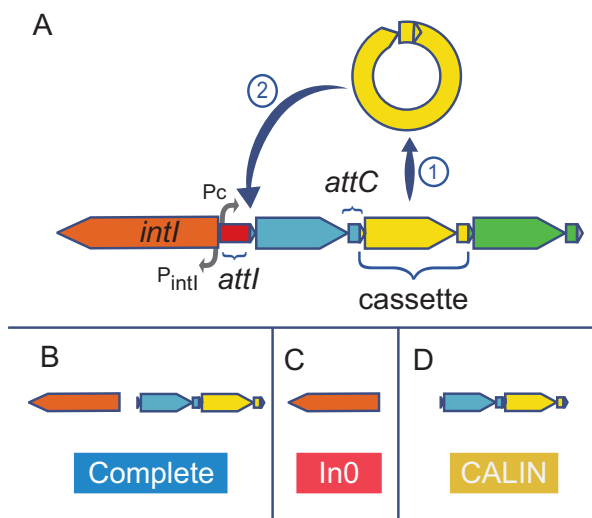
ponent is a cluster with up to 200 gene cassettes (6), most frequently transcribed in the opposite direction relative to the integrion-integrase (7). Typical gene cassettes have an open reading frame (ORF) surrounded by *attC* recombination sites (attachment site of the cassette), but the presence of the ORF is not mandatory. Cassettes carrying their own promoters are expressed independently of  $P_C$  (8).

Integrase-mediated recombination between two adjacent *attC* sites leads to the excision of a circular DNA fragment composed of an ORF and an *attC* site. The recombination of the *attC* site of this circular DNA fragment with an *attI* site leads to integration of the fragment at the location of the latter (9,10). Integrions can use this mechanism to capture cassettes from other integrions or to rearrange the order of their cassettes (11). The mechanism responsible for the creation of new cassettes is unknown.

The most distinctive features of the integrion are thus the integrion-integrase gene (*intI*) and a cluster of *attC* sites (Figure 1). The integrion-integrase is a site-specific tyrosine recombinase closely related to Xer proteins (12). Contrary to most other tyrosine recombinases, IntI recombinates nucleotide sequences of low similarity (13,14), by recognizing specific structural features of the *attC* site (15,16) (Figure 2A). This is partly caused by the presence of a ~35 residues domain near the patch III region of the integrion-integrase that is lacking in the other tyrosine recombinases (17). The integration of the *attC* site at *attI* produces chimeric *attI/attC* sites on one side and chimeric *attC/attC* sites on the other side of the cassette. This results in a cluster of chimeric *attC* sites with similar palindromic structures.

Previous literature focused on integrions carrying antibiotic resistance genes. These integrions are often mobile, due to their association with transposons, and carry few cassettes (18,19,20). Most of the so-called mobile integrions can be classed in five classes, numbered 1 to 5. The IntI within each class show little genetic diversity, indicating their recent emergence from a much larger and diverse pool of integrions (7,20–22), possibly chromosomally encoded (23). For example, prototypical class 1 integrions were found on many chromosomes of non-pathogenic soil and freshwater  $\beta$ -Proteobacteria (24). By contrast, so-called chromosomal

\*To whom correspondence should be addressed. Tel: +33 1 40 61 36 37; Email: jean.cury@normalesup.org



**Figure 1.** Schema of an integron and the three types of elements detected by IntegronFinder. (A) The integron is composed of a specific integrase gene (*intI*, orange), an *attI* recombination site (red), and an array of gene cassettes (blue, yellow and green). A cassette is typically composed of an ORF flanked by two *attC* recombination sites. The integrase has its own promoter ( $P_{intI}$ ). There is one constitutive promoter ( $P_C$ ) for the cluster of cassettes. Cassettes rarely contain promoters. The integrase can excise a cassette ① and/or integrate it at the *attI* site ②. (B) Complete integrons include an integrase and at least one *attC* site. (C) The In0 elements are composed of an integrase and no *attC* sites. (D) The clusters of *attC* sites lacking integrase-integrases (CALIN) are composed of at least two *attC* sites.

integrons are found in most strains of a species and carry cassettes encoding a wide range of functions. For example, the *Vibrio* spp. chromosomal integrons (initially called super-integrations due to the large number of cassettes they carry (25)) encode virulence factors, secreted proteins, and toxin-antitoxin modules (20). The high similarity of *attC* sites within chromosomal, but not mobile, integrons of *Vibrio* spp has prompted the hypothesis that cassettes are created by chromosomal and spread by mobile integrons (26). However, the dichotomy between chromosomal and mobile integrons has been criticized (7) because integrons encoded in the chromosome may be in mobile elements (27,28) and/or have small arrays of cassettes (29).

The analysis of metagenomics data has unraveled a vast pool of novel cassettes in microbial communities (30). Although antibiotic-resistance integrons were found to be abundant in human-associated environments such as sewage (31–33), most cassettes in environmental datasets encode different functions (or genes of unknown function) (31–33). The study of these functions, and of the adaptive impact of integrons, has been hindered by the difficulty in identifying integron cassettes. The bottleneck in these analyses is the recognition of *attC* sites, for which few tools were made available. The program XXR identifies *attC* sites in the large *Vibrio* integrons using pattern-matching techniques (20). The programs ACID (6) (no longer available) and ATTACCA (34) (now a part of RAC, available under private login) were designed to search for class 1 to class 3 mobile integrons. Such classical motif detection tools based

on sequence conservation identify *attC* sites only within restricted classes of integrons. They are inadequate to identify or align distantly related *attC* sites because their sequences are too dissimilar. Yet, these sites have conserved structural constraints that can be used to identify highly divergent sequences.

We built a program named IntegronFinder (Figure 3, [https://github.com/gem-pasteur/Integron\\_Finder](https://github.com/gem-pasteur/Integron_Finder)) to detect integrons and their most distinctive components: the integrase with the use of HMM profiles and the *attC* sites with the use of a covariance model (Figure 2). Covariance models use stochastic context-free grammars to model the constraints imposed by sequence pairing to form secondary structures. Such models have been previously used to detect structured motifs, such as tRNAs (35). They provide a good balance between sensitivity, the ability to identify true elements even if very diverse in sequence, and specificity, the ability to exclude false elements (36). They are ideally suited to model elements with high conservation of structure and poor conservation of sequence, such as *attC* sites. IntegronFinder also annotates known *attI* sites,  $P_{intI}$  and  $P_C$ , and any pre-defined type of protein coding genes in the cassettes (e.g., antibiotic resistance genes). IntegronFinder was built to accurately identify integron-integrases and *attC* sites of any generic integron. Importantly, we provide the program on a webserver that is free, requires no login, and has a long track record of stability (37) ([http://mobyte.pasteur.fr/cgi-bin/portal.py#forms::integron\\_finder](http://mobyte.pasteur.fr/cgi-bin/portal.py#forms::integron_finder)). We also provide a standalone application for large-scale genomics and metagenomics projects. We used IntegronFinder to identify integrons in bacterial genomes and to characterize their distribution and diversity.

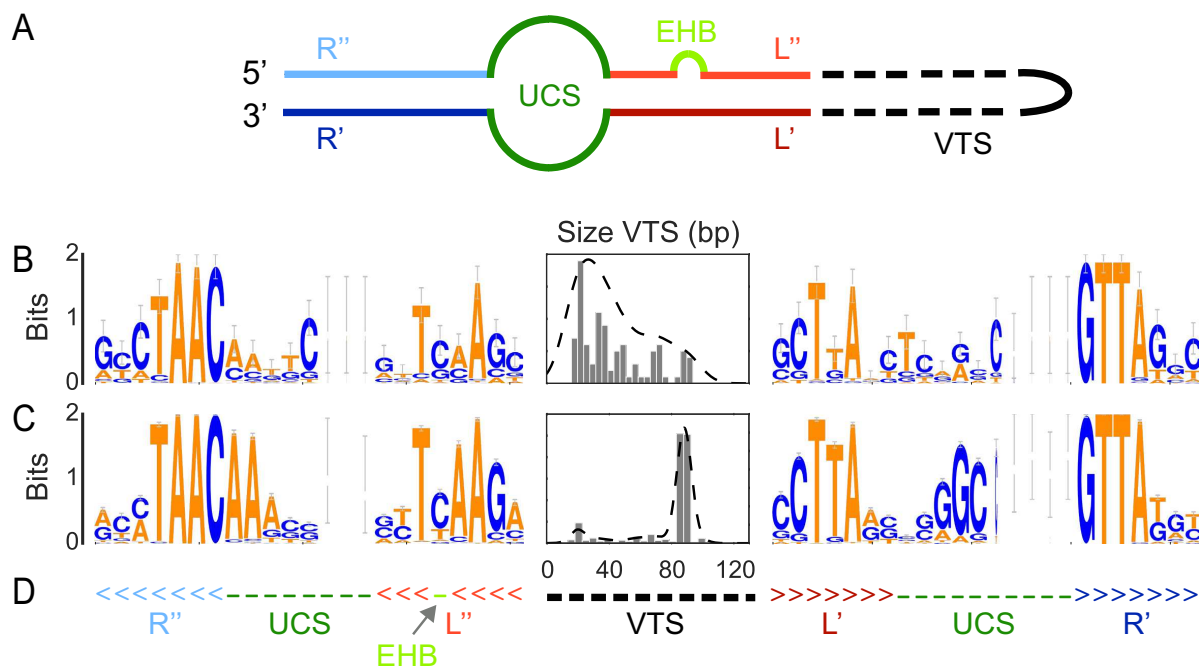
## MATERIALS AND METHODS

### Data

The sequences and annotations of complete genomes were downloaded from NCBI RefSeq (last accessed in November 2013, <http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/>). Our analysis included 2484 bacterial genomes (see Supplementary Table S1). We used the classification of replicons in plasmids and chromosomes as provided in the GenBank files. Our dataset included 2626 replicons labeled as chromosomes and 2006 as plasmids. The *attC* sites used to build the covariance model and the accession numbers of the replicons manually curated for the presence or absence of *attC* sites were retrieved from INTEGRALL, the reference database of integron sequences (<http://integrall.bio.ua.pt/>) (38). We used a set of 291 *attC* sites (Supplementary File 1) to build and test the model, and a set of 346 sequences with expert annotation of 596 *attC* sites to analyze the quality of the program predictions (Supplementary Tables S2a and S2b).

### Protein profiles

We built a protein profile for the region specific to the integron tyrosine recombinase. For this, we retrieved the 402 IntI homologs from the Supplementary file 11 of Cambray et al. (39). These proteins were clustered using uclust 3.0.617



**Figure 2.** Characteristics of the *attC* sites. (A) Scheme of the secondary structure of a folded *attC* site. EHB stands for Extra Helical Bases. (B) Analysis of the *attC* sites used to build the model, including the WebLogo (73) of the R and L box and unpaired central spacers (UCS) and the histogram (and kernel density estimation) of the size of the variable terminal structure (VTS). The Weblogo represents the information contained in a column of a multiple sequence alignment (using the log<sub>2</sub> transformation). The taller the letter is, the more conserved is the character at that position. The width of each column of symbols takes into account the presence of gaps. Thin columns are mostly composed of gaps. (C) Same as (B) but with the set of *attC* sites identified in complete integrans found in complete bacterial genomes. (D) Secondary structure used in the model in WUSS format, colors match those of (A).

(40) with a threshold of 90% identity to remove very closely related proteins (the largest homologs were kept in each case). The resulting 79 proteins were used to make a multiple alignment using MAFFT (41) (`-globalpair -maxiterate 1000`). The position of the specific region of the integron-integrase in *V. cholerae* was mapped on the multiple alignments using the coordinates of the specific region taken from (17). We recovered this section of the multiple alignment to produce a protein profile with hmmbuild from the HMMer suite version 3.1b1 (42). This profile was named intI\_Cterm (Supplementary File 2).

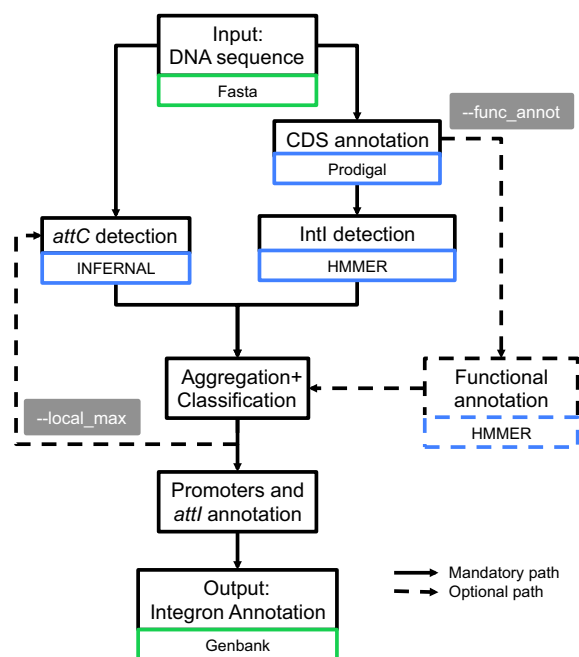
We used 119 protein profiles of the Resfams database (core version, last accessed on January 20, 2015 v1.1), to search for genes conferring resistance to antibiotics (<http://www.dantaslab.org/resfams>, (43)). We retrieved from PFAM the generic protein profile for the tyrosine recombinases (PF00589, phage\_integrase, <http://pfam.xfam.org/>, (44)). All the protein profiles were searched using hmmssearch from the HMMer suite version 3.1b1. Hits were regarded as significant when their e-value was smaller than 0.001 and their alignment covered at least 50% of the profile.

#### Construction and analysis of *attC* models

We built a covariance model for the *attC* sites (Supplementary File 3). These models score a combination of sequence and secondary structure consensus (35) (with the limitation that these are DNA not RNA structures). To produce

the *attC* models, 96 *attC* sites (33%) were chosen randomly from 291 known *attC* (see Data). The alignments were manually curated to keep the known conserved regions of the R and L boxes aligned in blocks. The unpaired central spacers (UCS) and the variable terminal structure (VTS) were not aligned because they were poorly conserved in sequence and length. Gaps were inserted in the middle of the VTS sequence as needed to keep the blocks of R and L boxes aligned. The consensus secondary structure was written in WUSS format beneath the aligned sequences (Supplementary File 4). The model was then built with INFERNAL 1.1 (36) using *cmbuild* with the option '`-hand`'. This option allows the user to set the columns of the alignment that are actual matches (consensus). This is crucial for the quality of the model, because most of the columns in the R and L boxes would otherwise be automatically assigned as inserts due to the lack of sequence conservation. The R-UCS-L sections of the alignment were chosen as the consensus region, and the VTS was designed as a gap region. We used *cmcalibrate* from INFERNAL 1.1 to fit the exponential tail of the covariance model e-values, with default options. The model was used to identify *attC* sites using INFERNAL with two alternative modes. The default mode uses heuristics to reduce the sequence space of the search. The Inside algorithm is more accurate, but computationally much more expensive (typically 10<sup>4</sup> times slower) (36). By default, the *attC* sites were kept for further analysis when





**Figure 3.** Diagram describing the different steps used by IntegronFinder to identify and annotate integrons. Solid lines represent the default mode, dotted lines optional modes. Blue boxes indicate the main dependency used for a given step. Green boxes indicate the format of the file needed for a given step.

their e-value were below 1. The user can set this value (option `--evalue_attc`).

### Identification of promoters and *attI* sites

The sequences of the  $P_c$  promoters, of the  $P_{intI}$  promoters and of the *attI* site were retrieved from INTEGRALL for the integrons of class 1, 2 and 3 when available (see Supplementary Table S3). We searched for exact matches of these sequences (accepting no indel nor mismatch) using pattern matching as implemented in the search function of the Bio.motifs module of Biopython v1.65 (45).

### Overview of IntegronFinder: a program for the identification of *attC* sites, *intI* genes, integrons and CALIN elements

The input of IntegronFinder is a sequence of DNA in FASTA format. The sequence is annotated with Prodigal v2.6.2 (46) using the default mode for replicons larger than 200 kb and the metagenomic mode for smaller replicons (`--p meta` in Prodigal) (Figure 3). In the present work, we omitted the annotation part and used the NCBI RefSeq annotations because they are curated. The annotation step is particularly useful to study newly acquired sequences or poorly annotated ones.

The program searches for the two protein profiles of the integron-integrase using `hmmsearch` with default parameters from HMMER suite version 3.1b1 and for the *attC* sites with the default mode of `cmsearch` from INFERNAL 1.1 (Figure 3). Two *attC* sites are put in the same cluster if

they are less than 4 kb apart on the same strand. The clusters are built by transitivity: an *attC* site less than 4 kb from any *attC* site of a cluster is integrated in that cluster. Clusters are merged when localized less than 4 kb apart. The threshold of 4kb was determined empirically as a compromise between sensitivity (large values decrease the probability of missing cassettes) and specificity (small values are less likely to put together two independent integrons). More precisely, the threshold is twice the size of the largest known cassettes (~2 kb (6)). This guarantees that even in the worst case (largest known cassettes) two *attC* sites will be clustered if an intervening site was not detected. Importantly, the user can set this threshold (`-- distance_thresh` in IntegronFinder).

The results of the searches for the elements of the integron are put together to class the loci in three categories (Figure 1 - B, C, D). (i) The elements with *intI* and at least one *attC* site were named complete integrons. The word complete is meant to characterize the presence of both elements; we cannot ascertain the functionality or expression of the integron. (ii) The *In0* elements have *intI* but no recognizable *attC* sites. We do not strictly follow the original definition of *In0*, which also includes the presence of an *attI* (47), because this sequence is not known for most integrons (and thus cannot be searched for). (iii) The cluster of *attC* site lacking integron-integrase (CALIN) has at least two *attC* sites and lacks nearby *intI*.

To obtain a better compromise between accuracy and running time, IntegronFinder can re-run INFERNAL to search for *attC* sites with more precision using the Inside algorithm (`-- max` option in INFERNAL), but only around previously identified elements (`-- local_max` option in IntegronFinder). More precisely, if a locus contains an integron-integrase and *attC* sites (complete integron), the search is constrained to the strand encoding *attC* sites between the end of the integron-integrase and 4 kb after its most distant *attC*. If other *attC* sites are found after this one, the search is extended by 4 kb in that direction until no more new sites are found. If the element contains only *attC* sites (CALIN), the search is performed on the same strand on both directions. If the integron is *In0*, the search for *attC* sites is done on both strands in the 4 kb flanking the integron-integrase on each side. The program then searches for promoters and *attI* sites near the integron-integrase. Finally, it can annotate the integron genes' cassettes (defined in the program as the CDS found between *intI* and 200 bp after the last *attC* site, or 200 bp before the first and 200 bp after the last *attC* site if there is no integron-integrase) using a database of protein profiles (option `-- func.annot`). For example, in the present study we used the ResFams database to search for antibiotic resistance genes. One can use any `hmm`-compatible profile databases with the program.

The program outputs tabular and GenBank files listing all the identified genetic elements associated with an integron. The program also produces a figure in pdf format representing each complete integron. For an interactive view of all the hits, one can use the GenBank file as input in specific programs such as Geneious (48).

The user can change the profiles of the integrases and the covariance model of the *attC* site. Thus, if novel models of

*attC* sites were to be built in the future, e.g., for novel types of *attC* sites, they could easily be plugged in IntegratorFinder.

### Phylogenetic analyses

We have made two phylogenetic analyses. One analysis encompasses the set of all tyrosine recombinases and the other focuses on IntI. The phylogenetic tree of tyrosine recombinases (Supplementary Figure S1) was built using 204 proteins, including: 21 integrases adjacent to *attC* sites and matching the PF00589 profile but lacking the intI\_Cterm domain, seven proteins identified by both profiles and representative of the diversity of IntI, and 176 known tyrosine recombinases from phages and from the literature (12). We aligned the protein sequences with Muscle v3.8.31 with default options (49). We curated the alignment with BMGE using default options (50). The tree was then built with IQ-TREE multicore version 1.2.3 with the model LG+I+G4. This model was the one minimizing the Bayesian Information Criterion (BIC) among all models available ('-m TEST' option in IQ-TREE). We made 10 000 ultra fast bootstraps to evaluate node support (Supplementary Figure S1, Tree S1).

The phylogenetic analysis of IntI was done using the sequences from complete integrons or In0 elements (*i.e.*, integrases identified by both HMM profiles) (Supplementary Figure S2). We added to this dataset some of the known integron-integrases of class 1, 2, 3, 4 and 5 retrieved from INTEGRALL. Given the previous phylogenetic analysis we used known XerC and XerD proteins to root the tree. Alignment and phylogenetic reconstruction were done using the same procedure; except that we built ten trees independently, and picked the one with best log-likelihood for the analysis (as recommended by the IQ-TREE authors (51)). The robustness of the branches was assessed using 1000 bootstraps (Supplementary Figure S2, Tree S2, Table S4).

### Pan-genomes

Pan-genomes are the full complement of genes in the species. They were built by clustering homologous proteins into families for each of the species (as previously described in (52)). Briefly, we determined the lists of putative homologs between pairs of genomes with BLASTP (53) (default parameters) and used the e-values ( $<10^{-4}$ ) to cluster them using SILIX (54). SILIX parameters were set such that a protein was homologous to another in a given family if the aligned part had at least 80% identity and if it included more than 80% of the smallest protein. We built pan-genomes for the 12 species having at least four complete genomes available in Genbank RefSeq and encoding at least one IntI. The genomes of these species carried 40% of the complete integrons in our dataset. We did not build a pan-genome for *Xanthomonas oryzae* because it contained too many rearrangements and repeated elements (55).

For a given species we computed the pattern of presence and absence of each integron-integrase protein family and the frequency of the integron-integrase within the species.

### Integron classification

We used two criteria to class integrons: frequency within the species' genomes (Supplementary Figure S3) and number of cassettes (Supplementary Figure S4). Integrons were classed as sedentary chromosomal integrons (as named by (4)) when their frequency in the pan-genome was 100%, or when they contained more than 19 *attC* sites. They were classed as mobile integrons when missing in more than 40% of the species' genomes, when present on a plasmid, or when the integron-integrase was from classes 1 to 5. The remaining integrons were classed as 'other'.

### Pseudo-genes detection

We translated the six reading frames of the region containing the CALIN elements (10 kb on each side) to detect *intI* pseudo-genes. We then ran hmmsearch with default options from HMMER suite v3.1b1 to search for hits matching the profile intI\_Cterm and the profile PF00589 among the translated reading frames. We recovered the hits with e-values lower than  $10^{-3}$  and alignments covering more than 50% of the profiles.

### IS detection

We identified insertion sequences (IS) by searching for sequence similarity between the genes present 4 kb around or within each genetic element and a database of IS from IS-Finder (56). Details can be found in (57).

### Detection of cassettes in INTEGRALL

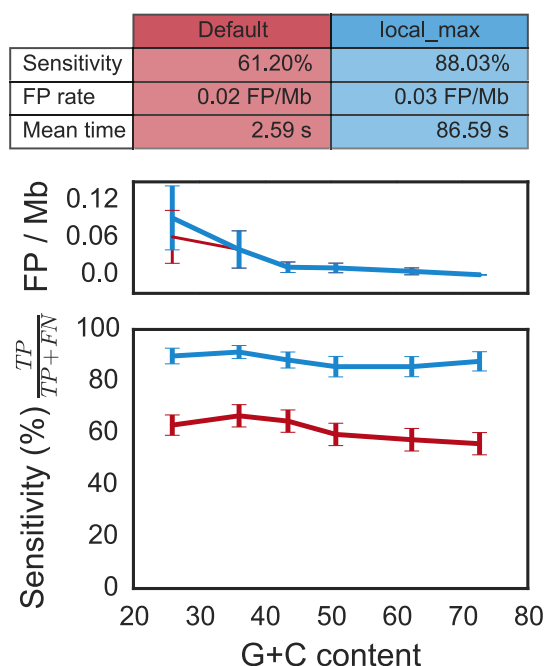
We searched for sequence similarity between all the CDS of CALIN elements and the INTEGRALL database using BLASTN from BLAST 2.2.30+. Cassettes were considered homologous to those of INTEGRALL when the BLASTN alignment showed more than 40% identity.

## RESULTS

### Models for *attC* sites

We selected a manually curated set of 291 *attC* sites representative of the diversity of sequences available in INTEGRALL (see Methods). We randomly sampled a third of them to build a covariance model of the *attC* site and set aside the others for subsequent validation. The characteristics of these sequences were studied in detail (Figure 2A), notably concerning the R and L boxes, the UCS and the EHB (15). The positions of the so-called Conserved Triplet (AAC and the complementary GTT) were more conserved than the others (Figure 2B and D). The length and sequence of the VTS were highly variable, between 20 and 100 nts long, as previously observed (1).

We used the covariance model to search for *attC* sites on 2484 complete bacterial genomes. The genomic *attC* sites showed stronger consensus sequences and more homogeneous VTS lengths than those used to build the model (Figure 2C). The analyses of sensitivity in the next paragraph show that our model missed very few sites. Hence, the differences between the initial and the genomic *attC* sites might



**Figure 4.** Quality assessment of the *attC* sites covariance model on pseudo-genomes with varying G+C content and depending on the run mode (default and ‘- - local\_max’). (Top) Table resuming the results. The mean time is the average running time per pseudo-genome on a Mac Pro, 2 × 2.4 GHz 6-Core Intel Xeon, 16 Gb RAM, with options - - cpu 20 and - - no-proteins. (Middle) Rate of false positives per mega-base (Mb) as function of the G+C content. (Bottom) Sensitivity (or true positive rate) as function of the G+C content. The red line depicts results obtained with the default parameters, and the blue line represents results obtained with the accurate parameters (‘- - local\_max’ option). Vertical lines represent standard error of the mean. There is no correlation with G+C content (all spearman  $\rho \in [-0.12; -0.04]$  and all  $P$ -values  $> 0.06$ ).

be due to our explicit option of using diverse sequences to build the model (to maximize diversity). They may also reflect differences between mobile integrons (very abundant in INTEGRALL) and integrons in sequenced bacterial genomes (where a sizeable fraction of cassettes were identified in *Vibrio* spp.).

We tested the ability of the covariance model to identify known sequences within pseudo-genomes built by randomizing dinucleotides from genomes with varying G+C content (Supplementary Table S5). We integrated in each pseudo-genome five *attC* sites (among the 195 of the validation set) at 2 kb intervals. We searched for *attC* sites in these genomes and found very few false positives in both run modes ( $\sim 0.03$  FP/Mb, Figure 4, see Methods for details). The proportion of true *attC* sites actually identified (sensitivity), was 61% for the default mode and 88% for the most accurate mode (with option ‘- - local\_max’). We identified at least two of the *attC* sites in 99% of the clusters (with the most accurate mode). Hence, clusters could be identified even when some *attC* sites were missed. The sensitivity of the model showed very little dependency on genome G+C composition in all cases (Figure 4).

We then searched for *attC* sites in sequences annotated for the presence of integrons in INTEGRALL (Supplemen-

tary Table S2a). The search was performed in 346 sequences containing 596 known *attC* sites. We found 570 *attC* sites with the most accurate mode (96% of sensitivity, Supplementary Table S2b). We missed 26 of the known *attC* sites, among which 15 were on the integron edges, and were probably missed because of the absence of R’ box on the 3’ side. All the 57 sequences annotated as In0 in INTEGRALL also lacked *attC* sites in our analysis. We found 247 *attC* sites missing in the annotations of INTEGRALL. If *attC* annotations in INTEGRALL were perfect and all these sites were false, the rate of false positives of our analysis would be 0.72 per Mb. However, about 90% of these non-annotated *attC* sites were found in clusters of two *attC* sites or more, which suggests that they are real *attC* sites. If all isolated *attC* sites were false positives (and the only ones), then the false positive rate would be 0.07 FP/Mb, i.e., less than one false *attC* site per genome.

These analyses showed a rate of false positives between 0.03 FP/Mb and 0.72 FP/Mb. The probability of having a cluster of two or more false *attC* sites by chance (within 4 kb) given this density of false positives is between  $4.10^{-6}$  and  $7.10^{-9}$  depending on the false positive rate (assuming a Poisson process). Hence, the clusters of *attC* sites given by our model are extremely unlikely to be false positives.

#### Identification of integron-integrases

We identified tyrosine recombinases using the protein profile PF00589 (from PFAM). To distinguish IntI from the other tyrosine recombinases, we built an additional protein profile corresponding to the IntI specific region near the patch III domain (17) (henceforth named intI.Cterm, see Methods). We found 215 proteins matching both profiles in the complete genomes of bacteria. Only six genes matched intI.Cterm but not PF00589. There were 18 808 occurrences of PF00589 not matching intI.Cterm, among which only 50 co-localized with an *attC* site. Among the latter, 29 were in genomes that encoded IntI elsewhere in the replicon (Supplementary Figure S5). The remaining 21 integrases were scattered in the phylogenetic tree of tyrosine recombinases, and only four of them were placed in an intermediate position between IntI and Xer (Supplementary Figure S1). These four sequences resembled typical phage integrases at the region of the patch III domain characteristic of IntI and they co-localized with very few *attC* sites (always less than three). This analysis strongly suggests that tyrosine recombinases lacking the intI.Cterm domain identified near *attC* sites are not IntI.

Most *intI* genes identified in bacterial genomes co-localized with *attC* sites (76%, Supplementary Figure S5). It is difficult to assess if the remaining *intI* genes are true or false, since In0 elements have often been described in the literature (47,58). We were able to identify IntI in the integrons of class 1 to class 5, as well as in well-known chromosomal integrons (e.g., in *Vibrio* super-integrons). We also identified all In0 elements in the INTEGRALL dataset mentioned above. Overall, these results show that IntI could be identified accurately using the intersection of both protein profiles.

We built a phylogenetic tree of the 215 IntI proteins identified in genomes (Supplementary Figure S2). Together with

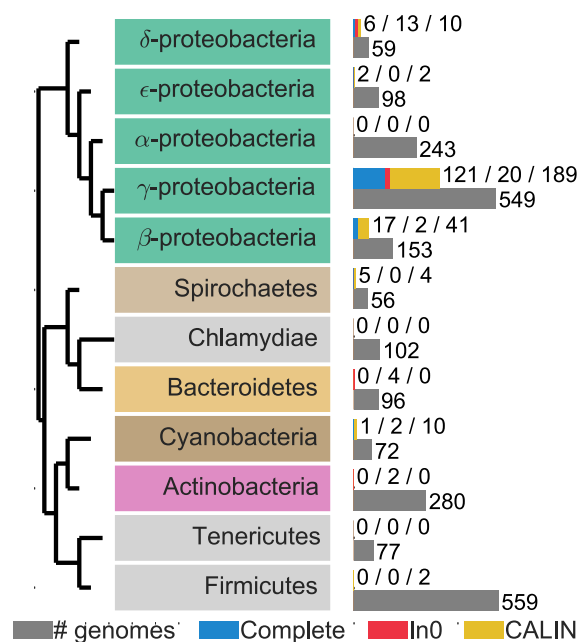
the analysis of the broader phylogenetic tree of tyrosine recombinases (Supplementary Figure S1), this extends and confirms previous analyses (1,7,22,59): (i) The XerC and XerD sequences are close outgroups. (ii) The IntI are monophyletic. (iii) Within IntI, there are early splits, first for a clade including class 5 integrons, and then for *Vibrio* super-integrons. On the other hand, a group of integrons displaying an integron-integrase in the same orientation as the *attC* sites (inverted integron-integrase group) was previously described as a monophyletic group (7), but in our analysis it was clearly paraphyletic (Supplementary Figure S2, column F). Notably, in addition to the previously identified inverted integron-integrase group of certain *Treponema* spp., a class 1 integron present in the genome of *Acinetobacter baumannii* 1656-2 had an inverted integron-integrase.

### Integrans in bacterial genomes

We built a program—IntegronFinder—to identify integrons in DNA sequences. This program searches for *intI* genes and *attC* sites, clusters them in function of their co-localization and then annotates cassettes and other accessory genetic elements (see Figure 3 and Methods). The use of this program led to the identification of 215 IntI and 4597 *attC* sites in complete bacterial genomes. The combination of this data resulted in a dataset of 164 complete integrons, 51 In0 and 279 CALIN elements (see Figure 1 for their description). The observed abundance of complete integrons is compatible with previous data (7). While most genomes encoded a single integron-integrase, we found 36 genomes encoding more than one, suggesting that multiple integrons are relatively frequent (20% of genomes encoding integrons). Interestingly, while the literature on antibiotic resistance often reports the presence of integrons in plasmids, we only found 24 integrons with integron-integrase (20 complete integrons, 4 In0) among the 2006 plasmids of complete genomes. All but one of these integrons were of class 1 (96%).

The taxonomic distribution of integrons was very heterogeneous (Figure 5 and Supplementary Figure S6). Some clades contained many elements. The foremost clade was the  $\gamma$ -Proteobacteria among which 20% of the genomes encoded at least one complete integron. This is almost four times as much as expected given the average frequency of these elements ( $\sim 6\%$ ,  $\chi^2$  test in a contingency table,  $P < 0.001$ ). The  $\beta$ -Proteobacteria also encoded numerous integrons ( $\sim 10\%$  of the genomes). In contrast, all the genomes of Firmicutes, Tenericutes and Actinobacteria lacked complete integrons. Furthermore, all 243 genomes of  $\alpha$ -Proteobacteria, the sister-clade of  $\gamma$  and  $\beta$ -Proteobacteria, were devoid of complete integrons, In0 and CALIN elements. Interestingly, much more distantly related bacteria such as Spirochaetes, Chlorobi, Chloroflexi, Verrucomicrobia and Cyanobacteria encoded integrons (Figure 5 and Supplementary Figure S6). The complete lack of integrons in one large phylum of Proteobacteria is thus very intriguing.

We searched for genes encoding antibiotic resistance in integron cassettes (see Methods). We identified such genes in 105 cassettes, i.e., in 3% of all cassettes from complete integrons (3116 cassettes). Most resistance cassettes were



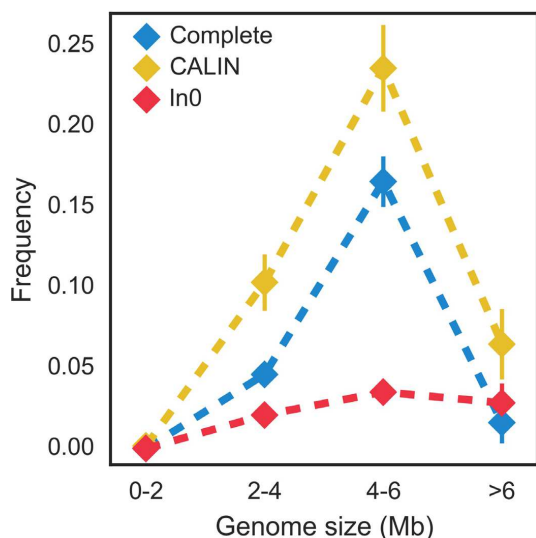
**Figure 5.** Taxonomic distribution of integrons in clades with more than 50 complete genomes sequenced. The gray bar represents the number of genomes sequenced for a given clade. The blue bar represents the number of complete integrons, the red bar the number of In0 and the yellow bar the number of CALIN. The colored text boxes refer to the colors in Supplementary Figure S2.

found in class 1 to 5 integrons (90% of them), even if the latter contained only 4.5% of all cassettes. This fits previous observations that integrons lacking antibiotic resistance determinants are very frequent in natural populations (24,30).

The association between genome size and the frequency of integrons has not been studied before. We binned the genomes in terms of their size and analyzed the frequency of complete integrons, In0 and CALIN. This showed a clearly non-monotonic trend (Figure 6). This distribution was not homogeneous in the different size categories ( $\chi^2$  test in a contingency table,  $P$ -values  $< 1.10^{-4}$  for complete, CALIN and In0). The same result was observed in a complementary analysis using only integrons from Gamma-Proteobacteria (Supplementary Figure S7). Very small genomes lack complete integrons, intermediate size genomes accumulate most of the integrons and the largest genomes encode few. Importantly, the same trends were observed for In0 and CALIN. Hence, the frequency of integrons is maximal for genomes of intermediate size (4–6 Mb).

### Unexpected abundance of CALIN elements

The number of *attC* sites lacking nearby integron-integrases was unexpectedly high. We found 431 occurrences of isolated single *attC* sites among the 1879 *attC* sites lacking an integrase. If these sites were all false, and were the only false ones, then the observed rate of false positives can be estimated at 0.047 FP/Mb. This is within the range of the rates of false positives observed in the sensitivity analysis (between 0.03 FP/Mb and 0.72 FP/Mb). The probability that



**Figure 6.** Frequency of integrons and related elements as a function of the genome size. Vertical bar represents standard error of the mean. The sample size in each bin is: 608 [0-2], 912 [2-4], 712 [4-6] and 247 [>6].

CALIN elements are false positives is exceedingly small for these rates of false positives. Therefore, we discarded single *attC* sites and kept the 279 clusters with two or more sites (CALIN) for the subsequent analyses. The CALIN resemble mobile integrons in terms of the number of cassettes: 83% had fewer than six *attC* sites and only 6.6% had more than 10 (Supplementary Figure S8). Nevertheless, some few CALIN were very large, with up to 114 *attC* sites. Furthermore, their cassettes were remarkably different from those of mobile integrons: only 147 out of the 1933 cassettes were homologous to those reported in INTEGRALL and only 31 carried antibiotic resistance genes (to be compared with 70% among class 1 to class 5 integrons and with 0.4% among the other complete integrons). Hence, CALIN are relatively small on average (5 *attC* sites) but may contain several tens of *attC* sites, and have many previously unknown gene cassettes.

The CALIN elements might have arisen from the loss of the integrase in a previously complete integron. Therefore, we searched for pseudogenes matching the specific IntI.Cterm domain less than 10 kb away from CALIN. We found such pseudo-genes near 15 out of 279 CALIN elements. It is worth noting that out of the 15 hits, 11 pseudo-genes were also matched by the PF00589 profile, which is consistent with the idea that they previously encoded IntI. Overall, our analysis showed that most CALIN (95%) are not close to recognizable *intI* pseudogenes.

We enquired on the possibility that some CALIN might actually be part of an integron and that we have missed this association because of the small 4 kb threshold used in the definition of the clusters. To test this hypothesis, we re-run IntegronFinder with a distance threshold of 10 kb. This analysis found 252 of the previously identified 279 CALIN elements, the remaining 27 being merged with a integron or In0 element with the 10 kb threshold. This shows that in-

creasing the distance threshold in the clustering procedure does not significantly change the observed abundance of CALIN.

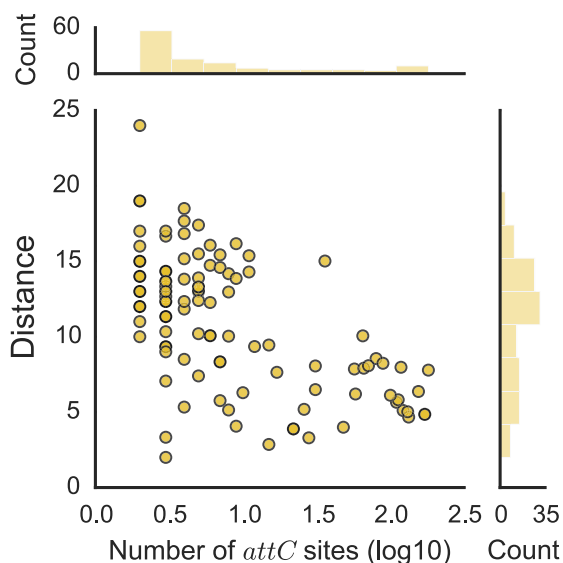
Chromosomal rearrangements (integrations, translocations or inversions) may split integrons and separate some cassettes from the neighborhood of the integron-integrase, thus producing CALIN elements in genomes encoding IntI. The CALIN elements might also result from integration of cassettes at secondary sites in the chromosome (9,10,60). We found some cases where IntI was actually encoded in another replicon (3.5% of CALIN). Overall, half of the CALIN were found in genomes encoding IntI and half in genomes lacking this gene.

Insertion sequences (IS) may create CALIN by promoting chromosomal rearrangements in a previously complete integron. The frequency of these events depends on the frequency of IS inside integrons. We therefore searched for IS inside or near CALIN, In0 and complete integrons (see Methods). We found that 12% of CALIN and 23% of the complete integrons encoded at least one IS within their cassettes. Upon IS-mediated rearrangements, the CALIN should be close to an IS. Indeed, 38% of the CALIN had a neighboring IS. Such co-localization was more frequent for CALIN in genomes encoding IntI than in the others ( $P < 0.001$ ,  $\chi^2$  contingency table). These results are consistent with the hypothesis that IS contribute to disrupt integrons and create CALIN. They may explain the origin of many CALIN elements, especially in the genomes encoding IntI in other locations.

#### Divergence of *attC* sites

Since *attC* sites were too poorly conserved in sequence to align using standard sequence alignment methods, we aligned them using the covariance model. We used these alignments to assess the sequence similarity between the R-UCS-L box and the difference in length of VTS sequences between *attC* sites. As expected, both measures showed that *attC* sites were more similar within than between integrons (Supplementary Figure S9). We then quantified the relationship between the number of *attC* sites in an integron and the average within-integron sequence dissimilarity in *attC* sites. The sequence similarity increased with the number of *attC* sites (Figure 7), i.e., the integrons carrying the longest arrays of cassettes had more homogeneous *attC* sites. Conversely, arrays of heterogeneous *attC* sites were almost always small.

Considering that many previous studies opposed mobile to chromosomal integrons, we tested if our results remained valid when following this dichotomy. We split our dataset into sedentary chromosomal integrons, mobile integrons and others (unclassified) (see Materials and Methods). Integrations from all three sets were found in the major clades of the IntI phylogeny (Supplementary Figure S2). Around 67% of the integrons encoded in chromosomes were classed as mobile in the species with computed pan-genomes (see Materials and Methods), showing that the separation between chromosomal and mobile integrons may be misleading. Expectedly, given their longer arrays of cassettes (Figure 7), the sedentary chromosomal integrons showed more similar *attC* sites than the mobile ones (Supplementary Fig-



**Figure 7.** Relationship between the number of *attC* sites in an integron and the mean sequence distance between *attC* sites within an integron. The x-axis is in log10 scale. The association is significant: spearman  $\rho = -0.53$ ,  $P < 0.001$ .

ure S9). The similarity of *attC* sites within CALIN elements was between that of sedentary and mobile integrons (Supplementary Figure S9). As proposed before (3), our results suggest that the dichotomy between sedentary chromosomal and mobile integrons may be informative because these two sets are quantitatively different, but may not reflect qualitative biological differences because there seems to be a continuum between large and small integrons.

## DISCUSSION

### IntegronFinder, limitations and perspectives

IntegronFinder identifies the vast majority of known *attC* sites and *intI* genes and is unaffected by genomic G+C content. The high sensitivity with which it identifies individual *attC* sites leads to a very small probability (0.02%) of missing all elements in a cluster of four *attC* sites. Nevertheless, it may be necessary to interpret with care the results of IntegronFinder in certain circumstances. For example, a genome rearrangement that splits an integron in two will result in the identification of a CALIN and an integron (eventually an In0 if the rearrangement takes place near the *attI* site). IntegronFinder accurately identifies these two genetic elements, which are independent from the transcriptional point of view since  $P_C$  cannot promote expression of the CALIN's cassettes. On the other hand, these elements may remain functionally linked because cassettes from the CALIN may be excised by the integron-integrase and reinserted in the integron at its *attI* site. It is unclear if the two elements should be regarded as independent, as it is done by default, or as a single integron. One should note that such cases might be difficult to distinguish from alternative evolutionary *scenarii* involving the loss of the integron-integrase in one of multiple integrons of a genome.

IntegronFinder detects few false positives among integrons and CALIN. Yet, we have identified 431 single *attC* sites in bacterial genomes whose relevance is less clear. Some of these sites might be false positives because their frequency in genomes is close to the upper limit of the false positive rates obtained in our validation procedure. Others might result from the genetic degradation of integron cassettes.

Our study was restricted to the analysis of complete bacterial genomes to avoid the complications of dealing with inaccurate genome assemblies. However, IntegronFinder can be used to analyze draft genomes or metagenomes as long as one is aware of the limitations of the procedure in such data. The difficulty in the analysis of draft genomes results from the presence of contig breaks that often coincide with repeated sequences, such as transposable elements. Their high frequency in integrons implicate that these might be scattered in different contigs. Under these circumstances, IntegronFinder will identify several genetic elements (typically an integron and several CALIN) even if the genome actually encodes one single complete integron. Metagenomics data are even more challenging because it includes numerous small contigs where it is difficult to identify complete integrons. Yet, since the models for *attC* sites and *intI* are very accurate they can be used to identify cassettes and integron-integrases in assembled metagenomes. This might dramatically improve the detection of novel gene cassettes in environmental data.

### Determinants of integron distribution

Our analysis highlighted associations between the frequency of integrons and certain genetic traits. The frequency of CALIN, complete integrons and In0 is often highly correlated in relation to all of these traits, e.g., all three types of elements show roughly similar distributions among bacterial phyla and in terms of genome size. This association between the three types of elements is most likely caused by their common evolutionary history.

Integrons have well-known roles in the spread of antibiotic resistance. Nevertheless, we identified very few known antibiotic resistance genes in complete integrons outside the class 1 to class 5 integrons. Interestingly, we also found few resistance genes in CALIN elements. This supports previous suggestions that integrons carry a much broader set of adaptive traits, than just antibiotic resistance, in natural populations (30).

We found an under-representation of integrons in both small and large bacterial genomes. Since integrons are gene-capturing platforms, one would expect a positive association between the frequency of integrons and that of horizontal transfer. Accordingly, the lack of integrons in bacteria with small genomes might be caused by the sexual isolation endured by these bacteria, which typically also have few or no transposable elements, plasmids or phages (61–63). The causes for the low frequency of integrons in the largest genomes must be different, since they are thought to engage in very frequent horizontal transfer (64,65). We can only offer a speculation to explain this puzzling result. Horizontal transfer is often brought by mobile genetic elements. These elements can be very large and costly, while encod-

ing few adaptive traits (if any) (66). The cost of these elements should scale with the inverse of genome size, if larger genomes have fewer constraints on the amount of incoming genetic material and if they select for more frequent horizontal transfer. Hence, the distribution of integrons might result from the combined effect of the frequency of transfer (increasing with genome size) and selection for compact transfer (decreasing with genome size). Further work will be necessary to test this hypothesis.

Most integrons with taxonomic identification available in INTEGRALL are from  $\gamma$ -proteobacteria (90%) (38). Our dataset is more diverse; we found many integrons in  $\beta$ -Proteobacteria and in other large phyla (such as Spirochaetes, Chloroflexi, Chlorobi or Planctomycetes). This shows that our method identifies integrons in clades distant from  $\gamma$ -proteobacteria. Surprisingly, the genomes from  $\alpha$ -Proteobacteria had no integrons, even if they encoded many tyrosine recombinases involved in the integration of a variety of mobile genetic elements. The complete absence of integrons, In0 and CALIN in  $\alpha$ -Proteobacteria is extremely puzzling. It cannot solely be ascribed to the frequency of small genomes in certain branches of  $\alpha$ -Proteobacteria, since our dataset included 99 genomes larger than 4 Mb in the clade. We also did not find complete integrons in Gram-positive bacteria. It is well known that differences in the translation machinery hinder the expression of transferred genetic information from Proteobacteria to Firmicutes (e.g., due specificities of the protein S1 (67)), but these differences cannot explain the lack of integrons in  $\alpha$ -Proteobacteria. Transfer of genetic information between clades of Proteobacteria and between Proteobacteria and Gram-positive bacteria is well documented (68,69). Accordingly, integrons have occasionally been identified in Firmicutes and  $\alpha$ -Proteobacteria (38,70), and we found CALIN in Firmicutes and In0 in Actinobacteria. Some unknown mechanism probably hinders the establishment of integrons in these lineages after transfer.

### The evolution of integrons

Our study sheds new light on the evolution of integrons. The use of the covariance model confirmed that *attC* sites are more similar within than between integrons. It also uncovered a positive association between the homogeneity of *attC* sites and the number of cassettes in integrons. If homogeneous *attC* sites result from the creation of cassettes by the integron-integrase of the same integron, then the largest integrons might be those creating more cassettes.

We found many CALIN in genomes. Previous works have identified *intI* pseudo-genes in bacterial genomes (22,29), and showed IntI-mediated creation of CALIN at secondary integration sites (9,10,60). However, the frequency of CALIN, and especially in genomes lacking integron-integrases, is surprising. These elements may have arisen in several ways. (i) By the unknown mechanism creating novel cassettes if this mechanism does not depend on IntI. (ii) By integration of cassettes at secondary integration sites by an integron encoded elsewhere in the genome. (iii) By loss of *intI*, even if most CALIN lacked recognizable neighboring pseudogenes of *intI*. (iv) By genome rearrangements separating a group of cassettes from the neighborhood of *intI*

(as observed in (71)). Mechanisms #3 and #4 are consistent with the presence of IS in a fourth of complete integrons. Mechanisms #1, #2 and #4 might explain why half of the CALIN are in replicons encoding IntI.

There are some similarities, but also key differences, between CALIN and mobile integrons. The average number of cassettes is similar in both elements, but the within-element sequence identity of *attC* sites is different and few CALIN have many cassettes. CALIN have very few cassettes homologous to those of mobile integrons and far fewer antibiotic resistant genes. This shows that most CALIN do not derive from the class 1 to 5 mobile integrons carrying antibiotic resistance genes.

The lack of an integron-integrase in CALIN elements does not imply that these cassettes cannot be mobilized. We found many co-occurrences of complete integrons, In0 and CALIN in genomes. They might facilitate the exchange of cassettes between elements. Integron-integrases have relaxed sequence similarity requirements to mediate recombination between divergent *attC* sites. It is thus tempting to speculate that integrons transferred into a genome encoding a CALIN might be able to integrate CALIN cassettes in their own cluster of cassettes. Alternatively, CALIN might provide cassettes to naturally transformable bacteria, in case their stable circular forms are able to survive in the environment and be taken up by transformation (72). If integrons often capture cassettes from CALIN elements then many genomes currently lacking integrons but carrying CALIN might be important reservoirs of novel cassettes.

It is unknown whether CALIN genes are expressed or have an adaptive value. Their high abundance in genomes suggests that at least some of them might provide advantageous traits for the host bacterium. CALIN with very degenerate *attC* sites might thus represent an intermediate step between the acquisition of a gene by an integron and its definitive stabilization in the genome by loss of the IntI-based cassette mobilizing activity.

### AVAILABILITY

The program was written in Python 2.7. It is freely available on a webserver (<http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::integron.finder>). The standalone program is distributed under an open-source GPLv3 license and can be downloaded from Github (<https://github.com/gem-pasteur/Integron.Finder/>) to be run using the command line. Supplementary materials include tables containing all integrons found at different level (elements, integrons, genomes, in Supplementary Tables S6, S7 and S8). It includes the list of the 596 *attC* sites with their annotated position (Supplementary Table S2a), and the corresponding file with observed position (Supplementary Table S2b). We provide the intI\_Cterm HMM profile (File S2) and the covariance model for the *attC* site (File S3).

### SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

J.C. is a member of the 'Ecole Doctorale Frontière du Vivant (FdV) – Programme Bettencourt'. We thank Didier Mazel, Jose A. Escudero, Céline Loot, Aleksandra Nivina, Philippe Glaser, Claudine Médigue, Alexandra Moura and Julian E. Davies for fruitful discussions and comments on the manuscript.

*Author contributions:* J.C. EPCR designed the study. J.C. made the analysis. J.C. and B.N. wrote the software and webserver. M.T. T.J. contributed with data. J.C. EPCR. drafted the manuscript. All authors contributed to the final text of the manuscript.

## FUNDING

European Research Council [EVOMOBILOME, 281605 to E.P.C.R.]. Funding for open access charge: ERC EVOMOBILOME.

*Conflict of interest statement.* None declared.

## REFERENCES

- Mazel,D. (2006) Integrans: agents of bacterial evolution. *Nat. Rev. Microbiol.*, **4**, 608–620.
- Partridge,S.R. (2011) Analysis of antibiotic resistance regions in Gram-negative bacteria. *FEMS Microbiol. Rev.*, **35**, 820–855.
- Gillings,M.R. (2014) Integrans: past, present, and future. *Microbiol. Mol. Biol. Rev.*, **78**, 257–277.
- Escudero,J.A., Loot,C., Nivina,A. and Mazel,D. (2015) The Integron: adaptation on demand. *Microbiol. Spectr.*, **3**, MDNA3-0019-2014.
- Collis,C.M. and Hall,R.M. (1995) Expression of antibiotic resistance genes in the integrated cassettes of integrans. *Antimicrob. Agents Chemother.*, **39**, 155–162.
- Joss,M.J., Koenig,J.E., Labbate,M., Polz,M.F., Gillings,M.R., Stokes,H.W., Doolittle,W.F. and Boucher,Y. (2009) ACID: annotation of cassette and integron data. *BMC Bioinformatics*, **10**, 118.
- Boucher,Y., Labbate,M., Koenig,J.E. and Stokes,H.W. (2007) Integrans: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.*, **15**, 301–309.
- Michael,C.A. and Labbate,M. (2010) Gene cassette transcription in a large integron-associated array. *BMC Genet.*, **11**, 82.
- Recchia,G.D., Stokes,H.W. and Hall,R.M. (1994) Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic Acids Res.*, **22**, 2071–2078.
- Recchia,G.D. and Hall,R.M. (1995) Plasmid evolution by acquisition of mobile gene cassettes: plasmid pIE723 contains the *aadB* gene cassette precisely inserted at a secondary site in the *incQ* plasmid RSF1010. *Mol. Microbiol.*, **15**, 179–187.
- Hall,R.M., Brookes,D.E. and Stokes,H.W. (1991) Site-specific insertion of genes into integrans: role of the 59-base element and determination of the recombination cross-over point. *Mol. Microbiol.*, **5**, 1941–1959.
- Nunes-Duby,S.E., Kwon,H.J., Tirumalai,R.S., Ellenberger,T. and Landy,A. (1998) Similarities and differences among 105 members of the *Int* family of site-specific recombinases. *Nucleic Acids Res.*, **26**, 391–406.
- Collis,C.M., Recchia,G.D., Kim,M.J., Stokes,H.W. and Hall,R.M. (2001) Efficiency of recombination reactions catalyzed by class I integron integrase *IntI1*. *J. Bacteriol.*, **183**, 2535–2542.
- MacDonald,D., Demarre,G., Bouvier,M., Mazel,D. and Gopaul,D.N. (2006) Structural basis for broad DNA-specificity in integron recombination. *Nature*, **440**, 1157–1162.
- Bouvier,M., Ducos-Galand,M., Loot,C., Bikard,D. and Mazel,D. (2009) Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genet.*, **5**, e1000632.
- Fruerie,C., Ducos-Galand,M., Gopaul,D.N. and Mazel,D. (2010) The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res.*, **38**, 559–569.
- Messier,N. and Roy,P.H. (2001) Integron integrases possess a unique additional domain necessary for activity. *J. Bacteriol.*, **183**, 6699–6706.
- Partridge,S.R., Tsafnat,G., Coiera,E. and Iredell,J.R. (2009) Gene cassettes and cassette arrays in mobile resistance integrans. *FEMS Microbiol. Rev.*, **33**, 757–784.
- Hall,R.M. and Collis,C.M. (1995) Mobile gene cassettes and integrans: capture and spread of genes by site-specific recombination. *Mol. Microbiol.*, **15**, 593–600.
- Rowe-Magnus,D.A., Guerout,A.M., Biskri,L., Bougie,P. and Mazel,D. (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the *Vibrionaceae*. *Genome Res.*, **13**, 428–442.
- Diaz-Mejia,J.J., Amabile-Cuevas,C.F., Rosas,I. and Souza,V. (2008) An analysis of the evolutionary relationships of integron integrases, with emphasis on the prevalence of class I integrans in *Escherichia coli* isolates from clinical and environmental origins. *Microbiology*, **154**, 94–102.
- Nemergut,D.R., Robeson,M.S., Kysela,R.F., Martin,A.P., Schmidt,S.K. and Knight,R. (2008) Insights and inferences about integron evolution from genomic data. *BMC Genomics*, **9**, 261.
- Hall,R.M. (2012) Integrans and gene cassettes: hotspots of diversity in bacterial genomes. *Ann. N Y Acad. Sci.*, **1267**, 71–78.
- Gillings,M., Boucher,Y., Labbate,M., Holmes,A., Krishnan,S., Holley,M. and Stokes,H.W. (2008) The evolution of class I integrans and the rise of antibiotic resistance. *J. Bacteriol.*, **190**, 5095–5100.
- Mazel,D., Dychinco,B., Webb,V.A. and Davies,J. (1998) A distinctive class of integron in the *Vibrio cholerae* genome. *Science*, **280**, 605–608.
- Rowe-Magnus,D.A., Guerout,A.M., Ploncard,P., Dychinco,B., Davies,J. and Mazel,D. (2001) The evolutionary history of chromosomal super-integrans provides an ancestry for multiresistant integrans. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 652–657.
- Hochhut,B., Lotfi,Y., Mazel,D., Faruque,S.M., Woodgate,R. and Waldor,M.K. (2001) Molecular analysis of antibiotic resistance gene clusters in *vibrio cholerae* O139 and O1 SXT constins. *Antimicrob. Agents Chemother.*, **45**, 2991–3000.
- Iwanaga,M., Toma,C., Miyazato,T., Insiengmay,S., Nakasone,N. and Ehara,M. (2004) Antibiotic resistance conferred by a class I integron and SXT constin in *Vibrio cholerae* O1 strains isolated in Laos. *Antimicrob. Agents Chemother.*, **48**, 2364–2369.
- Gillings,M.R., Holley,M.P., Stokes,H.W. and Holmes,A.J. (2005) Integrans in *Xanthomonas*: a source of species genome diversity. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 4419–4424.
- Holmes,A.J., Gillings,M.R., Nield,B.S., Mabbutt,B.C., Nevalainen,K.M. and Stokes,H.W. (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ. Microbiol.*, **5**, 383–394.
- Moura,A., Henriques,I., Ribeiro,R. and Correia,A. (2007) Prevalence and characterization of integrans from bacteria isolated from a slaughterhouse wastewater treatment plant. *J. Antimicrobial Chemother.*, **60**, 1243–1250.
- Stalder,T., Barraud,O., Casellas,M., Dagot,C. and Ploy,M.C. (2012) Integron involvement in environmental spread of antibiotic resistance. *Front. Microbiol.*, **3**, 119.
- Gillings,M.R., Gaze,W.H., Pruden,A., Smalla,K., Tiedje,J.M. and Zhu,Y.G. (2015) Using the class I integron-integrase gene as a proxy for anthropogenic pollution. *ISME J.*, **9**, 1269–1279.
- Tsafnat,G., Coiera,E., Partridge,S.R., Schaeffer,J. and Iredell,J.R. (2009) Context-driven discovery of gene cassettes in mobile integrans using a computational grammar. *BMC Bioinformatics*, **10**, 281.
- Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
- Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
- Neron,B., Menager,H., Maufrais,C., Joly,N., Maupetit,J., Letort,S., Carrere,S., Tuffery,P. and Letondal,C. (2009) Mobyly: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005–3011.
- Moura,A., Soares,M., Pereira,C., Leitao,N., Henriques,I. and Correia,A. (2009) INTEGRALL: a database and search engine for integrans, integrases and gene cassettes. *Bioinformatics*, **25**, 1096–1098.
- Cambay,G., Sanchez-Alberola,N., Campoy,S., Guerin,E., Da Re,S., Gonzalez-Zorn,B., Ploy,M.C., Barbe,J., Mazel,D. and Erill,I. (2011)




- Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mobile DNA*, **2**, 6.
40. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
  41. 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.
  42. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
  43. Gibson, M.K., Forsberg, K.J. and Dantas, G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
  44. Finn, R.D., Tate, J., Misty, J., Coghill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
  45. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
  46. 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.
  47. Bissonnette, L. and Roy, P.H. (1992) Characterization of In0 of *Pseudomonas aeruginosa* plasmid pVS1, an ancestor of integrons of multiresistance plasmids and transposons of gram-negative bacteria. *J. Bacteriol.*, **174**, 1248–1257.
  48. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.
  49. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
  50. Criscuolo, A. and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.*, **10**, 210.
  51. 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.
  52. Touchon, M., Cury, J., Yoon, E.-J., Krizova, L., Cerqueira, G.C., Murphy, C., Feldgarden, M., Wortman, J., Clermont, D., Lambert, T. *et al.* (2014) The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. *Genome Biol. Evol.*, **6**, 2866–2882.
  53. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
  54. Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
  55. Lee, B.M., Park, Y.J., Park, D.S., Kang, H.W., Kim, J.G., Song, E.S., Park, I.C., Yoon, U.H., Hahn, J.H., Koo, B.S. *et al.* (2005) The genome sequence of *Xanthomonas oryzae* pathovar *oryzae* KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Res.*, **33**, 577–586.
  56. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.*, **34**, D32–D36.
  57. Touchon, M. and Rocha, E.P. (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol. Biol. Evol.*, **24**, 969–981.
  58. Brown, H.J., Stokes, H.W. and Hall, R.M. (1996) The integrons In0, In2, and In5 are defective transposon derivatives. *J. Bacteriol.*, **178**, 4429–4437.
  59. Cambray, G., Guerout, A.M. and Mazel, D. (2010) Integrons. *Annu. Rev. Genet.*, **44**, 141–166.
  60. Segal, H., Victoria Francia, M., Garcia Lobo, J.M. and Elisha, G. (1999) Reconstruction of an active integron recombination site after integration of a gene cassette at a secondary site. *Antimicrob. Agents Chemother.*, **43**, 2538–2541.
  61. Silva, F.J., Latorre, A. and Moya, A. (2003) Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet.*, **19**, 176–180.
  62. Canback, B., Tamas, I. and Andersson, S.G. (2004) A phylogenomic study of endosymbiotic bacteria. *Mol. Biol. Evol.*, **21**, 1110–1122.
  63. McCutcheon, J.P. and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.*, **10**, 13–26.
  64. Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
  65. Cordero, O.X. and Hogeweg, P. (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc. Natl Acad. Sci. U.S.A.*, **106**, 21748–21753.
  66. Baltrus, D.A. (2013) Exploring the costs of horizontal gene transfer. *Trends Ecol. Evol.*, **28**, 489–495.
  67. Salah, P., Bisaglia, M., Aliprandi, P., Uzan, M., Sizun, C. and Bontems, F. (2009) Probing the relationship between Gram-negative and Gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res.*, **37**, 5578–5588.
  68. Mazodier, P. and Davies, J. (1991) Gene transfer between distantly related bacteria. *Annu. Rev. Genet.*, **25**, 147–171.
  69. Kloesges, T., Popa, O., Martin, W. and Dagan, T. (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol. Biol. Evol.*, **28**, 1057–1074.
  70. Nandi, S., Maurer, J.J., Hofacre, C. and Summers, A.O. (2004) Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc. Natl Acad. Sci. U.S.A.*, **101**, 7118–7122.
  71. Le Roux, F., Zouine, M., Chakroun, N., Binesse, J., Saulnier, D., Bouchier, C., Zidane, N., Ma, L., Rusniok, C., Lajus, A. *et al.* (2009) Genome sequence of *Vibrio splendidus*: an abundant planktonic marine species with a large genotypic diversity. *Environ. Microbiol.*, **11**, 1959–1970.
  72. Gestal, A.M., Liew, E.F. and Coleman, N.V. (2011) Natural transformation with synthetic gene cassettes: new tools for integron research and biotechnology. *Microbiology*, **157**, 3349–3360.
  73. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

### **2.2.2 Article 5: Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance**

This paper aims at better understanding the relationships between mobile integrons and sedentary chromosomal integrons and their relative dynamic when it comes to capture and exchange cassettes. This work was made in collaboration with experimentalist biologists, expert in the integron recombination system.



# Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance

Céline Loot,<sup>a,b</sup> Aleksandra Nivina,<sup>a,b,c</sup> Jean Cury,<sup>b,d</sup> José Antonio Escudero,<sup>a,b</sup> Magaly Ducos-Galand,<sup>a,b</sup> David Bikard,<sup>a,b</sup> Eduardo P. C. Rocha,<sup>b,d</sup>  Didier Mazel<sup>a,b</sup>

Unité de Plasticité du Génome Bactérien, Institut Pasteur, Paris, France<sup>a</sup>; Centre National de la Recherche Scientifique UMR 3525, Paris, France<sup>b</sup>; Université Paris Descartes, Sorbonne Paris Cité, Paris, France<sup>c</sup>; Microbial Evolutionary Genomics Unit, Institut Pasteur, Paris, France<sup>d</sup>

**ABSTRACT** Integrons ensure a rapid and “on demand” response to environmental stresses driving bacterial adaptation. They are able to capture, store, and reorder functional gene cassettes due to site-specific recombination catalyzed by their integrase. Integrons can be either sedentary and chromosomally located or mobile when they are associated with transposons and plasmids. They are respectively called sedentary chromosomal integrons (SCIs) and mobile integrons (MIs). MIs are key players in the dissemination of antibiotic resistance genes. Here, we used *in silico* and *in vivo* approaches to study cassette excision dynamics in MIs and SCIs. We show that the orientation of cassette arrays relative to replication influences *attC* site folding and cassette excision by placing the recombinogenic strands of *attC* sites on either the leading or lagging strand template. We also demonstrate that stability of *attC* sites and their propensity to form recombinogenic structures also regulate cassette excision. We observe that cassette excision dynamics driven by these factors differ between MIs and SCIs. Cassettes with high excision rates are more commonly found on MIs, which favors their dissemination relative to SCIs. This is especially true for SCIs carried in the *Vibrio* genus, where maintenance of large cassette arrays and vertical transmission are crucial to serve as a reservoir of adaptive functions. These results expand the repertoire of known processes regulating integron recombination that were previously established and demonstrate that, in terms of cassette dynamics, a subtle trade-off between evolvability and genetic capacitance has been established in bacteria.

**IMPORTANCE** The integron system confers upon bacteria a rapid adaptation capability in changing environments. Specifically, integrons are involved in the continuous emergence of bacteria resistant to almost all antibiotic treatments. The international situation is critical, and in 2050, the annual number of deaths caused by multiresistant bacteria could reach 10 million, exceeding the incidence of deaths related to cancer. It is crucial to increase our understanding of antibiotic resistance dissemination and therefore integron recombination dynamics to find new approaches to cope with the worldwide problem of multiresistance. Here, we studied the dynamics of recombination and dissemination of gene encoding cassettes carried on integrons. By combining *in silico* and *in vivo* analyses, we show that cassette excision is highly regulated by replication and by the intrinsic properties of cassette recombination sites. We also demonstrated differences in the dynamics of cassette recombination between mobile and sedentary chromosomal integrons (MIs and SCIs). For MIs, a high cassette recombination rate is favored and timed to conditions when generating diversity (upon which selection can act) allows for a rapid response to environmental conditions and stresses. In contrast, for SCIs, cassette excisions are less frequent, limiting cassette loss and ensuring a large pool of cassettes. We therefore confirm a role of SCIs as reservoirs of adaptive functions and demon-

Received 20 December 2016 Accepted 3 March 2017 Published 28 March 2017

**Citation** Loot C, Nivina A, Cury J, Escudero JA, Ducos-Galand M, Bikard D, Rocha EPC, Mazel D. 2017. Differences in integron cassette excision dynamics shape a trade-off between evolvability and genetic capacitance. *mBio* 8:e02296-16. <https://doi.org/10.1128/mBio.02296-16>.

**Editor** Julian E. Davies, University of British Columbia

**Copyright** © 2017 Loot et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Didier Mazel, [didier.mazel@pasteur.fr](mailto:didier.mazel@pasteur.fr).

C.L. and A.N. contributed equally to this work.

strate that the remarkable adaptive success of integron recombination system is due to its intricate regulation.

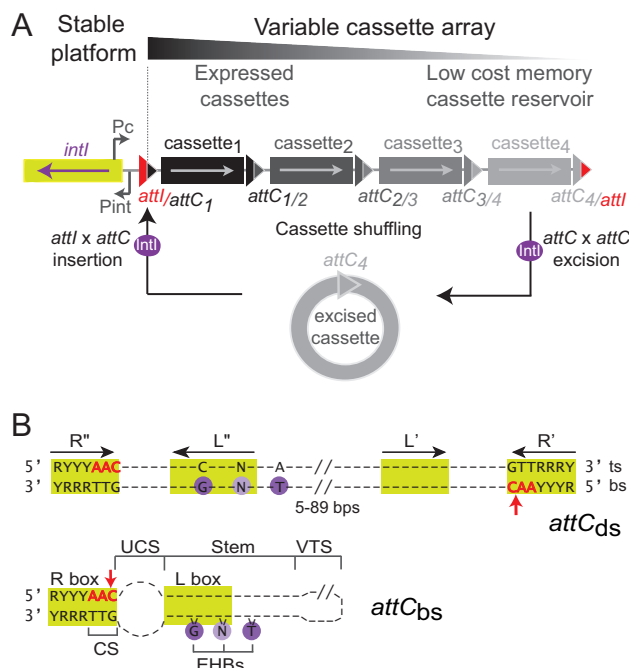
**KEYWORDS** integron, *attC* sites, cassette dynamics, replication, site-specific recombination

**A**ntibiotics are essential to the success of modern medicine, but their efficacy has been impeded by the emergence of multiresistant bacteria. In 1989, integrons were identified as systems responsible for the dissemination of resistance genes among Gram-negative bacterial pathogens (1, 2), primarily due to their association with transposable elements and conjugative plasmids. The aforementioned systems were later named mobile integrons (MIs) as opposed to sedentary chromosomally located integrons (SCIs), which are found in Gram-negative bacteria from various environments and play a general role in bacterial evolution (3). The integron is a powerful genetic system that enables bacterial evolution by capturing, stockpiling, and reordering cassette-encoding proteins with potentially advantageous functions for adaptation to changing environments (antibiotic resistance, virulence, interaction with phages [4–8]).

All integrons share a common structure composed of a stable platform and a variable cassette array. The stable platform contains the following: (i) a gene encoding the integron integrase (*intI*), a site-specific tyrosine recombinase which catalyzes cassette rearrangements; (ii) a primary recombination site for the insertion of cassettes, *attI*; and (iii) a promoter, *P<sub>c</sub>*, driving the expression of proximal cassettes in the array (Fig. 1A). The cassettes in the variable cassette array generally consist of a promoterless gene (coding sequence [CDS]) and a cassette recombination site (*attC*). Cassette arrays represent a low-cost repository of valuable functions for the cell and most likely reflect a history of adaptive events. The number of cassettes in the array can be very large in SCIs (more than 200), while it rarely exceeds eight in MIs (9, 10). Interestingly, *attC* sites found in SCI cassette arrays generally show a high degree of sequence identity, which increases with the number of cassettes (10). In contrast, *attC* sites of MI cassette arrays differ in length and sequence (11).

Integrations are atypical site-specific recombination systems. Unlike the *attI* sites, which are recombined in the classical double-stranded (ds) form, *attC* sites are recombined in a single-stranded (ss) folded form (Fig. 1B) (12–15). More precisely, the bottom strand of the *attC* site recombines about  $10^3$  times more frequently than the top strand (14). The preference for the bottom strand ensures that cassettes are inserted in the correct orientation relative to the *P<sub>c</sub>* promoter, allowing their expression (16, 17). In contrast to canonical site-specific recombination sites, the recognition of *attC* sites does not rely on the nature of their primary sequence but rather on the structure of their folded single-stranded DNA (ssDNA), as they share only 3 conserved nucleotides at the cleavage site (18). Folded *attC* sites form imperfect hairpins containing three unpaired structural features, the extrahelical bases (EHBs), the unpaired central spacer (UCS), and the variable terminal structure (VTS), which ensure strand selectivity and high levels of *attC* recombination (16–18). The variability of *attC* site length (from 57 to 141 bp) is mostly due to differences in the VTS loop (Fig. 1B) (19), which ranges from three unpaired nucleotides as in the *attC* site of the *aadA7* gene (*attC<sub>aadA7</sub>* site) to complex branched secondary structures in larger sites such as *Vibrio cholerae* repeat (VCR) sites (the *attC* sites from *V. cholerae* SCI [20]). Due to the ss nature of the *attC* site, the *attI* × *attC* recombination generates, after the first strand exchange, an atypical and asymmetric Holliday junction. To complete the recombination event, this Holliday junction has to be resolved through replication (21).

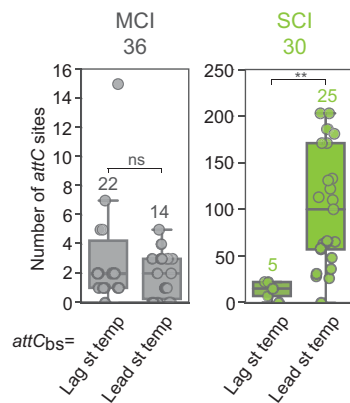
In terms of cassette dynamics, recombination between *attC* sites leads to the excision of a cassette from the array. Recombination between the *attC* site of the excised cassette and the *attI* site allows for the reinsertion of the cassette in the beginning of the array, placing it downstream of the *P<sub>c</sub>* promoter where it is more highly expressed. Hence, cycles of excision and insertion shuffle cassettes in the array and change their expression patterns



**FIG 1** The integron system. (A) Organization of integrons. Functional platform composed of the *intI* gene encoding the integrase (green rectangle), the cassette promoter ( $P_c$ ) and the integrase promoter ( $P_{int}$ ), as well as the primary *attI* recombination site (red triangle) are shown. Integrase (IntI; purple circle) catalyzes cassette excision (*attC* × *attC*) followed by insertion (*attI* × *attC*) of the excised cassette (gray circle). Hybrid *att* sites are indicated. Arrows inside the cassettes indicate the direction of their coding sequence (CDS), and the color intensity reflects the expression level of cassettes: only the first several cassettes of the array are expressed, while the subsequent ones can be seen as a low-cost cassette reservoir. (B) *attC* recombination sites. The double- and single-stranded *attC* sites (*attC*<sub>ds</sub> and *attC*<sub>bs</sub>) are shown. Green boxes show putative IntI1 binding sites, and red arrows show the cleavage point. For the *attC*<sub>ds</sub>, inverted repeats (R'', L'', L', and R') are indicated by black arrows. The conserved nucleotides are indicated, and violet circles show the conserved G nucleotide and the other bases, which constitute the extrahelical bases (EHBs) in folded *attC* sites (see *attC*<sub>bs</sub>). The top strand (ts) and bottom strand (bs) are marked. The structure of *attC*<sub>bs</sub> was determined by the RNAfold program from ViennaRNA 2 package (Materials and Methods). Structural features, namely, the unpaired central spacer (UCS), the EHBs, the stem, and the variable terminal structure (VTS), as well as the conserved sequence (CS), are indicated. R, purine; Y, pyrimidine; N, any base.

(Fig. 1A) (22, 23). In addition, excision and loss of cassettes close to  $P_c$  may increase the expression of downstream cassettes. Such events observed in clinical settings (24) are probably more cost-effective than excisions followed by insertions (25). Stress responses, especially the SOS response, increase the expression of the integrase and accelerate the dynamics of integron shuffling and dissemination. This ensures a rapid and “on demand” adaptation to novel environmental contexts and limits pleiotropic effects in the host bacterium (26, 27). The entrance of ssDNA into the cell by conjugation or transformation can also induce the SOS response, thus coupling integrase expression to moments when incoming DNA could supply novel cassettes (23, 28).

Overall cassette dynamics depend on both cassette excision and insertion. The balance between the two processes determines whether the array accumulates or loses cassettes over time. Since cassette excision is a prerequisite step for further cassette insertion and as its regulation influences the rate of both processes, we focused our studies on excision dynamics. The rate of cassette excision must be a result of a trade-off between evolvability and genetic capacitance. The rate needs to be high enough to ensure shuffling and dissemination of cassettes (and the adaptive functions they encode). However, if the rate is too high and the balance between cassette excision and insertion is shifted toward excision, then cassettes could be rapidly lost, decreasing the probability of their vertical transmission. Since cassette excision directly



**FIG 2** Distribution of the number of *attC* sites per integron (mobile chromosomal integron [MCI] and sedentary chromosomal integron [SCI]) as a function of *attC<sub>bs</sub>* orientation relative to replication. Circles correspond to the number of *attC* sites for each of the analyzed 36 MCIs and 30 SCIs. Tests of the differences between data sets were performed using the Wilcoxon rank sum test (\*\*,  $P$  value of  $1.25 \times 10^{-3}$ ; ns, not significant). bs, bottom strand; Lag st temp, lagging strand template; Lead st temp, leading strand template.

depends on simultaneous folding of consecutive *attC* sites, the regulation of cassette excision is dependent on *attC* site folding, meaning that bacteria must regulate it subtly. This is particularly important because the presence of stable and long hairpin structures can also be detrimental for the maintenance of bacterial genomes (29). We have previously demonstrated that there is a subtle equilibrium between opposite processes: on one hand, *attC* site integrity ensured by the single-stranded DNA binding (SSB) protein which hampers folding of *attC* sites in the absence of the integrase (30); on the other hand, *attC* site folding and recombination favored by the availability of ssDNA (for instance, during conjugation and replication) and by the propensity to form cruciform structures due to supercoiling (31).

In order to gain a better understanding of cassette excision dynamics in integrons, we performed both *in silico* analyses and *in vivo* experiments to study the parameters that play important roles in maintaining an adequate level of cassette excision: orientation of cassette arrays relative to replication, CDS lengths, and *attC* site properties. Finally, we discuss the results obtained for MCIs and SCIs in terms of integron evolutionary biology.

## RESULTS

***In silico* analyses of integrons. (i) Orientation of cassettes relative to replication.** The differences in ssDNA availability between the lagging and leading strands during DNA replication affect the formation of DNA secondary structures (32). When the bottom strand of an *attC* site (*attC<sub>bs</sub>*) is located on the lagging strand template in which large regions of ssDNA are available (i.e., between Okazaki fragments), its folding is favored, increasing the frequency of *attC*  $\times$  *attI* recombination (31). We previously observed that in all 10 analyzed sedentary chromosomally located integrons (SCIs), *attC* sites were oriented so that their bottom strands were located on the leading strand template, potentially limiting cassette rearrangements (31). Here, we broadened this analysis by comparing 30 SCIs with 36 mobile chromosomal integrons (MCIs) (Fig. 2; see Data Set S1 in the supplemental material). The latter corresponded to mobile integrons (carried on transposons) but located in chromosomes. We confirmed the previously observed trend: most SCIs (25/30) were oriented so that their *attC<sub>bs</sub>* were located on the leading strand template. In particular, this was always the case for *Vibrio* species. We did not observe such bias in orientation among MCIs.

We hypothesized that the orientation of the integron array relative to replication affects cassette dynamics by modifying their excision rate. Indeed, SCIs with *attC<sub>bs</sub>* on the leading strand template had larger arrays of cassettes than the others (Fig. 2). More

precisely, they carried up to 203 cassettes (median of 100), while arrays with  $attC_{bs}$  on the lagging strand template were always smaller (up to 22 cassettes, with a median of 15). In the data set of SCI-containing genomes, *Vibrio* genomes are overrepresented, reflecting the general bias toward human pathogen species. We therefore repeated this analysis after exclusion of *Vibrio* genomes and found that this difference remained significant (Fig. S1).

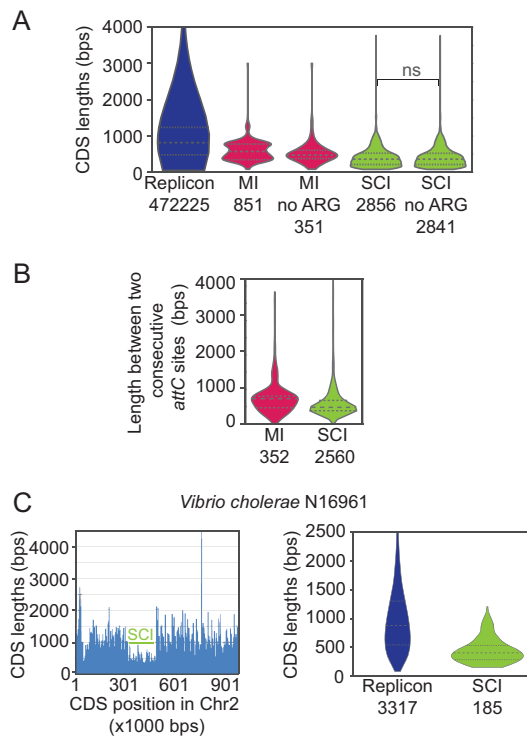
**(ii) Cassette lengths.** Cassette lengths were investigated as another means of control over their excision dynamics. When simultaneous folding of both flanking  $attC$  sites is promoted, e.g., when the distance between both sites was smaller than the length of ssDNA found between two Okazaki fragments or smaller than the size of supercoiled plectonemes, cassette excision is likely to be favored. Therefore, we decided to analyze the cassette length. We used two independent proxies for the length of cassettes to test this hypothesis: CDS length and the distance between identified  $attC$  sites. On the one hand, CDS length underestimates cassette length but is highly correlated with it, since most cassettes have one single CDS and small adjacent regions. On the other hand, the distance between  $attC$  sites provides an exact measure of cassette length but is affected by inaccuracies in the detection of  $attC$  sites (a missed  $attC$  can lead to the doubling of a cassette length). We performed these analyses for the 393 integrons identified by the IntegronFinder program (10) (Materials and Methods) and their respective replicons. CDSs in MIs and SCIs are significantly shorter than CDSs in replicons (median CDS lengths were 575, 362, and 818 bp, respectively [Fig. 3A]). Moreover, CDSs in SCIs are significantly shorter than those in MIs. We also performed this analysis by excluding known antibiotic resistance genes (ARGs) because they are overrepresented in MIs (500 out of 851 CDSs), upon which we observed a significant decrease in median MI CDS lengths. However, the non-ARG CDSs in MIs remain significantly longer than the non-ARG CDSs in SCIs (median CDS lengths were 473 and 362 bp, respectively).

We also measured the lengths of cassettes using the distance between two consecutive  $attC$  sites (Fig. 3B), which confirmed that cassettes of SCIs are smaller than those of MIs. We made three controls to validate these results. First, we excluded *Vibrio* genomes because they are overrepresented (Fig. S2A). Second, we controlled for interintegron and interreplicon variability by determining the mean values of CDS lengths per integron or per replicon and the mean lengths between two consecutive  $attC$  sites per integron (Fig. S2B and S2C). Third, we calculated pairwise differences in mean CDS lengths between a replicon and its associated integron to control for interspecies variability (Fig. S2D).

A clear example of this difference in lengths between CDSs of SCIs and replicons is the paradigmatic SCI of *V. cholerae*. CDSs in this SCI are significantly shorter than CDSs in the replicon (median CDS lengths are 405 and 882 bp, respectively [Fig. 3C]). We extended our analysis to SCIs present in other *Vibrio* species and observed similar trends (Fig. S3).

**(iii)  $attC$  site properties.** Cassette insertion frequency depends on the properties of  $attC$  sites involved in the reaction (17, 31). In order to be bound by the integrase,  $attC$  sites must adopt a recombinogenic structure, i.e., with paired R and L boxes (Fig. 1B). Based on DNA folding predictions, the probability of folding a recombinogenic structure can be calculated, which we call the pfold value (Materials and Methods). The presence of a large VTS can favor the formation of complex branched structures that do not reconstitute a recombinogenic  $attC$  site (31). Therefore, the length of  $attC$  sites could be an important parameter influencing their recombination. We compared the properties of  $attC$  sites in MIs and SCIs that are most likely to affect recombination levels: pfold, length of  $attC$  sites, and stability of the recombinogenic structure once folded ( $\Delta G$ ). These analyses were performed on 185  $attC$  sites from MIs and 1,744  $attC$  sites from SCIs (Fig. 4) (Materials and Methods).

We observed significant differences between MI and SCI  $attC$  site pfold values (Fig. 4A and S4A). The large majority of MI  $attC$  sites have a very high pfold value (71%



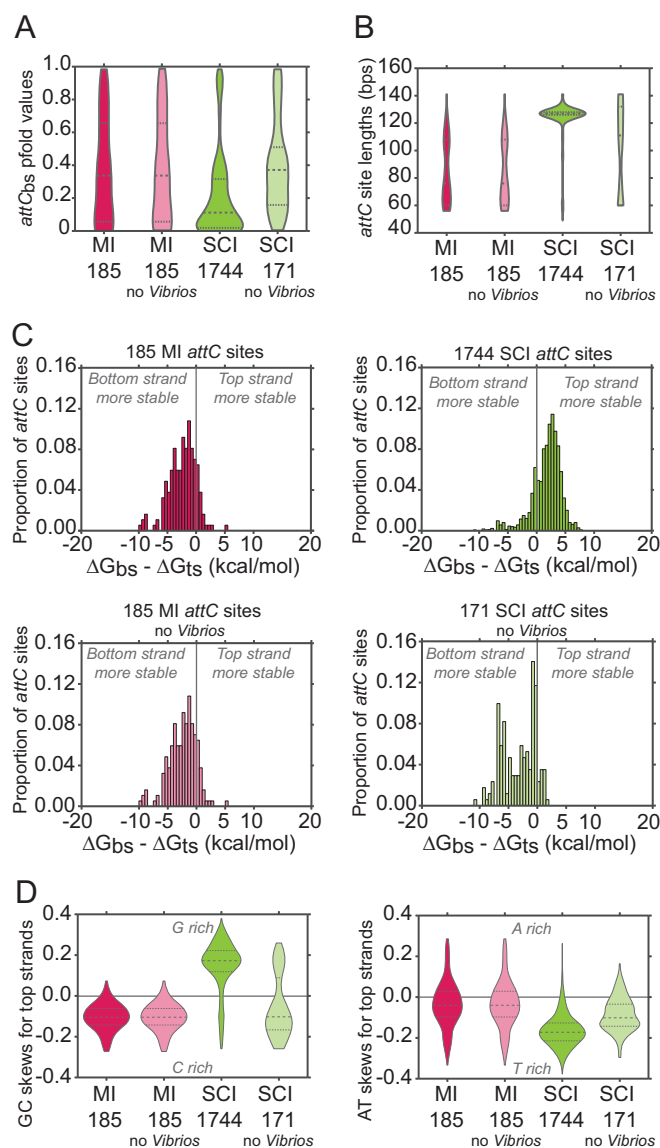
**FIG 3** Cassette and coding sequence (CDS) length analysis. The length of the CDS in replicons, mobile integrons (MI), and sedentary chromosomal integrons (SCI) with and without antibiotic resistance genes (ARG) is shown in base pairs (bps). The numbers below each violin diagram refer to the total number of CDSs or cassettes analyzed. (A) Violin plots showing the distribution of CDS lengths for replicons, MIs, and SCIs (excluding ARGs or not excluding ARGs). Tests of the differences between data sets were performed using the Wilcoxon rank sum test. Differences are significant ( $P$  values of  $<10^{-6}$ ), except between the two rightmost violin plots (differences between CDS lengths of all SCIs and SCIs excluding ARGs or not excluding ARG). ns, not significant. (B) Violin plots showing the distribution of lengths between two consecutive *attC* sites for MIs and SCIs. Tests of the differences between data sets were performed using the Wilcoxon rank sum test. Differences are significant ( $P$  values of  $<10^{-6}$ ). (C) CDS length analysis of the *Vibrio cholerae* N16961 strain. (Left) CDS lengths as a function of their positions in chromosome 2 (Chr2). The horizontal green bar indicates the position of the SCI. (Right) Violin plots showing the distribution of CDS lengths for the two replicons and the SCI. Tests of the differences between data sets were performed using the Wilcoxon rank sum test. Differences are significant ( $P$  values of  $<10^{-6}$ ).

sites with a pfold value of  $>0.1$ ), whereas SCIs contain fewer such sites (only 52%). Moreover, contrary to MIs, SCIs contain sites with an extremely low pfold value (6% sites with a pfold value between  $10^{-5}$  and  $10^{-7}$ ). These low-pfold *attC* sites are mostly found in *Vibrio* SCIs. When *attC* sites exclusively found in *Vibrio* strains were excluded from the data set, we found no significant difference among the pfold values of MI and SCI sites (Fig. 4A and S4A).

*attC* site length comparison between MIs and SCIs showed that the former have smaller VTSS, which can be as short as 3 nucleotides. The *attC* sites of SCIs often have longer VTSSs. The length of MI *attC* sites is relatively heterogeneous, ranging from 56 to 141 bp, with a majority of small *attC* sites ( $<100$  bp) (Fig. 4B and S4B). The *attC* sites of SCIs are more homogeneous, predominantly measuring between 120 and 129 bp (Fig. 4B and S4B). However, this size distribution is mostly due to *attC* sites of *Vibrio* spp. When the *Vibrio attC* sites are excluded, the length of *attC* sites is not significantly different between MIs and SCIs (Fig. 4B and S4B). We did not observe any correlation between the length and the pfold values of *attC* sites, even when excluding the *Vibrio attC* sites (Fig. S5A).

Our previous study of 263 MI *attC* sites from the INTEGRALL database (33) showed that the  $\Delta G$  of the folded bottom strands ( $\Delta G_{bs}$ ) is on average 2.12 kcal/mol lower than the  $\Delta G$  of the folded top strands ( $\Delta G_{ts}$ ), suggesting that folded bottom strands are more





**FIG 4** Analysis of 185 mobile integron (MI) and 1,744 sedentary chromosomal integron (SCI) *attC* site properties. *attC* sites were identified by IntegronFinder (Materials and Methods). The numbers below each violin diagram refer to the total number of *attC* sites analyzed. bs, bottom strand; ts, top strand; bps, base pairs. (A) Violin plots showing the distribution of *attC*<sub>bs</sub> pfold values for MIs and SCIs, excluding *Vibrio* strains or not excluding *Vibrio* strains. *attC*<sub>bs</sub> pfold values were calculated with RNAfold program from the ViennaRNA 2 package (Materials and Methods). Tests of the differences between data sets were performed using the Wilcoxon rank sum test. Differences are significant ( $P$  values of  $<10^{-6}$ ), except between MI *attC* sites excluding *Vibrio* strains or not excluding *Vibrio* strains, and between SCI *attC* sites excluding *Vibrio* strains and MI *attC* sites (excluding *Vibrio* strains or not excluding *Vibrio* strains). (B) Violin plots showing the distribution of *attC* site lengths for MIs and SCIs, excluding *Vibrio* strains or not excluding *Vibrio* strains. Tests of the differences between data sets were performed using the Wilcoxon rank sum test. Differences are significant ( $P$  values of  $<10^{-4}$ ) except between MI *attC* sites excluding *Vibrio* strains or not excluding *Vibrio* strains. (C) Proportion of *attC* sites as a function of the difference in  $\Delta G$  between bottom and top strands ( $\Delta G_{bs} - \Delta G_{ts}$ ) for MIs and SCIs, excluding *Vibrio* strains or not excluding *Vibrio* strains.  $\Delta G$  values (in kilocalories per mole) were calculated with RNAfold program from the ViennaRNA 2 package (Materials and Methods). (D) Violin plots showing the distribution of GC and AT skews calculated for the top strands of the *attC* sites for MIs and SCIs, excluding *Vibrio* strains or not excluding *Vibrio* strains. GC and AT skews were calculated as described in Materials and Methods. Negative skews correspond to an enrichment of purines (guanines [G] or adenines [A]) on the bottom strands. C, cytosine; T, thymine.

stable (17). The 185 MI *attC* sites from our genomic data set show similar differences (2.38 kcal/mol [Fig. 4C]). Surprisingly, the analysis of 1,744 SCI *attC* sites shows that  $\Delta G_{\text{bs}}$  is on average 1.71 kcal/mol higher than the  $\Delta G_{\text{ts}}$ , suggesting that folded top strands are more stable. Once again, this effect was due to the *attC* sites of *Vibrio* spp.: their exclusion reversed the trend toward higher  $\Delta G_{\text{ts}}$  (differences of 3.13 kcal/mol). This value was not significantly different from the one observed for MI *attC* sites (Fig. 4C). We observed a negative correlation between the length of *attC* sites and the  $\Delta G_{\text{bs}} - \Delta G_{\text{ts}}$  in MIs and in SCIs without *Vibrio attC* sites (Fig. S5B). This correlation was reversed for SCI *attC* sites, which is expected, given that their increased length is mostly due to a longer VTS, which together with the UCS produces this difference in  $\Delta G$  (Fig. S5B).

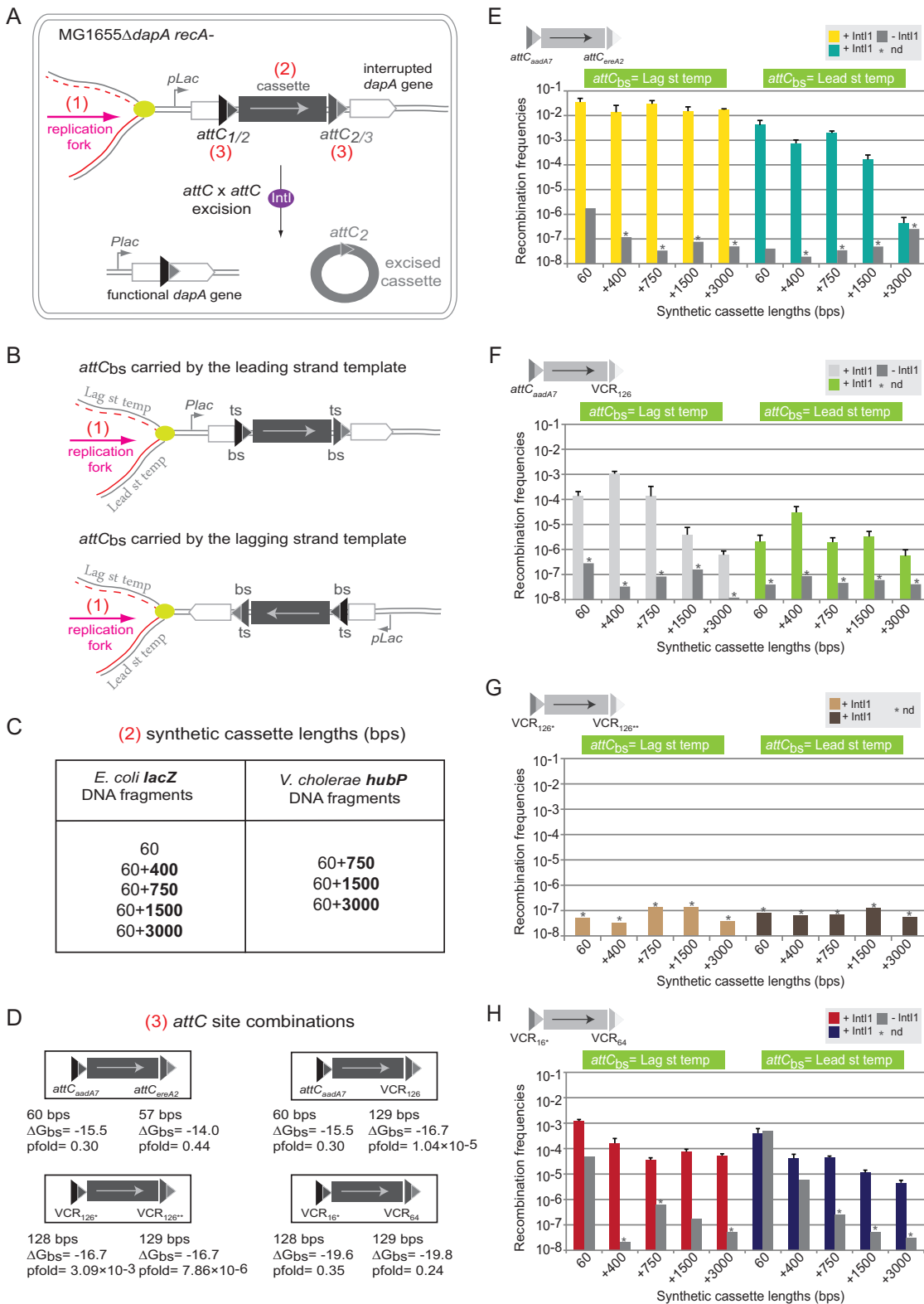
As previously described, the bottom strands of MI *attC* sites are enriched in purines, especially in guanines, which contribute to the difference in folded strand stability (17). Indeed, purines have a higher self-stacking tendency, thus stabilizing secondary structures (34). We confirmed negative GC and AT skews for MI *attC* sites from our data set (the skews were calculated relative to the top strands) (Fig. 4D). However, for SCI *attC* sites, we observed positive GC and negative AT skews, meaning that the bottom strands were C and A rich. This difference in nucleotide skews of *attC* sites in MIs and SCIs could explain, at least in part, the difference in folded strand stability between bottom and top strands. When *attC* sites from *Vibrio* spp. were excluded from the analysis, the GC skew of the remaining SCI *attC* sites became negative as in MI sites, even though there was a small subpopulation of SCI *attC* sites with C-rich bottom strands (Fig. 4D). Additionally, these remaining SCI *attC* sites showed a more homogeneous negative AT skew, resembling that of MI sites.

**In vivo analysis of cassette excision. (i) Cassette excision assay.** In order to better understand the biological significance of our *in silico* analyses, we performed *in vivo* excision tests for several synthetic cassettes using the previously described excision assay (23) (Fig. 5A). In this assay, excision of cassettes between *attC* sites leads to reconstitution of the essential *dapA* gene, allowing recombinants to grow on media lacking 2,6-diaminopimelic acid (DAP) (the reticulating agent of peptidoglycan in *Escherichia coli*). Comparison of the number of clones growing with and without DAP yields a recombination frequency for a given reaction. Corresponding strains without integrase were used as controls to assess the rate of false-positive events potentially due to replication slippage.

We tested the influence of three parameters on cassette excision (Fig. 5A): (i) orientation of cassettes relative to replication, (ii) cassette lengths, and (iii) different *attC* site combinations. To study the effect of cassette orientation, we inserted synthetic cassettes into the  $\lambda$  *attB* site of the MG1655 $\Delta$ *dapA* strain in both orientations, so that the bottom strands of both *attC* sites were on either the leading or lagging strand template (Fig. 5B). For the leading strand template orientation, the only possibility for *attC* sites to fold was by extrusion from dsDNA. For the lagging strand template orientation, *attC* sites could fold either by extrusion from dsDNA or directly from ssDNA generated from discontinuous replication. We varied the cassette length by introducing DNA fragments of various lengths (from 60 to 3,060 bp) between two *attC* sites (*E. coli lacZ* DNA fragments) (Fig. 5C). Finally, we also used four combinations of *attC* sites: *attC*<sub>aadA7</sub>-*attC*<sub>ereA2</sub>, *attC*<sub>aadA7</sub>-VCR<sub>126</sub>, VCR<sub>126</sub>\*-VCR<sub>126</sub>\*\*<sub>r</sub>, and VCR<sub>16</sub>\*-VCR<sub>64</sub> (VCR names explained in the legend to Fig. 5D) (Fig. 5D and S6). While *attC*<sub>aadA7</sub> and *attC*<sub>ereA2</sub> are found in MIs, the VCRs correspond to *attC* sites from the *V. cholerae* SCI.

We performed two additional controls on *attC*<sub>aadA7</sub>-*attC*<sub>ereA2</sub> and *attC*<sub>aadA7</sub>-VCR<sub>126</sub> cassettes. First, we tested that the expression of *dapA* from the *P*<sub>lac</sub> promoter (Fig. 5A) did not interfere with recombination by performing excision reactions without isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG). Second, we tested for sequence-specific effects by replacing selected *lacZ* DNA fragments by *hubP* fragments from *V. cholerae* (Fig. 5C). In all these controls, we found no significant difference from our results (Fig. S7).

Cassette Excision in Integrations



**FIG 5** Cassette excision assay. (A) Experimental set-up of the cassette excision assay. The *dapA* gene (white rectangle) is interrupted by a synthetic cassette containing a DNA fragment (gray rectangle) flanked by two *attC* sites (triangles). Recombination mediated by the IntI integrase (purple oval) leads to the excision of the cassette (excised cassette [gray circle]) and restores a functional *dapA* gene. Three parameters are varied: (1) orientation of cassettes relative to replication; (2) cassette lengths; and (3) different *attC* site combinations. (B) Orientation of cassettes relative to replication. The orientation of the bottom strands (bs) of *attC* sites relative to replication are shown.

(Continued on next page)

**(ii) Effects of regulatory network on cassette excision.** (a)  $attC_{aadA7}$ - $attC_{ereA2}$  synthetic cassettes. The first set of synthetic cassettes was flanked by  $attC_{aadA7}$  and  $attC_{ereA2}$  MI sites (Fig. 5D). The pfold value of both sites is  $>0.1$ , implying that their most stable structures are recombinogenic (Fig. 5D and S6). Recombination occurred at high frequency for all cassettes when  $attC_{bs}$  were carried on the lagging strand template and for all but the largest cassette when  $attC_{bs}$  were carried on the leading strand template (Fig. 5E). This suggests that both tested  $attC$  sites could be efficiently and simultaneously extruded from dsDNA (Fig. 5E). In the absence of integrase, we observed excision events only for small cassettes (60-bp-long cassette) in both orientations, probably due to replication slippage (Fig. S8) (30).

(b)  $attC_{aadA7}$ - $VCR_{126}$  synthetic cassettes. In order to assess the importance of the  $attC$  site pfold value on cassette excision, we tested the excision of cassettes flanked on one side by the previously used  $attC_{aadA7}$  site and on the other side by the  $VCR_{126}$  site.  $VCR_{126}$  has a very low pfold value (pfold value of  $1.04 \times 10^{-5}$ ) that is significantly lower than that of  $attC_{ereA2}$  (Fig. 5D and S6), and it is also unlikely to fold into a recombinogenic structure from dsDNA (31). Accordingly, we observed significantly lower recombination rates for this cassette independently of cassette orientation and fragment type (Fig. 5F and S7D). As for the previous set of cassettes, when  $attC_{bs}$  were carried on the lagging strand template, the frequency of recombination was higher than in the inverse orientation. However, for this set of cassettes, recombination also depended on cassette length (Fig. 5F). Due to the low propensity of the VCR site to extrude from dsDNA, we observed a relatively constant low frequency of cassette excision when the bottom strands of  $attC$  sites were carried on the leading strand template.

(c) VCR-VCR synthetic cassettes. We also tested two sets of cassettes flanked on both sides by VCR sites. First, we combined  $VCR_{126^+}$  and  $VCR_{126^{**}}$  sites. The most stable structures of these sites are non-recombinogenic (pfold values of  $3.09 \times 10^{-3}$  and  $7.86 \times 10^{-6}$  [Fig. 5D and S6]), and we expected that the low pfold values would not allow simultaneous folding and cassette excision from the ds pathway or even the ss pathway. Indeed, we did not detect excision of these cassettes (Fig. 5G). Second, we combined  $VCR_{16^+}$  and  $VCR_{64}$  sites for which the most stable structures are recombinogenic (pfold values of 0.35 and 0.24 [Fig. 5D and S6]). In this case, we observed high rates of recombination from the ds and/or ss pathway (Fig. 5H), presumably because both VCR sites could fold efficiently. We observed higher recombination rates when  $attC_{bs}$  were carried on the lagging strand template. The frequency of recombination, similarly to  $attC_{aadA7}$ - $VCR_{126}$  cassettes, depended on cassette length when  $attC_{bs}$  were carried on the lagging strand template (from  $1.17 \times 10^{-3}$  for 60 bp to  $5.25 \times 10^{-5}$  for +3,000-bp cassette lengths), but this effect was less pronounced for  $VCR_{16^+}$ - $VCR_{64}$  cassettes (Fig. 5F and H). As for other sites, in the absence of integrase, we observed recombination events only for small cassettes (60- and +400-bp cassette lengths) in both orientations. These events were likely due to replication slippage favored by the high sequence identity between the two sites (83% [Fig. S8]).

### FIG 5 Legend (Continued)

Lag st temp, lagging strand template; Lead st temp, leading strand template. (C) Synthetic cassette lengths. Lengths in base pairs (bps) of the supplementary *E. coli lacZ* and *V. cholerae hubP* DNA fragments introduced in synthetic cassettes are indicated in bold numbers. (D)  $attC$  site combinations. The four  $attC$  site combinations used are shown.  $\Delta G_{bs}$  (in kilocalories per mole) and  $attC_{bs}$  pfold were calculated with RNAfold program from the ViennaRNA 2 package (Materials and Methods, and Fig. S6).  $\Delta G_{bs}$  is calculated for the most stable structure with constraints (recombinogenic) (Fig. S6). The numbering of each VCR indicates its position in the *V. cholerae* N16961 SCI. The single and double asterisks indicate that the wild-type  $attC$  site has been slightly modified (by removing 1 nucleotide [nt] [single asterisk] or changing 1 nt [double asterisk]) in order to generate a functional *dapA* fusion after recombination events. bs, bottom strand. (E to H) Recombination frequencies for  $attC_{aadA7}$ - $attC_{ereA2}$  (E),  $attC_{aadA7}$ - $VCR_{126}$  (F),  $VCR_{126^+}$ - $VCR_{126^{**}}$  (G), and  $VCR_{16^+}$ - $VCR_{64}$  (H) synthetic cassettes. Synthetic cassette lengths and orientation of  $attC_{bs}$  relative to replication are indicated. Experiments were performed in the presence (+) or absence (-) of Int1 integrase. bs, bottom strand; Lag st temp, lagging strand template; Lead st temp, leading strand template; bps, base pairs; nd, not detected.

## DISCUSSION

Our study aimed to understand the rules that govern cassette array dynamics in SCIs and MIs and to determine the cause of shorter CDS lengths in cassettes compared to the rest of the genome.

**Replication controls cassette dynamics in integrations.** The *in silico* analyses reveal that  $attC_{bs}$  in SCIs are predominantly carried on the leading strand template and that these SCI arrays are significantly larger, up to 200 cassettes in SCIs of *Vibrio* spp. Our *in vivo* results show that such orientation relative to replication limits cassette excision and thus stabilizes large cassette arrays. This may explain why this orientation was much less frequent among MCIs, for which a higher cassette mobility might be favored over the preservation of a larger reservoir of genetic functions that will stay accessible through horizontal gene transfer. Interestingly, the five SCIs identified in inverse orientation were found exclusively in *Xanthomonas* species and their number of cassettes did not exceed 22. Comparison of SCIs in these two closely related genera is particularly interesting, since according to phylogenetic analyses, the acquisition of integron systems in both *Vibrio* and *Xanthomonas* occurred independently and thus can be regarded as two single ancestral events (3, 35). Contrary to the *Vibrio* SCIs, the *Xanthomonas* SCIs are subjected to genetic erosion, and it is tempting to speculate that this is a consequence of their orientation and that the frequent inactivation of their integrases was selected to freeze cassette excisions (36).

**Cassette length constraints in integrations.** Our analyses revealed that CDSs carried on SCI and MI cassettes were on average shorter than the remaining replicon's CDSs. In order to clarify the origin(s) of this characteristic, in particular whether the reduced length of CDSs in cassettes reflected a constraint in maximal distance between two consecutive  $attC$  sites for efficient recombination, we tested the effect of cassette length on their excision *in vivo*. We would then expect to see a significant drop in recombination frequency for cassettes larger than most SCI and MI cassettes and smaller than most replicon CDSs (that is, between 500 and 800 bp), at least in one orientation. However, this was not the case (Fig. 5). When bottom strands were carried on the lagging strand template, even though cassette length can have an impact on cassette excision frequency, we observed a consistent drop in recombination frequencies only for very large cassettes ( $\geq 1,500$  bp), when at least one of the adjacent sites has a low pfold value (Fig. 5F). Also, we did not observe any consistent drop in recombination frequencies for cassettes up to 1,500 bp when  $attC_{bs}$  were carried on the leading strand template (Fig. 5). Taken together, these observations indicate that the reduced length of CDSs in cassettes is not due to a recombination-related limitation. It is possible that the prevalence of small CDSs in cassettes reflects constraints during cassette genesis. The process of *de novo* cassette formation is largely unknown, and the proposed hypotheses have many incongruities (discussed in reference 7). Thus, we can only speculate on the underlying processes and related constraints. However, the limitations in maximal cassette length are not stringent, since long cassettes such as ARGs encoding class D  $\beta$ -lactamases are found among MIs (37). Long cassettes might be less likely to be created, and their presence could reflect their strong selection and confer important selective advantages.

**$attC$  site properties control cassette dynamics in integrations.** By using different  $attC$  sites in our synthetic cassettes, we tested the impact of their properties on cassette excision. When cassettes were flanked by two  $attC$  sites with high pfold values, cassette length did not influence the recombination rate when the bottom strands were carried on the lagging strand templates. Because of their high pfold values, simultaneous folding of the two  $attC$  sites could occur in three possible ways. (i) Both sites fold from ssDNA during the passage of the replication fork (the "ss pathway"). (ii) Both sites are extruded from dsDNA (the "ds pathway"). (iii) One site is folded from ssDNA, and the other is extruded from dsDNA. The third pathway could explain the very high efficiency of recombination that we obtained for +3,000-bp cassettes when bottom strands were carried on the lagging strand template. On the other hand, the difficulties of recom-

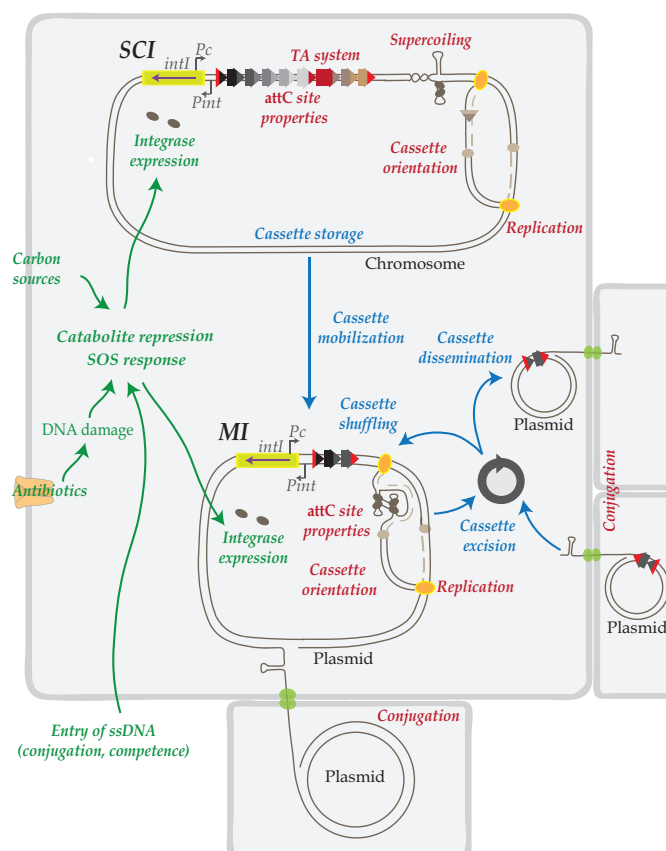
binning +3,000-bp cassettes when both high-pfold  $attC_{bs}$  were carried on the leading strand template, and therefore must have been extruded from dsDNA, might be explained by topological constraints such as the presence of independent topological domains in bacterial chromosomes (38). Another hypothesis is that cruciform extrusion induces DNA structural transitions, restricting the slithering of the molecule and reducing the possibility of distant sites to come into contact (39).

When cassettes are flanked by at least one  $attC$  site with a low pfold value, we observed a decrease in recombination efficiency. Moreover, when cassettes are flanked on both sides by low-pfold  $attC$  sites, their excision frequency is even further decreased. In addition, the excision rate of cassettes flanked by high-pfold VCRs is decreased compared to that of cassettes flanked by high-pfold MI  $attC$  sites. These differences could be due either to the influence of other folding-related VCR properties or to host factor binding. Once folded, large VCR sites could be efficiently targeted by hairpin or cruciform-binding proteins (29).

Interestingly, we also observed an effect of cassette length on the excision frequency of cassettes flanked by at least one low-pfold  $attC$  site and oriented with the bottom strand carried on the lagging strand template. Under these conditions, cassette length correlates with cassette excision frequency, possibly because of a higher chance for two consecutive  $attC_{bs}$  sites to be located within the ss region at the replication fork (between Okazaki fragments [40]). This effect has previously been observed for the IS608 insertion sequence, which also requires ssDNA substrates to recombine (41).

These results show that  $attC$  site pfold, and more generally  $attC$  site biophysical properties, control cassette excision dynamics. Moreover, the *in silico* analyses demonstrated that MIs mostly contain small  $attC$  sites with high pfold values and folded bottom strands which are more stable than folded top strands. This ultimately favors their recombination by the integrase (17). This is also true for many SCIs, but surprisingly, not for *Vibrio* SCIs. The reason for such discrepancy is unknown, but the genomic architecture of vibrios, with their two-chromosome replication being highly regulated and coordinated and their unique physical organization (42, 43) might be at the origin of several specific traits.

**Cassette dynamics: evolutionary considerations and trade-off.** In these studies, we show that cassette excision is highly regulated by the cell replication process and the properties of cassette recombination sites. We demonstrate that differential dynamics of cassette excision are ensured by integron properties that shape a trade-off between evolvability and genetic capacitance. In MIs, efficient cassette recombination is favored and timed to conditions when generating diversity upon which selection can act ensures a rapid response to environmental stresses. In contrast, in SCIs, cassette dynamics favor the maintenance of large cassette arrays and vertical transmission. Interestingly, even in large SCI arrays, there are very few pseudogenes among cassette CDs. Several studies of *V. cholerae* have shown that the SCI array was the most variable locus among isolates (44, 45). This suggests that on a global time scale, cassettes must be regularly rearranged and tested for selective advantage, thus explaining the preservation of cassette gene functionality, even when promoterless. We therefore confirmed the role of SCIs as a reservoir of adaptive functions. Indeed, the evolutionary history of integrons suggests that SCIs could constitute a cassette reservoir and that subsequent harvesting of cassettes from various SCI sources leads to contemporary MIs (35). SCI  $attC$  sites display a strikingly high degree of sequence relatedness (around 80% identity), unlike their MI counterparts (3, 10). This suggests a link between the host and the sequences of  $attC$  recombination sites, and more precisely, it suggests that the formation of *de novo* integron cassettes most likely occurs in SCI hosts. Moreover, SCI cassettes can become substrates of MI integrases and therefore be directly recruited into MIs as demonstrated for class 1 integrons (3, 20, 46). Thus, the most recombinogenic cassettes in SCIs would be more likely mobilized in MIs and further selected because of their higher capacity to disseminate the associated adaptive functions. The ensemble of these regulation processes can have a direct effect on integrase stability.



**FIG 6** Regulatory network in integrations. Representation of cassette dynamics, namely, cassette storage in SCIs, cassette mobilization from SCIs to MIs, and cassette excision, shuffling, and dissemination in MIs is shown in blue. Representation of the regulatory network is shown in red. Toxin-antitoxin (TA) systems stabilize cassette arrays in SCIs and supercoiling, replication (lagging strand template), and conjugation favor *attC* site folding and cassette dynamics. Representation of the connections between integrations and bacterial physiology is shown in green. Conjugation, competence, and antibiotics induce the SOS response and integrase expression, and carbon sources initiate catabolite repression and integrase expression.

A mathematical model has suggested that, while integrases in MIs are selectively maintained by the antibiotic pressure, integrases in SCIs are maintained because they enable their hosts to use cassette arrays efficiently as a reservoir of standing genetic variability (47).

Taken together, these results extend the list of processes intimately connecting the integron system with its host cell physiology (Fig. 6). This complex and extensive network of regulation processes constitutes a powerful and daunting system, making it increasingly difficult to limit the spread of multidrug resistance among bacteria.

## MATERIALS AND METHODS

**In silico analysis. (i) Data.** The sequences and annotations of complete genomes were downloaded from NCBI RefSeq (last accessed in November 2013, <http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/>). Using the IntegronFinder program ([https://github.com/gem-pasteur/Integron\\_Finder](https://github.com/gem-pasteur/Integron_Finder)), we analyzed 2,484 bacterial genomes, including 2,626 replicons labeled as chromosomes and 2,006 replicons labeled as plasmids. IntegronFinder ensures an automatic and accurate identification of integrations, cassette arrays, and *attC* sites.

**(ii) SCI and MI classification.** Several criteria were used to determine the integron classification (10). Briefly, integrations were considered SCIs when they were present in the genomes of all the available sequenced strains of the species or when they contained more than 19 *attC* sites. They were considered MIs when they were absent in more than 40% of the sequenced genomes of the species, when they were present on a plasmid, or when the integrase was from one of the five classes of MIs.

**(iii) CDS analysis in sequenced strains containing integrations.** The set of MIs had 851 CDSs, whereas the set of SCIs had 2,856 CDSs, together belonging to 393 integrations. The replicons containing integrations

had 472,225 CDSs. Leading and lagging strands were determined using the information from the OriC prediction database (48). The leading strand was defined as the strand with an increasing gradient of GC disparity (the GC disparity is a measure of the Z-curve representing the excess of G over C, a similar measure to the GC skew). The complementary strand was defined as the lagging strand.

**(iv) CDS analysis in *Vibrio* strains.** Annotated genomes of the indicated *Vibrio* strains were downloaded from NCBI RefSeq. CDS lengths were those from the RefSeq annotations. NCBI reference sequences of the genomes and position and length of integrases are presented in Table S1A in the supplemental material.

**(v) attC sites used for the analysis.** For MI and SCI attC site comparison, we used the attC sites published and classified in reference 10 (185 different MI attC sites and 1,744 different sedentary CI attC sites). We reran IntegronFinder with the clustering parameters of 15 kb (-dt 15000 instead of 4 kb by default) for six known SCIs that IntegronFinder could not properly aggregate because of this threshold. The sequences of attC sites included the full 7-bp-long R box (the variable 4 bp of the R' sequence not necessarily being complementary to their counterparts in the R'' sequence, as is the case for attC sites in the integron array).

**(vi) attC site folding, ΔG and pfold predictions, and skews.** All folding predictions were obtained by RNAfold program from the ViennaRNA 2 package (49) with the set of DNA folding parameters derived from reference 50. We used the -p option to compute the partition functions and the -c option to add constraints required for a recombinogenic structure: pairing of the L' and L'' sequences and pairing of the 4 bases 5'-YAAC-3' in the R' sequence with the 4 bases 5'-GTTR-3' in the R'' sequence. We report the structures and ΔG values of the minimal free energy (MFE) structure. In order to obtain the pfold (probability of folding a recombinogenic structure) values, folding predictions were performed with and without constraints. pfold was calculated as the Boltzmann probability of the

constrained (recombinogenic) structure in the ensemble:  $P_{\text{fold}} = e^{\frac{E_u - E_c}{RT}}$ , where  $E_u$  is the Gibbs free energy of the unconstrained (total) ensemble,  $E_c$  is the Gibbs free energy of the constrained (recombinogenic) ensemble,  $R$  is the gas constant, and  $T$  is the absolute temperature (51). GC skew measures the abundance of guanines (G) compared to cytosines (C) on the top strand:  $(G - C)/(G + C)$ ; AT skew measures the abundance of adenines (A) compared to thymines (T) on the top strand:  $(A - T)/(A + T)$ .

**In vivo studies. (i) Bacterial strains and media.** Bacterial strains used in this study are described in Table S1B.

**(ii) Plasmids and primers.** Plasmids, primers, and synthetic fragments used in this study are described in Tables S1C and S1D.

**(iii) Integron cassette excision assay.** The pBAD::intI1 plasmid (p3938) was introduced by transformation into the MG1655ΔdapA derivative strains containing insertions (in the attB lambda site) of plasmids carrying a dapA gene interrupted by the synthetic cassettes (Tables S1C and S1D). These strains are unable to synthesize 2,6-diaminopimelic acid (DAP), and as a result, they are not viable without DAP supplement in the medium. Recombination between attC sites causes excision of the synthetic cassette, restoring a functional dapA gene and allowing the strain to grow on DAP-free medium. After overnight growth in the presence of appropriate antibiotics (spectinomycin [Sp], carbenicillin [Carb]), DAP, and glucose, strains were cultivated for 6 h in the presence of the appropriate antibiotic (Carb), DAP, L-arabinose (Ara), and IPTG to allow intI1 expression (Pbad) and dapA expression (P<sub>lac</sub> promoter), respectively. Then, cultures were plated on agar containing either LB with IPTG and Sp or LB with DAP, IPTG, and Sp. Recombination activity was calculated as the ratio of the number of cells growing in the absence of DAP over the total number of cells. For each reaction, we confirmed cassette excision by performing PCRs with SW23begin/DapA-R primers generating products of 715, 763, 784, and 832 bp for attC<sub>aadA7</sub>-attC<sub>ereA2</sub>, attC<sub>aadA7</sub>-VCR<sub>126</sub>, VCR<sub>126</sub>-VCR<sub>126\*\*</sub>, and VCR<sub>16</sub>-VCR<sub>64</sub> cassettes, respectively.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mBio.02296-16>.

**FIG S1**, EPS file, 0.5 MB.

**FIG S2**, EPS file, 1 MB.

**FIG S3**, PDF file, 0.6 MB.

**FIG S4**, EPS file, 1.3 MB.

**FIG S5**, JPG file, 1 MB.

**FIG S6**, PDF file, 0.2 MB.

**FIG S7**, EPS file, 1.1 MB.

**FIG S8**, EPS file, 1 MB.

**TABLE S1**, PDF file, 0.16 MB.

**DATA SET S1**, XLSX file, 0.1 MB.

## ACKNOWLEDGMENTS

We thank Claire Vit for helpful discussions and Sebastian Aguilar Pierlé for reading the manuscript.

This work was supported by the Institut Pasteur, the Centre National de la Recherche



Scientifique, Fondation pour la Recherche Médicale (project DBF20160635736), the French Government Investissement d'Avenir program Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-IBEID), the French National Research Agency (ANR-12-BLAN-DynamiNT), the European Union Seventh Framework Program (FP7-HEALTH-2011-single-stage), the "Evolution and Transfer Of Antibiotic Resistance" (EvoTAR); Paris Descartes University—Sorbonne Paris Cité, Fondation pour la Recherche Médicale (FDT20150532465) and École Doctorale Frontières du Vivant (FdV)—Programme Bettencourt (to A.N.); Marie Curie Intra-European Fellowship for Career Development (FP-7-PEOPLE-2011-IEF, ICADIGE) (to J.A.E.); and European Research Council grant (EVOMOBILOME grant 281605) (to E.P.C.R. and J.C.).

## REFERENCES

- Stokes HW, Hall RM. 1989. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Mol Microbiol* 3:1669–1683. <https://doi.org/10.1111/j.1365-2958.1989.tb00153.x>.
- Fluit AC, Schmitz FJ. 2004. Resistance integrons and super-integrons. *Clin Microbiol Infect* 10:272–288. <https://doi.org/10.1111/j.1198-743X.2004.00858.x>.
- Mazel D. 2006. Integrons: agents of bacterial evolution. *Nat Rev Microbiol* 4:608–620. <https://doi.org/10.1038/nrmicro1462>.
- Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Baptiste E, Lopez P, Tarr CL, Polz MF. 2011. Local mobile gene pools rapidly cross species boundaries to create endemicity within global *Vibrio cholerae* populations. *mBio* 2:e00335-10. <https://doi.org/10.1128/mBio.00335-10>.
- Cambray G, Guerout AM, Mazel D. 2010. Integrons. *Annu Rev Genet* 44:141–166. <https://doi.org/10.1146/annurev-genet-102209-163504>.
- Rapa RA, Labbate M. 2013. The function of integron-associated gene cassettes in *Vibrio* species: the tip of the iceberg. *Front Microbiol* 4:385. <https://doi.org/10.3389/fmicb.2013.00385>.
- Escudero JA, Loot C, Nivina A, Mazel D. 2015. The integron: adaptation on demand. *Microbiol Spectr* 3:MDNA3-0019-2014. <https://doi.org/10.1128/microbiolspec.MDNA3-0019-2014>.
- Partridge SR, Tsafnat G, Coiera E, Iredell JR. 2009. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 33:757–784. <https://doi.org/10.1111/j.1574-6976.2009.00175.x>.
- Naas T, Mikami Y, Imai T, Poirel L, Nordmann P. 2001. Characterization of In53, a class 1 plasmid- and composite transposon-located integron of *Escherichia coli* which carries an unusual array of gene cassettes. *J Bacteriol* 183:235–249. <https://doi.org/10.1128/JB.183.1.235-249.2001>.
- Cury J, Jové T, Touchon M, Néron B, Rocha EP. 2016. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* 44:4539–4550. <https://doi.org/10.1093/nar/gkw319>.
- Recchia GD, Hall RM. 1995. Gene cassettes: a new class of mobile element. *Microbiology* 141:3015–3027. <https://doi.org/10.1099/13500872-141-12-3015>.
- Francia MV, Zabala JC, de la Cruz F, Garcia-Lobo JM. 1999. The Int1 integron integrase preferentially binds single-stranded DNA of the *attC* site. *J Bacteriol* 181:6844–6849.
- Johansson C, Kamali-Moghaddam M, Sundström L. 2004. Integron integrase binds to bulged hairpin DNA. *Nucleic Acids Res* 32:4033–4043. <https://doi.org/10.1093/nar/gkh730>.
- Bouvier M, Demarre G, Mazel D. 2005. Integron cassette insertion: a recombination process involving a folded single strand substrate. *EMBO J* 24:4356–4367. <https://doi.org/10.1038/sj.emboj.7600898>.
- Escudero JA, Loot C, Parissi V, Nivina A, Bouchier C, Mazel D. 2016. Unmasking the ancestral activity of integron integrases reveals a smooth evolutionary transition during functional innovation. *Nat Commun* 7:10937. <https://doi.org/10.1038/ncomms10937>.
- Bouvier M, Ducos-Galand M, Loot C, Bikard D, Mazel D. 2009. Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genet* 5:e1000632. <https://doi.org/10.1371/journal.pgen.1000632>.
- Nivina A, Escudero JA, Vit C, Mazel D, Loot C. 2016. Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of *attC* recombination sites. *Nucleic Acids Res* 44:7792–7803. <https://doi.org/10.1093/nar/gkw646>.
- Frumerie C, Ducos-Galand M, Gopaul DN, Mazel D. 2010. The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res* 38:559–569. <https://doi.org/10.1093/nar/gkp990>.
- Stokes HW, O'Gorman DB, Recchia GD, Parsekian M, Hall RM. 1997. Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* 26:731–745. <https://doi.org/10.1046/j.1365-2958.1997.6091980.x>.
- Mazel D, Dychinco B, Webb VA, Davies J. 1998. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* 280:605–608. <https://doi.org/10.1126/science.280.5363.605>.
- Loot C, Ducos-Galand M, Escudero JA, Bouvier M, Mazel D. 2012. Replicative resolution of integron cassette insertion. *Nucleic Acids Res* 40:8361–8370. <https://doi.org/10.1093/nar/gks620>.
- Collis CM, Hall RM. 1995. Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* 39:155–162. <https://doi.org/10.1128/AAC.39.1.155>.
- Baharoglu Z, Bikard D, Mazel D. 2010. Conjugative DNA transfer induces the bacterial SOS response and promotes antibiotic resistance development through integron activation. *PLoS Genet* 6:e1001165. <https://doi.org/10.1371/journal.pgen.1001165>.
- Hocquet D, Llanes C, Thouvez M, Kulasekara HD, Bertrand X, Plésiat P, Mazel D, Miller SI. 2012. Evidence for induction of integron-based antibiotic resistance by the SOS response in a clinical setting. *PLoS Pathog* 8:e1002778. <https://doi.org/10.1371/journal.ppat.1002778>.
- Barraud O, Ploy MC. 2015. Diversity of class 1 integron gene cassette rearrangements selected under antibiotic pressure. *J Bacteriol* 197:2171–2178. <https://doi.org/10.1128/JB.02455-14>.
- Guerin E, Cambray G, Sanchez-Alberola N, Campoy S, Erill I, Da Re S, Gonzalez-Zorn B, Barbé J, Ploy MC, Mazel D. 2009. The SOS response controls integron recombination. *Science* 324:1034. <https://doi.org/10.1126/science.1172914>.
- Cambray G, Sanchez-Alberola N, Campoy S, Guerin E, Da Re S, González-Zorn B, Ploy MC, Barbé J, Mazel D, Erill I. 2011. Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mob DNA* 2:6. <https://doi.org/10.1186/1759-8753-2-6>.
- Baharoglu Z, Krin E, Mazel D. 2012. Connecting environment and genome plasticity in the characterization of transformation-induced SOS regulation and carbon catabolite control of the *Vibrio cholerae* integron integrase. *J Bacteriol* 194:1659–1667. <https://doi.org/10.1128/JB.05982-11>.
- Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* 74:570–588. <https://doi.org/10.1128/MMBR.00026-10>.
- Loot C, Parissi V, Escudero JA, Amarir-Bouhram J, Bikard D, Mazel D. 2014. The integron integrase efficiently prevents the melting effect of *Escherichia coli* single-stranded DNA-binding protein on folded *attC* sites. *J Bacteriol* 196:762–771. <https://doi.org/10.1128/JB.01109-13>.
- Loot C, Bikard D, Rachlin A, Mazel D. 2010. Cellular pathways controlling integron cassette site folding. *EMBO J* 29:2623–2634. <https://doi.org/10.1038/emboj.2010.151>.
- Wolfson J, Dressler D. 1972. Regions of single-stranded DNA in the growing points of replicating bacteriophage T7 chromosomes. *Proc Natl Acad Sci U S A* 69:2682–2686. <https://doi.org/10.1073/pnas.69.9.2682>.
- Moura A, Soares M, Pereira C, Leitão N, Henriques I, Correia A. 2009. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* 25:1096–1098. <https://doi.org/10.1093/bioinformatics/btp105>.

34. Sigel A, Operschall BP, Sigel H. 2014. Comparison of the pi-stacking properties of purine versus pyrimidine residues. Some generalizations regarding selectivity. *J Biol Inorg Chem* 19:691–703. <https://doi.org/10.1007/s00775-013-1082-5>.
35. Rowe-Magnus DA, Guerout AM, Ploncard P, Dychinco B, Davies J, Mazel D. 2001. The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrations. *Proc Natl Acad Sci U S A* 98:652–657. <https://doi.org/10.1073/pnas.98.2.652>.
36. Gillings MR, Holley MP, Stokes HW, Holmes AJ. 2005. Integrations in *Xanthomonas*: a source of species genome diversity. *Proc Natl Acad Sci U S A* 102:4419–4424. <https://doi.org/10.1073/pnas.0406620102>.
37. Poirel L, Naas T, Nordmann P. 2010. Diversity, epidemiology, and genetics of class D beta-lactamases. *Antimicrob Agents Chemother* 54:24–38. <https://doi.org/10.1128/AAC.01512-08>.
38. Postow L, Hardy CD, Arsuaga J, Cozzarelli NR. 2004. Topological domain structure of the *Escherichia coli* chromosome. *Genes Dev* 18:1766–1779. <https://doi.org/10.1101/gad.1207504>.
39. Shlyakhtenko LS, Hsieh P, Grigoriev M, Potaman VN, Sinden RR, Lyubchenko YL. 2000. A cruciform structural transition provides a molecular switch for chromosome structure and dynamics. *J Mol Biol* 296:1169–1173. <https://doi.org/10.1006/jmbi.2000.3542>.
40. Johnson A, O'Donnell M. 2005. Cellular DNA replicases: components and dynamics at the replication fork. *Annu Rev Biochem* 74:283–315. <https://doi.org/10.1146/annurev.biochem.73.011303.073859>.
41. Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M. 2010. Single-stranded DNA transposition is coupled to host replication. *Cell* 142:398–408. <https://doi.org/10.1016/j.cell.2010.06.034>.
42. Soler-Bistué A, Mondotte JA, Bland MJ, Val ME, Saleh MC, Mazel D. 2015. Genomic location of the major ribosomal protein gene locus determines *Vibrio cholerae* global growth and infectivity. *PLoS Genet* 11:e1005156. <https://doi.org/10.1371/journal.pgen.1005156>.
43. Val ME, Marbouty M, de Lemos Martins F, Kennedy SP, Kemble H, Bland MJ, Possoz C, Koszul R, Skovgaard O, Mazel D. 2016. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Sci Adv* 2:e1501914. <https://doi.org/10.1126/sciadv.1501914>.
44. Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Ren Y, Cheng J, Wang W, Wang J, Qian W, Li D, Wang L. 2008. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS One* 3:e4053. <https://doi.org/10.1371/journal.pone.0004053>.
45. Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, Haley BJ, Taviani E, Jeon YS, Kim DW, Lee JH, Brettin TS, Bruce DC, Challacombe JF, Detter JC, Han CS, Munk AC, Chertkov O, Meincke L, Saunders E, Walters RA, Huq A, Nair GB, Colwell RR. 2009. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci U S A* 106:15442–15447. <https://doi.org/10.1073/pnas.0907787106>.
46. Rowe-Magnus DA, Guerout AM, Mazel D. 2002. Bacterial resistance evolution by recruitment of super-integration gene cassettes. *Mol Microbiol* 43:1657–1669. <https://doi.org/10.1046/j.1365-2958.2002.02861.x>.
47. Engelstädter J, Harms K, Johnsen PJ. 2016. The evolutionary dynamics of integrations in changing environments. *ISME J* 10:1296–1307. <https://doi.org/10.1038/ismej.2015.222>.
48. Gao F, Luo H, Zhang CT. 2013. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res* 41:D90–D93. <https://doi.org/10.1093/nar/gks990>.
49. Lorenz R, Bernhart SH, Höner Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA package 2.0. *Algorithms Mol Biol* 6:26. <https://doi.org/10.1186/1748-7188-6-26>.
50. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* 101:7287–7292. <https://doi.org/10.1073/pnas.0401799101>.
51. Hofacker IL, Lorenz R. 2013. Predicting RNA structure: advances and limitations. *Methods Mol Biol* 1086:1–19. [https://doi.org/10.1007/978-1-62703-667-2\\_1](https://doi.org/10.1007/978-1-62703-667-2_1).

## 2.3 Conclusion

This chapter presents the first tool to accurately identify all types of integrans and their components with high sensitivity and specificity. It led to the most exhaustive study of integrans, allowing a precise quantification of integrans traits, such as size, *attC* sites similarities, distribution and frequency in bacterial genomes. Notably, we showed an unexpected pattern of integran frequency with the size of the genome, where large genomes are less likely to carry integrans, although they are known to engage in numerous horizontal gene transfer events [147]. This works also reveal the puzzling absence of integrans in  $\alpha$ -Proteobacteria and on the large clades of monoderm bacteria. This suggests a specific mechanism hampering the stabilization of integran in those lineages since class 1 integran have been observed [139]. Another interesting outcome of this study is the quantification of CALINs elements. Although the integration of the cassettes at secondary sites is known for a long time [159], these elements have been overlooked, notably because of the lack of tools to identify them. IntegronFinder allowed to detect CALINs, and stressed their abundance in bacterial genomes. They represent a large pool of cassettes with a variety of genes, mainly of unknown function, which could be mobilized by other integrans. Their role in integran evolution remains to be elucidated.

In another study, we took the advantage of the developed tool to complement experimental data on the dynamic of excision of the integran cassette. This work shows that the relative position of the recombination site with respect to the replication fork's orientation has an impact on integran dynamics. Depending on the chromosomal strand on which the recombinogenic strand of the *attC* site is, the cassettes will be more or less mobilizable by an integran-integrase. This study also highlighted a bias for short CDS compared to their host genome, whose reason is not linked to the recombination mechanism, and remains unexplained.

Overall, these results improve our understanding of integran evolution. They raise new questions on the life cycle of the cassettes. Creation of cassettes seems to require short CDS as most of their CDS are small. Cassettes are thought to be generated by sedentary chromosomal integrans if their creation is dependent on homogeneous *attC* sites. However, the hypothesis suggesting that newly created cassettes should have similar recombination sites is just as likely as the opposite (different recombination sites). The former hypothesis relies on the fact that sedentary chromosomal integrans emerge earlier, tend to carry many cassettes with highly similar *attC* sites, which serve as a pool of cassettes for mobile integrans. The high similarity among recombination sites of the sedentary chromosomal integran suggests that the creation relies on a copy-pasting mechanism. But the repetition of similar recombination sites can be explained by duplication events of cassettes upon insertion of the excised cassette in the conserved integran deriving from the replication of the top strand [66]. Although this process creates cassettes, it does not create *de novo* cassettes, by associating a recombination sites to a new gene. If the process of cassette creation does not require similar *attC* sites, then, generation

of *attC* sites might be no different from the process creating other specific recombination sites, starting with *attI* site, whose creation process remains unknown. This process could be random, where sequences looking like *attC* sites are generated randomly by mutations. In presence of an integron-integrase the pseudo-*attC* site endure positive selection for a proper *attC* site. Given the false positive rate of IntegronFinder, there is about one false *attC* site every 20Mb, or about one every 4 to 5 genomes. These *attC* sites are considered false by precaution, but could be functional, or at the beginning or end of their recombination site life-cycle. An evolution experiment with those *attC* sites would verify whether they are functional, and if the integron-integrase can impose a positive selective pressure on them. The combination with another close pseudo-*attC* might lead the creation of a cassette.

The fate of gene cassettes is also poorly understood. The fact that the integron-integrase can act in *trans* renders the presence of CALINs as full entity of integrons along with In0 elements and complete integrons. Besides, *attC* sites within CALINs tend to be more homogeneous than those within a mobile integron, thus, if the creation of cassettes involves similar *attC* sites, then CALINs could be a intermediate of choice. It is nonetheless likely that some CALINs at least are the results of integrons degradation, due to the over-representation of IS and the presence of degraded integrase nearby. It could be interesting to study the dynamic of CALINs from degraded integrons, and assess whether the gene or the recombination site degrades faster, or if both are degraded simultaneously. If the gene is lost first, one can imagine, that the *attC* might be recycled for another cassette. The frequency of gene cassettes domesticated by bacterial chromosome would also provide valuable information on bacterial genome evolution. In depth analysis of single *attC* sites, and CALIN in general, could provide hints toward *de novo attC* site creation and cassette degradation. For instance, the study of the association between an *attC* site and the associated CDS would be pertinent to comprehend the fate of cassettes. Indeed, the *attC* site in its excised form is always the same for a given cassette. Following how frequent the same *attC* site is linked to the same gene would provide useful data to comprehend these elements and their dynamic.

Another interesting perspective from this work is that integrons are not that abundant after all. Different major clades such as Firmicutes or  $\alpha$ -Proteobacteria are devoid of these elements, whereas pathogens with multi-resistance to antibiotics are found in those clades. This observation suggests that integrons are counter selected in those lineages and therefore they must carry a high cost to their host or encounter incompatibility issues. For instance, a mechanism involving host factors could destabilize the integron leading to its loss. Another possibility would be that the integron system is not well regulated in these lineages and becomes toxic for its host. A third hypothesis would be that clades with integrons are able to produce cassettes and thus integrons, while clades devoid of integrons cannot produce them. Assuming a degradation rate of integron equivalent in both groups (creating integrons or not), the net

result is a deficit of integrans in absence of creation. Given the high diversity of the bacteria life styles, all three hypotheses are probably intertwined. The comprehension of this system preventing integrans persistence could improve our understanding of integrans biology, but could also provide insights in developing strategies to eradicate class 1 integrans or to control their dissemination.

Finally, most results presented here rely on the typical dichotomy between mobile and sedentary integrans. Although this dichotomy shows quantitative differences which inform on their biology, a continuum of integrans exists in between. Non-classified integrans actually represent more than 50% of the dataset used in the first article, and more than 300 corresponding cassettes have never been studied in detail. Thus, increasing our knowledge on integron evolution and notably on integron-cassettes relies on the study of less known integrans for which we already have access, but also on sequencing data over short time-scale to be able to follow the quick evolutionary pace of this system. Improving further IntegronFinder to make it more accessible will allow more people to tackle these questions and will hopefully provide better understanding of integron evolution, leading to better management of antibiotic resistance dissemination mediated by this mechanism.



## **Part III**

# **Conclusions and Perspectives**





This work aimed at better understanding two genetic elements, conjugative elements and integrons, infamous for their important role in the capture and spread of antibiotic resistance genes. Part of the research focused on variants of these elements associated with antibiotic resistance. I developed and made available for the community accurate tools and methods to reveal the breadth of diversity of conjugative elements and integrons. The analysis of these elements revealed similarities and differences between them.

Both conjugative elements and integrons are widespread among bacterial clades. An updated distribution of these elements is depicted in Figure 2.2. This dataset<sup>1</sup> is still highly biased toward pathogenic bacteria, as highlighted by the high relative abundance of genomes in Proteobacteria (split in its five main phyla) and Firmicutes. Conjugative elements are frequent in major clades, and few of them lack conjugative elements. Importantly, clades where we did not identify conjugative elements may actually have them. In particular, several known conjugative systems in Firmicutes and Actinobacteria were left undetected because the relaxase is not known. Furthermore, some genomes lack conjugative elements but have mobilizable elements suggesting that conjugation is also at least occasionally occurring in these species. Integrons are also found in diverse clades, but often at very small frequency. Large, highly sampled clades such as Firmicutes and  $\alpha$ -Proteobacteria lack complete integrons. The presence of CALINs

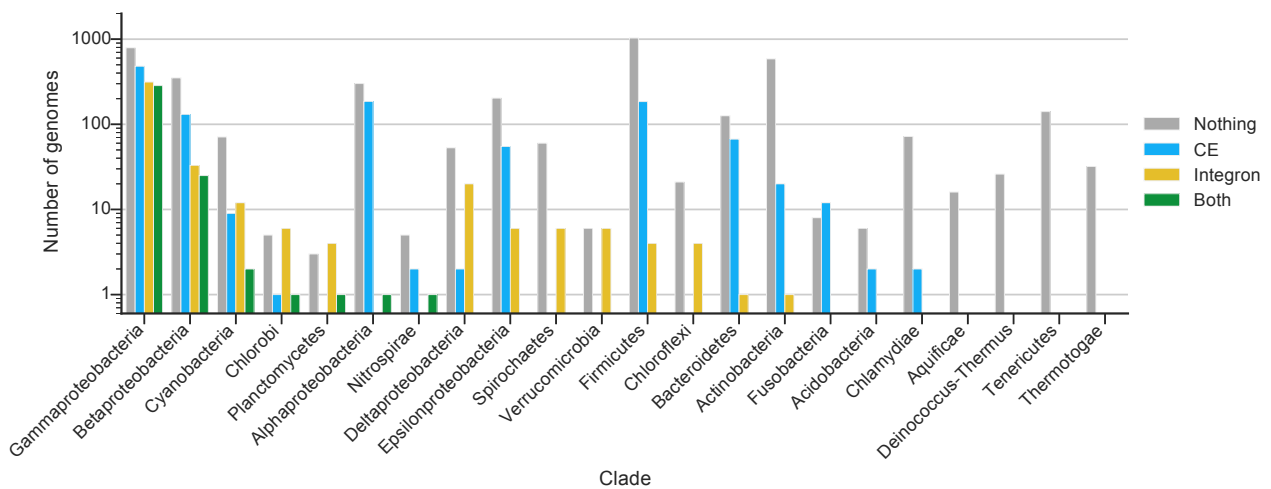


Figure 2.2: Distribution of conjugative elements (CE) and integrons in bacterial clades. The Y-axis is shown in a  $\log_{10}$  scale. The grey bars represent the number of genomes with neither integron nor CE. The blue bars represent the number of genomes having only at least one CE. The yellow bars represent the number of genomes having only at least one integron (complete, In0 or CALIN). The green bars represent the number of genomes having both CE and integron. Planctomycetes, Nitrospirae, Firmicutes, Chloroflexi, Bacteroidetes, Actinobacteria and Acidobacteria do not contain complete integrons (only In0 and/or CALINs).

<sup>1</sup>The dataset is composed of 5562 bacterial genomes obtained from NCBI RefSeq in November 2016

and In0, but not complete integrons (Figure 2.2, legend), in these clades suggest that either the former are more mobile, or they result from local genetic degradation of the latter.

This is in line with the hypothesis that complete integrons cannot establish themselves in these genetic backgrounds. A key question for further research remains the why of this incompatibility.

Although both elements are widespread, they do not share the same distribution when looking at their host's genome size (Figure 2.3). The positive correlation of the size of the conjugative elements with the size of their host is explained by the fact that larger genomes engage more frequently in horizontal transfer (although one might also argue that these genes endure more transfer because they have more mobile elements) [147, 41]. The smallest genomes endure little horizontal gene transfer, often because they have endosymbiosis lifestyles leading to their sexual isolation [134]. While smaller genomes also have fewer integrons, there is no good correlation between the frequency of integrons and genome size for the larger genomes. This was an unexpected result because the frequency of transfer in these genomes should allow them to accumulate more integrons. One would have expected a similar trend between integrons and conjugative elements. As seen in Figure 2.2, most of the integrons and conjugative elements are in  $\gamma$ -Proteobacteria, and the trends we observe might be due to this clade. To better comprehend the relation between the accumulation of a mobile genetic element with the size of the genome, it is important to take into account the sampling bias, by weighting the frequencies with the phylogenetic inertia. For instance, the high frequency of integrons in the range of genome sizes from 4 to 6 Mb might be due to the oversampling of nosocomial bacteria in which class 1 integrons are frequent.

Another feature shared between integrons and conjugative elements is the high modularity of the system, characteristic of other mobile genetic elements. Modularity facilitates the gain and loss of an accessory function without interrupting other functions. It also facilitates recombination between mobile elements [191]. Integrons are by definition modular, due to their array of cassettes in which each cassette forms a module. ICEs are also modular, they have an integration module, a conjugation module, and one or more cargo regions. CPs are modular too [191], but proper quantification of their modular organization would be needed, to assess whether they have a higher or lower modularity than ICEs. One would expect CPs to have more smaller modules since we have seen that they engage in more gene exchanges. The consequence of this modularity is the fact that outside of the focal region (*e.g.* the conjugative system), the other modules are versatile, and in case of mobile genetic elements, this variability is often sizeable. Indeed, most of the integron cassettes and genes in cargo regions of conjugative elements are of unknown function. Further work leading to their accurate annotation would be necessary to comprehend fully the potential of these elements. However, this task remains complicated exactly because of their high genetic diversity and their lack of homologs of known function for

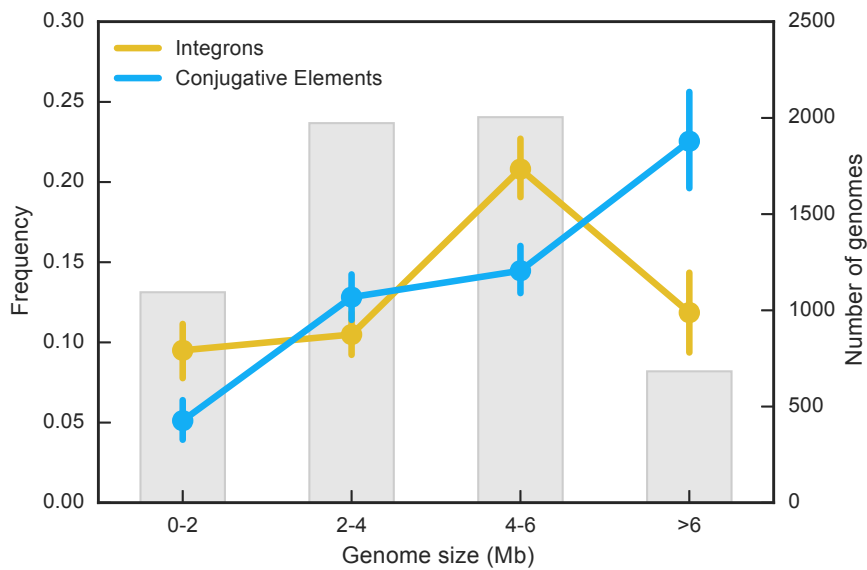


Figure 2.3: Frequency of integrons and conjugative elements as function of the host's genome size. The line plot represents the frequency (left axis) of a given element for a given genome size category. Small colored vertical bars represent the 95% confidence interval for the mean, represented as a point. The barplot represent the number of genomes (right axis) within a given genome size category.

most of their genes. The annotation of these genes in integrons is even made harder because these genes are smaller and lack genetic context (they are not embedded in operons with genes of known function).

The modularity of mobile elements also facilitates the activation of certain functions by other elements in the genome. Conjugative elements can lose part of the conjugative system and become mobilizable, likewise, the two platforms of integrons can split and create In0 and CALINs elements. The study of these derivative elements is of importance because they can reveal whether the element is being degraded, in formation, or a singular entity. Their study is still in the infancy and should deserve more attention. Notably because of their evolutionary potential for the host in case of domestication or for the MGE otherwise. However, modularity can also complicate the task of their detection. Indeed, CALINs elements were not described before the appropriate tool was made to reveal their abundance. Besides, the exhaustive detection of mobilizable elements by conjugation (integrative or not) seems complicated in a near future. Indeed the minimum component necessary for mobilization is the *oriT*, which has a small size and evolves fast [72]. Currently, only experimental methods can identify an *oriT* with certainty. The detection of mobilizable or conjugative elements can even be more complicated given that some ICEs are split upon integration in the genome [98], making their complete identification impossible with the current tools.

The prediction of the type of mobility and the possible integration sites of an element given its sole sequence could have major impact on public health surveillance. For instance, many elements like small plasmids potentiate the evolution of antibiotic resistance genes [166] but their type of mobility remains unknown [174]. Apart from antibiotic resistance or other cargo genes, the transfer and integration themselves can have a great impact on the host's phenotype as observed in a strain of *Elizabethkingia anophelis*, where an ICE inserted within a gene involved in DNA repair may have increased the mutation rate of an outbreak's strain [154].

Finally, the task of identifying integrons and conjugative elements can be improved, especially for ICEs, which are imprecisely delimited. During this thesis I tried to tackle this problem with two approaches. First, I tried to identify the precise integration sites of ICEs (attL and attR) using statistical methods such as Gibbs sampling [118] or expectation maximization methods [11], but we abandoned rapidly given the small size and the degenerated nature of the recombination sites, the large size of the region of interest and the relatively small amount of sequences we had, which were not sufficient for these statistical methods. I then tried a method based on an idea of Nick Croucher, which suggested to look for excised ICEs in next generation sequencing data, with the goal of finding reads overlapping the beginning and end of my delimited ICE (in terms of gene content). Such reads would reveal the precise site of insertion in the chromosome and the recombination site of the ICE. However, after some trials, this turned out more complicated than expected. A reason was probably that the ICE we were investigating did not excise sufficiently frequently to be sequenced in the (usually excellent) growth conditions used previous to sequencing. Finally, more genomic data with higher genetic diversity would provide valuable information on the conserved pattern of the ICE limits. Furthermore, more powerful statistical tools exist nowadays that were not conceivable for biological applications a few years ago. These deep learning methods are now starting to be used in biology for the detection of DNA motifs [5, 7]. However, they require a huge amount of annotated sequence data to be trained on before making any prediction. To do so, they often rely on high throughput experiments [5], or on simulated data [171]. Semi-automatic methods as the one allowing the delimitation of ICEs could fill this gap by providing sufficient labelled data which could be used to train more complex machine learning methods.

The tools, methods and data I produced in this thesis could be valuable for the field of microbial genomics. First, they allowed detailed analyses of two major genetic elements and brought new questions to the field. Second, they might improve genome assembly pipelines where contigs tend to break near repeated sequences, abundant in MGE, preventing the complete assembly of a genome. Given more detection of integrons or conjugative elements, one might be able to assemble contigs that likely belong to the same mobile genetic element. Finally, these methods could fill the gap between experimental biology producing highly qualitative data in low quantity, and the need for high quality data in huge quantity to train more sophisticated

statistical models.



# List of Figures

|     |  |     |
|-----|--|-----|
| 1   | DNA . . . . .  | 14  |
| 2   | Number of prokaryotic genomes sequenced . . . . .                                      | 15  |
| 3   | The various mobile genetic elements and corresponding transfer mechanisms . .          | 19  |
| 4   | Schematic representation of the Pan- and Core-genomes . . . . .                        | 22  |
| 5   | Schema of an organized bacterial genome . . . . .                                      | 23  |
| 6   | The matryoshka's nature of MGE . . . . .   | 27  |
| 7   | Different possible outcomes of site-specific recombination . . . . .                   | 34  |
| 8   | Schema of a tyrosine-recombinase strand exchange mechanism . . . . .                   | 35  |
| 9   | Schema of a toy HMM and of a small HMM profile . . . . .                               | 39  |
| 10  | Covariance Model example . . . . .   | 41  |
| 1.1 | Phylogeny of VirB4 defining eight MPF types and their associated replicon type         | 47  |
| 1.2 | Schema of the conjugation mechanism and its main components . . . . .                  | 48  |
| 2.1 | Schema of an integron and an <i>attC</i> site sequence and secondary structure . . . . | 131 |
| 2.2 | Distribution of conjugative elements and integrons in bacterial clades . . . . .       | 169 |
| 2.3 | Frequency of integrons and conjugative elements as function of the genome size .       | 171 |





# Bibliography



- 
- [1] Sophie S Abby, Bertrand Néron, Hervé Ménager, Marie Touchon, and Eduardo P C Rocha. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS one*, 9(10):e110726, jan 2014.
- [2] E.P. Abraham and E. Chain. An Enzyme from Bacteria able to Destroy Penicillin. *Nature*, 3713:837, 1940.
- [3] H. W. Ackermann and D. Prangishvili. Prokaryote viruses studied by electron microscopy. *Archives of Virology*, 157(10):1843–1849, 2012.
- [4] Tomoichiro Akiba, Kotaro Koyoma, Yoshito Ishiki, Sadao Kimura, and Toshio Fukushima. On the Mechanism of the development of multiple drug resistant clones of *Shigella*. *Japan J. Microb.*, 4(2), 1960.
- [5] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8):831–838, 2015.
- [6] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
- [7] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Mol Syst Biol*, 12, 2016.
- [8] J. Arthur Harris. Pearl And Jennings On Assortative Conjugation In The Protozoa. *Science*, 35(906):740–745, 1912.
- [9] Oswald T Avery, Colin M Macleod, and Maclyn Mccarty. Studies On The Chemical Nature Of The Substance Inducing Transformation Of Pneumococcal Types. *Journal of Experimental Medicine*, 79(2):137–158, 1944.
- [10] P Ayoubi, A O Kilic, and M N Vijayakumar. Tn5253, the pneumococcal omega (cat tet) BM6001 element, is a composite structure of two conjugative transposons, Tn5251 and Tn5252. *J. Bacteriol.*, 173(5):1617–1622, mar 1991.
- [11] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue):W202–8, jul 2009.
- [12] Xavier Bellanger, Sophie Payot, Nathalie Leblond-bourget, and Gérard Guédon. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiology Reviews*, 38:720–760, 2014.
- [13] Andrea J Berman, Satwik Kamtekar, Jessica L Goodman, José M Lázaro, Miguel de Vega, Luis Blanco, Margarita Salas, and Thomas A Steitz. Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. *The EMBO Journal*, 26(14):3494–3505, 2007.
- [14] David Bikard, Stéphane Julié-Galau, Guillaume Cambray, and Didier Mazel. The synthetic integron: an in vivo genetic shuffling device. *Nucleic acids research*, 38(15):e153, aug 2010.
- [15] L. Bissonnette and P. H. Roy. Characterization of In0 of *Pseudomonas aeruginosa* plasmid pVS1, an ancestor of integrons of multiresistance plasmids and transposons of gram- negative bacteria. *Journal of Bacteriology*, 174(4):1248–1257, 1992.
- [16] G Blakely, S Colloms, G May, M Burke, and D Sherratt. *Escherichia coli* XerC recombinase is required for chromosomal segregation at cell division. *The New biologist*, 3(8):789–98, aug 1991.

- [17] Louis-Marie Bobay, Marie Touchon, and Eduardo P. C. Rocha. Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences*, 111(33):12127–12132, 2014.
- [18] Louis-Marie Marie Bobay, Eduardo P C Rocha, and Marie Touchon. The adaptation of temperate bacteriophages to their host genomes. *Molecular biology and evolution*, 30(4):737–51, apr 2013.
- [19] Yan Boucher, Maurizio Labbate, Jeremy E Koenig, and H W Stokes. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in microbiology*, 15(7):301–9, jul 2007.
- [20] Marie Bouvier, Gaëlle Demarre, and Didier Mazel. Integron cassette insertion: a recombination process involving a folded single strand substrate. *The EMBO journal*, 24(24):4356–67, dec 2005.
- [21] E Fidelma Boyd, Salvador Almagro-Moreno, and Michelle a Parent. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends in microbiology*, 17(2):47–53, mar 2009.
- [22] D. J. Brenner, G. R. Fanning, G. V. Miklos, and A. G. Steigerwalt. Polynucleotide Sequence Relatedness Among Shigella Species. *International Journal of Systematic Bacteriology*, 23(1):1–7, 1973.
- [23] D. J. Brenner, A. G. Steigerwalt, H. G. Wathen, R. J. Gross, and B. Rowe. Confirmation of aerogenic strains of Shigella boydii 13 and further study of Shigella serotypes by DNA relatedness. *Journal of Clinical Microbiology*, 16(3):432–436, 1982.
- [24] Mathieu Brochet, Christophe Rusniok, Elisabeth Couvé, Shaynoor Dramsi, Claire Poyart, Patrick Trieu-Cuot, Frank Kunst, and Philippe Glaser. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 105(41):15961–6, 2008.
- [25] Vincent Burrus, Guillaume Pavlovic, Bernard Decaris, and Gérard Guédon. Conjugative transposons : the tip of the iceberg. *Molecular Microbiology*, 46:601–610, 2002.
- [26] Guillaume Cambray, Anne-Marie Guerout, and Didier Mazel. Integrons. *Annual review of genetics*, 44:141–66, jan 2010.
- [27] Allan Campbell. The future of bacteriophage biology. *Nature reviews. Genetics*, 4(6):471–477, 2003.
- [28] Nicolas Carraro and Vincent Burrus. The dualistic nature of integrative and conjugative elements. *Mobile Genetic Elements*, 5(6):98–102, 2015.
- [29] 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. *PLOS Genetics*, 11(6):e1005298, 2015.
- [30] Somdatta Chatterjee, Ayan Mondal, Shravani Mitra, and Sulagna Basu. Acinetobacter baumannii transfers the bla NDM-1 gene via outer membrane vesicles. *Journal of Antimicrobial Chemotherapy*, 72:2201–2207, 2017.
- [31] Noam Chomsky. *Transformational Analysis*. PhD thesis, University of Pennsylvania, 1955.
- [32] Noam Chomsky. Three Models For the Description of Language. *IRE Transact. Information Theory*, 2:113–124, 1956.
- [33] Noam Chomsky. On certain formal properties of grammars. *Information and Control*, 2(2):137–167, jun 1959.
- [34] Anthony Rm Coates, Gerry Halls, and Yanmin Hu. Novel classes of antibiotics or more of the same? *British Journal of Pharmacology*, 163(1):184–194, 2011.

- 
- [35] Tom Coenye and Peter Vandamme. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 228(1):45–49, 2003.
- [36] S. N. Cohen, A. C. Y. Chang, H. W. Boyer, and R. B. Helling. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences*, 70(11):3240–3244, 1973.
- [37] Marta Colomer-Lluch, Juan Jofre, and Maite Muniesa. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS ONE*, 6(3), 2011.
- [38] Stephen Cooper and Charles E. Helmstetter. Chromosome Replication and the division cycle of *Escherichia coli* B/r. *Journal of Molecular Biology*, 31:519–540, 1968.
- [39] Tim F Cooper and Jack A Heinemann. Postsegregational killing does not increase plasmid stability but acts to mediate the exclusion of competing plasmids. *Proc Natl Acad Sci U S A*, 97(23):12643–12648, 2000.
- [40] Tim F Cooper, Tiago Paixã O, and Jack A Heinemann. Within-host competition selects for plasmid-encoded toxin –antitoxin systems. *Proc. R. Soc. B*, 277:3149–3155, 2010.
- [41] O. X. Cordero and P. Hogeweg. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of the National Academy of Sciences*, 106(51):21748–21753, 2009.
- [42] Francis H. C. Crick, Leslie Barnett, Sydney Brenner, and R.J. Watts-Tobin. General Nature Of The Genetic Code For Proteins. *Nature*, 4809, 1961.
- [43] L C Crossman. Plasmid replicons of *Rhizobium*. *Biochemical Society Transactions*, 33(1):157–158, 2005.
- [44] Nicholas J Croucher, Simon R Harris, Lars Barquist, Julian Parkhill, Stephen D Bentley, and Xavier Didelot. A High-Resolution View of Genome-Wide Pneumococcal Transformation. *PLoS Pathog*, 8(6), 2012.
- [45] Nicholas J. Croucher, Rafal Mostowy, Christopher Wymant, Paul Turner, Stephen D. Bentley, and Christophe Fraser. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. *PLoS Biology*, 14(3):e1002394, mar 2016.
- [46] Lloyd Czaplewski, Richard Bax, Martha Clokie, Mike Dawson, Heather Fairhead, Vincent A Fischetti, Simon Foster, Brendan F Gilmore, Robert E W Hancock, David Harper, Ian R Henderson, Kai Hilpert, Brian V Jones, Aras Kadioglu, David Knowles, Sigríður Ólafsdóttir, David Payne, Steve Projan, Sunil Shaunak, Jared Silverman, Christopher M Thomas, Trevor J Trust, Peter Warn, and John H Rex. Alternatives to antibiotics - a pipeline portfolio review. *The Lancet Infectious Diseases*, 16:239–251, 2016.
- [47] Julian Davies. Inactivation of antibiotics and the dissemination of resistance genes. *Science*, 264(5157):375–382, 1994.
- [48] Arthur L Delcher, Douglas Harmon, Simon Kasif, Owen White, and Steven L Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23):4636–4641, 1999.
- [49] Keith M Derbyshire and Todd A Gray. Distributive Conjugal Transfer : New Insights into Horizontal Gene Transfer and Genetic Exchange in Mycobacteria. *Microbiology Spectrum*, 2(1):1–19, 2014.
- [50] Tatiana Dimitriui, Chantal Lotton, Julien Bénard-Capelle, Dusan Misevic, Sam P Brown, Ariel B Lindner, and François Taddei. Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30):11103–8, 2014.

- [51] Matthew S. Dodd, Dominic Papineau, Tor Grenne, John F. Slack, Martin Rittner, Franco Pirajno, Jonathan O’Neil, and Crispin T. S. Little. Evidence for early life in Earth’s oldest hydrothermal vent precipitates. *Nature*, 543(7643):60–64, 2017.
- [52] Sara Domingues, Gabriela Jorge Da Silva, Kaare Magne Nielsen, and Kaare M Nielsen. Global dissemination patterns of common gene cassette arrays in class 1 integrons. *Microbiology*, 131:1313–1337, 2015.
- [53] Emilie Dordet Frisoni, Marc Serge Marends, Eveline Sagné, Laurent Xavier Nouvel, Romain Guérillot, Philippe Glaser, Alain Blanchard, Florence Tardy, Pascal Sirand-Pugnet, Eric Baranowski, and Christine Citti. ICEA of *Mycoplasma agalactiae*: A new family of self-transmissible integrative elements that confers conjugative properties to the recipient strain. *Molecular Microbiology*, 89(6):1226–1239, 2013.
- [54] P Doty, H Boedtker, Jr Fresco, R Haselkorn, and M Litt. Secondary structure in ribonucleic acids. *Proc Natl Acad Sci U S A*, 45:482–499, 1959.
- [55] François Drouin, Josiane Mélançon, and Paul H. Roy. The *intI*-like tyrosine recombinase of *Shewanella oneidensis* is active as an integron integrase. *Journal of Bacteriology*, 184(6):1811–1815, 2002.
- [56] Gyanendra P Dubey, Ganesh Babu, Malli Mohan, Anna Dubrovsky, Eilon Sherman, Ohad Medalia, and Ben-Yehuda Correspondence. Architecture and Characteristics of Bacterial Nanotubes. *Developmental Cell*, 36:453–461, 2016.
- [57] Gyanendra P Dubey and Sigal Ben-Yehuda. Intercellular Nanotubes Mediate Bacterial Communication. *Cell*, 144:590–600, 2011.
- [58] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence analysis*. Cambridge University press, 1998.
- [59] Judith Ebel-Tsipis, David Botstein, and Maurice S. Fox. Generalized transduction by phage P22 in *Salmonella typhimurium*. *Journal of Molecular Biology*, 71(2):433–448, nov 1972.
- [60] Sean R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6(9):755–763, jun 1996.
- [61] Sean R Eddy. Profile hidden Markov models. *Bioinformatics Review*, 14(9):755–763, 1998.
- [62] Sean R. Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Computational Biology*, 4(5), 2008.
- [63] Sean R. Eddy. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195, oct 2011.
- [64] François Enault, Arnaud Briet, Léa Bouteille, Simon Roux, Matthew B Sullivan, and Marie-Agnès Petit. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME Journal*, 11(1):237–247, 2017.
- [65] Patricia Escobar-Paamo, Catherine Giudicelli, Claude Parsot, and Erick Denamur. The Evolutionary History of *Shigella* and Enteroinvasive *Escherichia coli* Revised. *Journal of molecular evolution*, 57:140–148, 2003.
- [66] José Antonio Escudero, Céline Loot, Aleksandra Nivina, and Didier Mazel. The Integron: Adaptation On Demand. *Microbiology spectrum*, 3(2):1–22, 2015.
- [67] Edward J Feil. Small change: keeping pace with microevolution. *Nature reviews. Microbiology*, 2(6):483–495, 2004.
- [68] Alexander Fleming. Penicillin, Nobel Lecture. 1945.

- [69] A. Carolin Frank, Cecilia M. Alsmark, Mikael Thollessen, and Siv G. E. Andersson. Functional Divergence and Horizontal Transfer of Type IV Secretion Systems. *Molecular Biology and Evolution*, 22(5):1325–1336, feb 2005.
- [70] A. E. Franke and D. B. Clewell. Evidence for conjugal transfer of a *Streptococcus faecalis* transposon (Tn916) from a chromosomal site in the absence of plasmid DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 45 Pt 1(1):77–80, 1981.
- [71] Rémi Fronzes, Peter J. Christie, and Gabriel Waksman. The structural biology of type IV secretion systems. *Nature Reviews Microbiology*, 7(10):703–714, 2009.
- [72] L S Frost. Conjugation, Bacterial. *Encyclopedia of Microbiology (Third Edition)*, pages 517–531, 2009.
- [73] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature reviews. Microbiology*, 3(9):722–32, sep 2005.
- [74] José L. García and Eduardo Díaz. Plasmids as Tools for Containment. *Microbiology Spectrum*, 2(5), 2014.
- [75] C Genetello, N Van Larebeke, Marcelle Holsters, A De Picker, M Van Montagu, and J Schell. Ti plasmid of *Agrobacterium* as conjugative plasmids. *Nature*, 269:585–586, 1977.
- [76] Kenn Gerdes, Alexander P Gulyaev, Thomas Franch, Kim Pedersen, and Nikolaj D Mikkelsen. Antisense Rna-Regulated Programmed Cell Death. *Annual Review of Genetics*, 31:1–31, 1997.
- [77] Michael Gillings, Yan Boucher, Maurizio Labbate, Andrew Holmes, Samyuktha Krishnan, Marita Holley, and H. W. Stokes. The evolution of class 1 integrons and the rise of antibiotic resistance. *Journal of Bacteriology*, 190(14):5095–5100, 2008.
- [78] Michael R Gillings. Integrons: past, present, and future. *Microbiology and molecular biology reviews*, 78(2):257–77, 2014.
- [79] Michael R. Gillings. Lateral gene transfer, bacterial genome evolution, and the Anthropocene. *Annals of the New York Academy of Sciences*, 1389(1):1–17, 2016.
- [80] Michael R. Gillings. Class 1 integrons as invasive species. *Current Opinion in Microbiology*, 38:10–15, 2017.
- [81] Michael R Gillings, Marita P Holley, H W Stokes, and Andrew J Holmes. Integrons in *Xanthomonas* : A source of species genome diversity. *Proceedings of the National Academy of Sciences*, 102(12):4419–4424, 2005.
- [82] Stephen Jay Gould and Elisabeth S Vrba. Exaptation—a Missing Term in the Science of Form. *Paleobiology*, 8(1):4–15, 1982.
- [83] MB Griffith, Fred. The significance of pneumococcal types. *Journ. of Hyg.*, XXVII(2), 1928.
- [84] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*, 33(Database issue):D121–4, jan 2005.
- [85] Nigel D F Grindley, Katrine L Whiteson, and Phoebe a Rice. Mechanisms of Site-Specific Recombination. *Annu. Rev. Biochem.*, 75:567–605, jan 2006.
- [86] Eduardo A Groisman and Howard Ochman. Pathogenicity Islands: Bacterial Evolution in Quantum Leaps. *Cell*, 87:791–794, 1996.

- [87] Danxia Gu, Ning Dong, Zhiwei Zheng, Di Lin, Man Huang, Lihua Wang, Edward Wai-Chi Chan, Lingbin Shu, Jiang Yu, Rong Zhang, and Sheng Chen. A fatal outbreak of ST11 carbapenem-resistant hyper-virulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. *The Lancet Infectious Diseases*, 3099(17):1–10, 2017.
- [88] Julien Guglielmini, Fernando de la Cruz, and Eduardo P C Rocha. Evolution of conjugation and type IV secretion systems. *Molecular biology and evolution*, 30(2):315–31, feb 2013.
- [89] Julien Guglielmini, Bertrand Néron, Sophie S Abby, María Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P C Rocha. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic acids research*, pages gku194–, mar 2014.
- [90] Julien Guglielmini, Leonor Quintais, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P C Rocha. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS genetics*, 7(8):e1002222, aug 2011.
- [91] Feng Guo, Deshmukh N Gopaul, and Gregory D Van Duyne. Structure of Cre recombinase complexed with DNA in a site- specific recombination synapse. *Nature*, 384:40–46, 1997.
- [92] Sébastien Halary, Jessica W Leigh, Bachar Cheaib, Philippe Lopez, and Eric Bapteste. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):127–32, jan 2010.
- [93] Ruth M. Hall. Integrons and gene cassettes: Hotspots of diversity in bacterial genomes. *Annals of the New York Academy of Sciences*, 1267:71–78, 2012.
- [94] Ruth M. Hall and H. W. Stokes. Integrons or super integrons ? *Microbiology*, 150(1):3–4, 2004.
- [95] Holly L. Hamilton and Joseph P. Dillard. Natural transformation of *Neisseria gonorrhoeae*: From DNA donation to homologous recombination. *Molecular Microbiology*, 59(2):376–385, jan 2006.
- [96] Ellie Harrison and Michael A. Brockhurst. Plasmid-mediated horizontal gene transfer is a coevolutionary process. *Trends in microbiology*, 20(6):262–7, jun 2012.
- [97] Peter W Harrison, Ryan P J Lower, Nayoung K D Kim, and J Peter W Young. Introducing the bacterial ‘chromid’: not a chromosome, not a plasmid. *Trends in microbiology*, 18(4):141–8, apr 2010.
- [98] Timothy L Haskett, Jason J Terpolilli, Amanuel Bekuma, Graham W O’Hara, John T Sullivan, Penghao Wang, Clive W Ronson, and Joshua P Ramsay. Assembly and transfer of tripartite integrative and conjugative genetic elements. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1613358113–, oct 2016.
- [99] W. Hayes. What are episomes and plasmids. In *Bacterial Episomes and Plasmids*, pages 4–8. J. & A. Churchill Ltd., 1969.
- [100] John F Heidelberg, Jonathan a Eisen, William C Nelson, Rebecca a Clayton, Michelle L Gwinn, Robert J Dodson, Daniel H Haft, Å Tettelin, Erin K Hickey, Jeremy D Peterson, Lowell Umayam, Steven R Gill, Karen E Nelson, Timothy D Read, Delwood Richardson, Maria D Ermolaeva, Jessica Vamathevan, Steven Bass, Haiying Qin, Ioana Dragoi, Patrick Sellers, Lisa Mcdonald, Teresa Utterback, Robert D Fleishmann, William C Nierman, Owen White, Steven L Salzberg, Hamilton O Smith, Rita R Colwell, John J Mekalanos, J Craig Venter, Claire M Fraser, H Tettelin, Delwood Richardson, Maria D Ermolaeva, Jessica Vamathevan, Steven Bass, Haiying Qin, Ioana Dragoi, Patrick Sellers, Lisa Mcdonald, Teresa Utterback, Robert D Fleishmann, William C Nierman, Owen White, Steven L Salzberg, Hamilton O



- Smith, Rita R Colwell, John J Mekalanos, J Craig Venter, and Claire M Fraser. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, 406(6795):477–483, 2000.
- [101] Ferdi L. Hellweger, Robert J. Clegg, James R. Clark, Caroline M. Plugge, and Jan-Ulrich Kreft. Advancing microbial sciences by individual-based modelling. *Nature Reviews Microbiology*, 14(7):461–471, 2016.
- [102] Jennifer M Henke and Bonnie L Bassler. Bacterial social engagements. *Trends in Cell Biology*, 14(11):648–656, 2004.
- [103] Bianca Hochhut, Yasmin Lotfi, Didier Mazel, Shah M Faruque, Roger Woodgate, Matthew K Waldor, M Shah, Roger Woodgate, Matthew K Waldor, and Shah M Faruque. Molecular Analysis of Antibiotic Resistance Gene Clusters in *Vibrio cholerae* O139 and O1 SXT Constins. *Antimicrobial Agents and Chemotherapy*, 45(11):2991–3000, 2001.
- [104] Dirk Hofreuter, Stefan Odenbreit, and Rainer Haas. Natural transformation competence in *Helicobacter pylori* is mediated by the basic components of a type IV secretion system. *Molecular Microbiology*, 41(2):379–391, dec 2001.
- [105] Andrew J. Holmes, Michael R. Gillings, Blair S. Nield, Bridget C. Mabbutt, K. M H Nevalainen, and H. W. Stokes. The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environmental Microbiology*, 5(5):383–394, 2003.
- [106] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, jan 2010.
- [107] F Jacob and E L Wollman. Les épisomes, éléments génétiques ajoutés. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, 247(1):154–156, 1958.
- [108] Rasmus Bugge Jensen and Kenn Gerdes. Programmed cell death in bacteria: proteic plasmid stabilization systems. *Mol Microbiol*, 17(2):205–210, 1995.
- [109] Christopher M. Johnson and Alan D. Grossman. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual Review of Genetics*, 49(1):annurev-genet-112414-055018, 2015.
- [110] Calum Johnston, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology*, 12, 2014.
- [111] Michael J Joss, Jeremy E Koenig, Maurizio Labbate, Martin F Polz, Michael R Gillings, Harold W Stokes, W Ford Doolittle, and Yan Boucher. ACID: annotation of cassette and integron data. *BMC bioinformatics*, 10:118, jan 2009.
- [112] A Kerr. Transfer of Virulence between Isolates of *Agrobacterium*. *Nature*, 223(5211):1175–1176, 1969.
- [113] Jana Klimentová and Jiří Stulík. Methods of isolation and purification of outer membrane vesicles from gram-negative bacteria. *Microbiological Research*, 170:1–9, 2015.
- [114] Igor Konieczny, Katarzyna Bury, Aleksandra Wawrzycka, and Katarzyna Wegrzyn. Iterons Plasmids. In *Plasmids—Biology and Impact in Biotechnology and Discovery*. American Society for Microbiology, 2015.
- [115] Yohann Lacotte, Marie-Cécile Ploy, and Sophie Raheison. Class 1 integrons are low-cost structures in *Escherichia coli*. *The ISME Journal*, 1138:1535–1544, 2017.

- [116] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprasad Kora, Trudy Wassenaar, Suresh Poudel, and David W Ussery. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*, 15:141–161, 2015.
- [117] Andrew S Lang, Olga Zhaxybayeva, and J Thomas Beatty. Gene transfer agents: phage-like elements of genetic exchange. *Nature Reviews Microbiology*, 10, 2012.
- [118] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, and J. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, oct 1993.
- [119] J Lederberg and E L Tatum. Gene recombination in *Escherichia coli.*, 1946.
- [120] Joshua Lederberg. Plasmid (1952 – 1997). *Plasmid*, 9:1–9, 1998.
- [121] Catherine a Lee, Ana Babic, and Alan D Grossman. Autonomous plasmid-like replication of a conjugative transposon. *Molecular microbiology*, 75(2):268–79, jan 2010.
- [122] Joshua Lilly and Manel Camps. Mechanisms of Theta Plasmid Replication. In *Plasmids—Biology and Impact in Biotechnology and Discovery*. American Society for Microbiology, 2015.
- [123] Małgorzata B Łobocka, Debra J Rose, Guy Plunkett, Marek Rusin, Arkadiusz Samojedny, Hansjörg Lehnerr, Michael B Yarmolinsky, and Frederick R Blattner. Genome of Bacteriophage P1. *Journal of Bacteriology*, 186(21):7032–7068, 2004.
- [124] Céline Loot, Aleksandra Nivina, Jean Cury, José Antonio Escudero, Magaly Ducos-Galand, David Bikard, Eduardo P. C. Rocha, and Didier Mazel. Differences in Integron Cassette Excision Dynamics Shape a Trade-Off between Evolvability and Genetic Capacitance. *mBio*, 8(2):e02296–16, 2017.
- [125] M G Lorenz and W Wackernagel. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev*, 58(3):563–602, 1994.
- [126] Quinn Lu. Plasmid Vectors for Gene Cloning and Expression. In *Plasmid Biology*, chapter 27. American Society for Microbiology, 2004.
- [127] Douglas MacDonald, Gaëlle Demarre, Marie Bouvier, Didier Mazel, and Deshmukh N Gopaul. Structural basis for broad DNA-specificity in integron recombination. *Nature*, 440(7088):1157–62, apr 2006.
- [128] Jacques Mahillon and Michael Chandler. Insertion Sequences. *Microbiology And Molecular Biology Reviews*, 62(3):725–774, 1998.
- [129] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*, 41(9), 2013.
- [130] Lionel Makart, Florian Commans, Annika Gillis, and Jacques Mahillon. Horizontal transfer of chromosomal markers mediated by the large conjugative plasmid pXO16 from *Bacillus thuringiensis* serovar israelensis. *Plasmid*, 91(April):76–81, 2017.
- [131] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.
- [132] Didier Mazel, Broderick Dychinco, Vera A Webb, and Julian Davies. A Distinctive Class of Integron in the *Vibrio cholerae* Genome. *Science*, 280(5363):605–608, 1998.
- [133] P Mazodier and J Davies. Gene transfer between distantly related bacteria. *Annual review of genetics*, 25(27):147–171, 1991.

- 
- [134] John P. McCutcheon and Nancy A. Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13–26, 2011.
- [135] James O. McInerney, Alan McNally, Mary J. O’Connell, A. T. Lloyd, and A. Eyre-Walker. Why prokaryotes have pangenomes. *Nature Microbiology*, 2(4):17040, mar 2017.
- [136] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Massignani, and Rino Rappuoli. The microbial pan-genome. *Current Opinion in Genetics and Development*, 15(6):589–594, 2005.
- [137] Carolyn a Michael and Maurizio Labbate. Gene cassette transcription in a large integron-associated array. *BMC genetics*, 11:82, 2010.
- [138] Alex Mira, Howard Ochman, and Nancy A. Moran. Deletional bias and the evolution of bacterial genomes, 2001.
- [139] Alexandra Moura, Mário Soares, Carolina Pereira, Nuno Leitão, Isabel Henriques, and António Correia. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics (Oxford, England)*, 25(8):1096–8, apr 2009.
- [140] Eric P Nawrocki, Diana L Kolbe, and Sean R Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, 25(10):1335–7, may 2009.
- [141] Eric Paul Nawrocki. *Structural RNA Homology Search and Alignment Using Covariance Models*. PhD thesis, 2009.
- [142] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O’Neal. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, 53(5):1161–8, 1965.
- [143] Teresa Nogueira, Daniel J. Rankin, Marie Touchon, François Taddei, Sam P. Brown, and Eduardo P C Rocha. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Current biology*, 19(20):1683–91, nov 2009.
- [144] A. Norman, L. H. Hansen, and S. J. Sorensen. Conjugative plasmids: vessels of the communal gene pool. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527):2275–2289, 2009.
- [145] S E Nunes-Düby, H J Kwon, R S Tirumalai, T Ellenberger, and a Landy. Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic acids research*, 26(2):391–406, jan 1998.
- [146] K Ochiai, T Yamanaka, K Kimura, and O Sawada. Inheritance of drug resistance (and its transfer) between Shigella strains and Between Shigella and E. coli strains. *Hihon Iji Shimpo (Japanese)*, 1861(34), 1959.
- [147] H Ochman, J G Lawrence, and E A Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, may 2000.
- [148] Pedro H Oliveira, Marie Touchon, Jean Cury, and Eduardo P. C. Rocha. The chromosomal organization of horizontal gene transfer in Bacteria. *Nature communications*, 8(841):1–10, 2017.
- [149] Pedro H Oliveira, Marie Touchon, and Eduardo P C Rocha. Regulation of genetic flux between bacteria by restriction-modification systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(20):5658–63, may 2016.

- [150] Sarah P. Otto and Yannis Michalakis. The evolution of recombination in changing environments. *Trends in Ecology and Evolution*, 13(4):145–151, 1998.
- [151] R Overbeek, M Fonstein, M D’Souza, G D Pusch, and N Maltsev. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2896–901, 1999.
- [152] Samay Pande, Shraddha Shitut, Lisa Freund, Martin Westermann, Felix Bertels, Claudia Colesie, Ilka B Bischofs, and Christian Kost. Metabolic cross-feeding via intercellular nanotubes among bacteria. *Nature Communications*, 6(6238), 2015.
- [153] John H Paul and Matthew B Sullivan. Marine phage genomics: what have we learned? *Current Opinion in Biotechnology*, 16:299–307, 2005.
- [154] Amandine Perrin, Elise Larsonneur, Ainsley C Nicholson, David J Edwards, Kristin M Gundlach, Anne M Whitney, Christopher A Gulvik, Melissa E Bell, Olaya Rendueles, Jean Cury, Perrine Hugon, Dominique Clermont, Vincent Enouf, Vladimir Loparev, Phalasy Juieng, Timothy Monson, David Warshauer, Lina I Elbadawi, Maroya Spalding Walters, Matthew B Crist, Judith Noble-Wang, Gwen Borlaug, Eduardo P C Rocha, Alexis Criscuolo, Marie Touchon, Jeffrey P Davis, Kathryn E Holt, John R McQuiston, and Sylvain Brisse. Evolutionary dynamics and genomic features of the Elizabethkingia anophelis 2015 to 2016 Wisconsin outbreak strain. *Nature communications*, 8(May):15483, 2017.
- [155] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome biology*, 9(1):R4, 2008.
- [156] Mark Ptashne and A Genetic Switch. *Phage Lambda and Higher Organisms*. 1992.
- [157] Jasna Rakonjac. Filamentous Bacteriophages: Biology and Applications. *eLS*, 2012.
- [158] Rita A Rapa, Maurizio Labbate, and Daniela Ceccarelli. The function of integron-associated gene cassettes in Vibrio species: the tip of the iceberg. *Frontiers in Microbiology*, 125(18):15–1, 2013.
- [159] G D Recchia, H W Stokes, and R M Hall. Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic acids research*, 22(11):2071–2078, 1994.
- [160] R.R. Reed and N.D.F. Grindley. Transposon-mediated site-specific recombination in vitro: DNA cleavage and protein-DNA linkage at the recombination site. *Cell*, 25(3):721–728, sep 1981.
- [161] Eduardo P.C. Rocha. Inference and analysis of the relative stability of bacterial chromosomes. *Molecular Biology and Evolution*, 23(3):513–522, 2006.
- [162] Igor B Rogozin, Kira S Makarova, Janos Murvai, Eva Czabarka, Yuri I Wolf, Roman L Tatusov, Laszlo a Szekely, and Eugene V Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic acids research*, 30(10):2212–2223, 2002.
- [163] Dean A Rowe-magnus, Anne-marie Guerout, Latefa Biskri, and Philippe Bouige. Comparative Analysis of Superintegrons : Engineering Extensive Genetic Diversity in the Vibrionaceae. *Genome research*, pages 428–442, 2003.
- [164] Dean A Rowe-magnus, Anne-marie Guerout, Pascaline Ploncard, Broderick Dychinco, and Julian Davies. The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrons. *Proceedings of the National Academy of Sciences*, 2001.

- 
- [165] José A. Ruiz-Maso, Cristina Machon, Lorena Bordanaba-Ruiseco, Manuel Espinosa, Miquel Coll, and Gloria del Solar. Plasmid Rolling-Circle Replication. In *Plasmids—Biology and Impact in Biotechnology and Discovery*. American Society for Microbiology, 2015.
- [166] Alvaro San Millan, Jose Antonio Escudero, Danna R Gifford, Didier Mazel, and R Craig Maclean. Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nature Ecology & Evolution*, 1(0010):1–8, 2016.
- [167] Borja Sánchez, Philippe Bressollier, and María C Urdaci. Exported proteins in probiotic bacteria: adhesion to intestinal surfaces, host immunomodulation and molecular cross-talking with the host. *FEMS Immunol Med Microbiol*, 54:1–17, 2008.
- [168] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci/ USA*, 74(12):5463–5467, 1977.
- [169] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4):463–481, sep 1951.
- [170] Ron Sender, Shai Fuchs, Ron Milo, T Lee, H Ahn, and S Baek. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biology*, 14(8):e1002533, aug 2016.
- [171] Sara Sheehan and Yun S. Song. Deep Learning for Population Genetic Inference. *PLoS Computational Biology*, 12(3):e1004845, mar 2016.
- [172] Patricia Signier, Edith Goubeyre, and Mick Chandler. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Review*, 38:865–891, 2014.
- [173] Simon Silver and Tapan K Misra. Plasmid-Mediated Heavy Metal Resistances. *Ann. Rev. Microbiol*, 42:717–43, 1988.
- [174] Chris Smillie, M Pilar Garcillán-Barcia, M Victoria Francia, Eduardo P C Rocha, and Fernando de la Cruz. Mobility of plasmids. *Microbiology and molecular biology reviews : MMBR*, 74(3):434–52, sep 2010.
- [175] Erwin F. Smith and C. O. Townsend. A Plant-Tumor of Bacterial Origin. *Nature*, 25(643):671–673, 1907.
- [176] Gerald R. Smith. Conjugational recombination in *E. coli*: Myths and mechanisms. *Cell*, 64(1):19–27, 1991.
- [177] Margaret C. M. Smith and Helena M. Thorpe. Diversity in the serine recombinases. *Molecular microbiology*, 44(2):299–307, apr 2002.
- [178] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Publishing Group*, 16, 2015.
- [179] Anjana Srivatsan, Ashley Tehranchi, David M. MacAlpine, and Jue D. Wang. Co-orientation of replication and transcription preserves genome integrity. *PLoS Genetics*, 6(1), 2010.
- [180] Ofer Stempler, Amit K Baidya, Saurabh Bhattacharya, Ganesh Babu, Malli Mohan, Elhanan Tzipilevich, Lior Sinai, Gideon Mamou, and Sigal Ben-Yehuda. Interspecies nutrient extraction and toxin delivery between bacteria. *Nature Communications*, 8(315), 2017.
- [181] Cameron R Strachan and Julian Davies. The Whys and Wherefores of Antibiotic Resistance. *Cold Spring Harbor perspectives in medicine*, 7(2):a025171, feb 2017.

- [182] J. T. Sullivan, J. R. Trzebiatowski, R. W. Cruickshank, J. Gouzy, S. D. Brown, R. M. Elliot, D. J. Fleetwood, N. G. McCallum, U. Rossbach, G. S. Stuart, J. E. Weaver, R. J. Webby, F. J. de Bruijn, and C. W. Ronson. Comparative Sequence Analysis of the Symbiosis Island of *Mesorhizobium loti* Strain R7A. *Journal of Bacteriology*, 184(11):3086–3095, jun 2002.
- [183] E L Tatum and Joshua Lederberg. Gene recombination in the bacterium *Escherichia coli*. *Journal of Bacteriology*, 53:673–684, 1947.
- [184] Evelien M. te Poele, Henk Bolhuis, and Lubbert Dijkhuizen. Actinomycete integrative and conjugative elements. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 94(1):127–143, 2008.
- [185] Christopher M Thomas and Kaare M Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews. Microbiology*, 3(9):711–21, sep 2005.
- [186] Marie Touchon, Sophie Charpentier, Olivier Clermont, Eduardo P C Rocha, Erick Denamur, and Catherine Branger. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *Journal of Bacteriology*, 193(10):2460–2467, 2011.
- [187] Marie Touchon, Jean Cury, Eun-Jeong Yoon, Lenka Krizova, Gustavo C Cerqueira, Cheryl Murphy, Michael Feldgarden, Jennifer Wortman, Dominique Clermont, Thierry Lambert, Catherine Grillot-Courvalin, Alexandr Nemeč, Patrice Courvalin, and Eduardo Pc Rocha. The Genomic Diversification of the Whole *Acinetobacter* Genus: Origins, Mechanisms, and Consequences. *Genome biology and evolution*, 6(10):2866–2882, oct 2014.
- [188] Marie Touchon, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, Odile Bouvet, Alexandra Calteau, Hélène Chiapello, Olivier Clermont, Stéphane Cruveiller, Antoine Danchin, Médéric Diard, Carole Dossat, Meriem El Karoui, Eric Frapy, Louis Garry, Jean Marc Ghigo, Anne Marie Gilles, James Johnson, Chantal Le Bouguéneç, Mathilde Lescat, Sophie Mangenot, Vanessa Martinez-Jéhanne, Ivan Matic, Xavier Nassif, Sophie Oztas, Marie Agnès Petit, Christophe Pichon, Zoé Rouy, Claude Saint Ruf, Dominique Schneider, Jérôme Turret, Benoit Vacherie, David Vallenet, Claudine Médigue, Eduardo P C Rocha, and Erick Denamur. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS genetics*, 5(1):e1000344, jan 2009.
- [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. *Current Opinion in Microbiology*, 38:66–73, 2017.
- [190] Marie Touchon and Eduardo P C Rocha. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harbor Perspectives in Biology*, 8(1):1–18, 2016.
- [191] Ariane Toussaint and Christophe Merlin. Mobile Elements as a Combination of Functional Modules. *Plasmid*, 47(1):26–35, 2002.
- [192] Todd J Treangen, Ole Herman Ambur, Tone Tonjum, and Eduardo PC Rocha. The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biology*, 9(3):R60, 2008.
- [193] Todd J Treangen and Eduardo P C Rocha. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS genetics*, 7(1):e1001284, jan 2011.

- 
- [194] Guy Tsafnat, Enrico Coiera, Sally R Partridge, Jaron Schaeffer, and Jon R Iredell. Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. *BMC bioinformatics*, 10:281, jan 2009.
- [195] Alan M. Turing. The Chemical Basis of Morphogenesis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 237(641):37–72, 1952.
- [196] F. W. Twort. An Investigation On The Nature Of Ultra-Microscopic Viruses. *The Lancet*, 186(4814):1241–1243, 1915.
- [197] Antony van Leewenhoeck. Concerning Little Animals by Him Observed in Rain-Well-Sea and Snow Water; as Also in Water Wherein Pepper Had Lain Infused. *Philosophical Transactions*, 12:821–831, 1677.
- [198] Sara Vieira-Silva and Eduardo P C Rocha. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS genetics*, 6(1):e1000808, jan 2010.
- [199] Matthew K Waldor and John J Mekalanos. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science*, 272:1910–1914, 1996.
- [200] K.A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).
- [201] William B Whitman, David C Coleman, and William J Wiebe. Perspective Prokaryotes: The unseen majority. *Proc Natl Acad Sci U S A*, 95:6578–6583, 1998.
- [202] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.
- [203] E.-L. Wollman, F. Jacob, and W. Hayes. Conjugation and Genetic Recombination in Escherichia coli K-12. *Cold Spring Harbor Symposia on Quantitative Biology*, 21:141–162, 1956.
- [204] World Health Organization. Antimicrobial resistance - Global Report on Surveillance. Technical report, 2014.
- [205] Qiu E. Yang and Timothy R. Walsh. Toxin–antitoxin systems and their role in disseminating and maintaining antimicrobial resistance. *FEMS Microbiology Reviews*, 187:6094–105, mar 2017.
- [206] Norton D Zinder and Joshua Lederberg. Genetic Exchange in Salmonella. *Journal of Bacteriology*, 64:679–699, 1952.





# Annexes



## Co-authored Manuscript 1: Touchon et al, 2014

This paper represent the work I have done during a 6 months internship in my current laboratory, and finished during my PhD. I contributed to the curation of the data, to the construction of the core and pan-genomes and the phylogenetic trees of the genus *Acinetobacter* as well as the species *Acinetobacter baumannii*.

# The Genomic Diversification of the Whole *Acinetobacter* Genus: Origins, Mechanisms, and Consequences

Marie Touchon<sup>1,2</sup>, Jean Cury<sup>1,2</sup>, Eun-Jeong Yoon<sup>3</sup>, Lenka Krizova<sup>4</sup>, Gustavo C. Cerqueira<sup>5</sup>, Cheryl Murphy<sup>5</sup>, Michael Feldgarden<sup>5</sup>, Jennifer Wortman<sup>5</sup>, Dominique Clermont<sup>6</sup>, Thierry Lambert<sup>3</sup>, Catherine Grillot-Courvalin<sup>3</sup>, Alexandr Nemeč<sup>4,\*</sup>, Patrice Courvalin<sup>3,\*</sup>, and Eduardo P.C. Rocha<sup>1,2,\*</sup>

<sup>1</sup>Microbial Evolutionary Genomics, Institut Pasteur, Paris, France

<sup>2</sup>CNRS, UMR3525, Paris, France

<sup>3</sup>Unité des Agents Antibactériens, Institut Pasteur, Paris, France

<sup>4</sup>Laboratory of Bacterial Genetics, National Institute of Public Health, Prague, Czech Republic.

<sup>5</sup>Broad Institute of Harvard and MIT, Cambridge, Massachusetts

<sup>6</sup>Collection de l'Institut Pasteur, Institut Pasteur, Paris, France

\*Corresponding author. E-mail: anemec@szu.cz, patrice.courvalin@pasteur.fr, erocha@pasteur.fr.

Accepted: October 6, 2014

Data deposition: This project has been deposited in GenBank according to the accession numbers provided in [supplementary material](#) online.

## Abstract

Bacterial genomics has greatly expanded our understanding of microdiversification patterns within a species, but analyses at higher taxonomical levels are necessary to understand and predict the independent rise of pathogens in a genus. We have sampled, sequenced, and assessed the diversity of genomes of validly named and tentative species of the *Acinetobacter* genus, a clade including major nosocomial pathogens and biotechnologically important species. We inferred a robust global phylogeny and delimited several new putative species. The genus is very ancient and extremely diverse: Genomes of highly divergent species share more orthologs than certain strains within a species. We systematically characterized elements and mechanisms driving genome diversification, such as conjugative elements, insertion sequences, and natural transformation. We found many error-prone polymerases that may play a role in resistance to toxins, antibiotics, and in the generation of genetic variation. Surprisingly, temperate phages, poorly studied in *Acinetobacter*, were found to account for a significant fraction of most genomes. Accordingly, many genomes encode clustered regularly interspaced short palindromic repeats (CRISPR)-Cas systems with some of the largest CRISPR-arrays found so far in bacteria. Integrons are strongly overrepresented in *Acinetobacter baumannii*, which correlates with its frequent resistance to antibiotics. Our data suggest that *A. baumannii* arose from an ancient population bottleneck followed by population expansion under strong purifying selection. The outstanding diversification of the species occurred largely by horizontal transfer, including some allelic recombination, at specific hotspots preferentially located close to the replication terminus. Our work sets a quantitative basis to understand the diversification of *Acinetobacter* into emerging resistant and versatile pathogens.

**Key words:** comparative genomics, bacterial genus, evolution, mobile genetic elements, nosocomial pathogens.

## Introduction

In the last few years, a number of studies have harnessed the power of high-throughput sequencing to study epidemiological and evolutionary patterns within bacterial species. Such studies have uncovered patterns of transmission of multidrug resistant clones, the emergence of virulent strains, and their within-host evolution (e.g., Harris et al. 2010; Kennemann et al. 2011; Mather et al. 2013; McGann et al. 2014). There has been considerably less emphasis on the large-scale

sampling and sequencing of broader taxonomic units such as genera. Yet, this is an important level of analysis to understand the emergence of pathogenic species, because genera often include pathogens, commensals, and free-living (environmental) bacteria. The genus *Acinetobacter* is a good example of this as it includes a broad group of biochemically and physiologically versatile bacteria that occupy different natural ecosystems and play an increasing causative role in opportunistic human infections. The taxonomy of the genus consists

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

currently of 33 distinct validly named species (<http://www.bacterio.net/acinetobacter.html>, last accessed June 2014), about ten provisionally termed genomic species delineated by DNA–DNA hybridization (DDH) (Dijkshoorn et al. 2007), several putative new species but not validly named, and a number of strains of an as-yet unknown taxonomic status (Rodriguez-Bano et al. 2006; Yamahira et al. 2008; Smet et al. 2012). Following seminal DDH measures of relatedness (Bouvet and Jeanjean 1989; Tjernberg and Ursing 1989), several distinct phylogenies of the *Acinetobacter* genus have been estimated using one or a few phylogenetic marker genes (Rainey et al. 1994; Yamamoto et al. 1999; Krawczyk et al. 2002; La Scola et al. 2006; Diancourt et al. 2010). These analyses suggested the existence of several groups encompassing phylogenetically close species, including the species of the *Acinetobacter calcoaceticus*–*Acinetobacter baumannii* (ACB) complex (Nemec et al. 2011), proteolytic genomic species (Bouvet and Jeanjean 1989; Nemec et al. 2009), *Acinetobacter guillouiae* and *Acinetobacter bereziniae* (Nemec et al. 2010), or *Acinetobacter nectaris* and *Acinetobacter boissieri* (Alvarez-Perez et al. 2013). However, these previous studies showed very diverse phylogenetic scenarios at higher taxonomic levels with weakly supported internal nodes (Yamamoto et al. 1999; Krawczyk et al. 2002; La Scola et al. 2006; Diancourt et al. 2010). Recent genome-wide analyses produced phylogenies showing that *A. baumannii* is well separated from the other strains of the ACB complex and these from other species in the genus (Chan et al. 2012). Nevertheless, the lack of a robust phylogenetic scenario encompassing all known *Acinetobacter* spp. and the unknown position of their last common ancestor (root of the tree) seriously hampers the understanding of the diversification of this genus.

*Acinetobacter* spp. are among the most frequent causes of hospital-acquired bacterial infections (Peleg et al. 2008). Community-acquired infections are less frequent but have also been reported (Falagas et al. 2007; Eveillard et al. 2013). Although *A. baumannii* is the most frequently identified nosocomial pathogen in the genus, several other species cause occasionally infections in humans including *Acinetobacter nosocomialis*, *Acinetobacter pittii*, and less frequently *Acinetobacter ursingii*, *Acinetobacter haemolyticus*, *Acinetobacter lwoffii*, *Acinetobacter parvus*, and *Acinetobacter junii* (Nemec et al. 2001, 2003; Dijkshoorn et al. 2007; Peleg et al. 2008; Turton et al. 2010; Karah et al. 2011). Clinical cases typically involve not only ventilator-associated pneumonia and septicemia, but also endocarditis, meningitis, burn and surgical wound infections, and urinary tract infections. Importantly, *Acinetobacter* spp. are isolated in the environment and asymptotically associated with humans, albeit the precise environmental reservoirs are unknown. Some species, notably *Acinetobacter baylyi*, are becoming emerging model organisms because of their transformability, metabolic versatility, and genome plasticity

(Barbe et al. 2004; Metzgar et al. 2004; de Berardinis et al. 2008). Interestingly, these same traits, together with intrinsic resistance to toxins and antibiotics, are thought to cause the increasing frequency of *Acinetobacter* spp. as agents of nosocomial infections (Peleg et al. 2012).

In the last decade, a small number of complete genome sequences and a large number of draft sequences of *Acinetobacter* spp. have become available (Barbe et al. 2004; Fournier et al. 2006; Antunes et al. 2013). These studies focused heavily on *A. baumannii* and the ACB complex and on what distinguishes them. Although the diversity of the genus remains largely unexplored, several studies have shown that *A. baumannii* gene repertoires are very diverse, with fewer than half of the genes being part of the species' core-genomes (Adams et al. 2008; Imperi et al. 2011; Sahl et al. 2011; Farrugia et al. 2013). Like for many bacterial clades, a large fraction of the genes of the pan-genome are of unknown function. Accessory functions involved in transport and genetic regulation are also highly variable between strains (Adams et al. 2008; Imperi et al. 2011). Finally, genome-wide comparisons between *A. baumannii* and other species suggest the existence of high genetic diversity in the genus (Vallenet et al. 2008; Fondi et al. 2013).

Somewhat surprisingly, most known virulence factors of *A. baumannii* are found in its core-genome and are present in other species of the genus (Antunes et al. 2011). The multifactorial basis of virulence and the independent emergence of pathogenic strains in the genus suggest that the genetic background has an important role in *Acinetobacter* evolution. To understand the emergence of pathogenic *Acinetobacter*, it is therefore important to study the mechanisms of genetic diversification of the genus. For this, we have carefully sampled a large number of representative strains. Following the sequencing of their genomes, we characterized their genetic diversity and built a robust phylogeny of the entire genus. This information was used to assess pending taxonomic issues, to guide evolutionary studies, and to sample the genus for key mechanisms generating genetic variability. With these data at hand we focused on the genome dynamics of *A. baumannii*.

## Materials and Methods

### Choice of Strains

We analyzed 13 complete genomes retrieved from GenBank RefSeq in February 2013 (Pruitt et al. 2007), two *A. baumannii* genomes sequenced at the Pasteur Institute and at Walter Reed, and 118 genome sequences derived from 116 strains (two sequences were obtained from each of the type strains of *Acinetobacter indicus* and *Acinetobacter brisouii*) which were sequenced at the Broad Institute (see details in [supplementary table S1, Supplementary Material](#) online) (Perichon et al. 2014). The 116 strains were selected from the collections

of *A. Nemece* (strains designated NIPH or ANC) or of the Institut Pasteur (CIP strains) based on polyphasic taxonomic analyses to reflect the currently known breadth of the diversity of the genus *Acinetobacter* at the species level. Overall, 83 strains belonged to 29 validly named species (including *Acinetobacter grimontii*, a junior synonym of *A. junii*, but devoid of *A. boissieri*, *A. harbinensis*, *Acinetobacter puyangensis*, and *Acinetobacter qingfengensis*, which were unavailable at the time), 16 strains to eight genomic species as defined by DDH, and ten strains to seven tentative novel species termed *Acinetobacter* taxons 18–23 and 26. The name "*Acinetobacter bohemicus*" has recently been proposed for taxon 26 (Krizova et al. 2014). Seven remaining strains were closely related to one of the species/taxa but were considered taxonomically unique at the species level based on our previous taxonomic analysis and the average nucleotide identity (ANI) data obtained in this study (these strains are termed *A. calcoaceticus*-like, *A. brisouii*-like, *A. pittii*-like, or Taxon 18-like). The *Acinetobacter* taxons are working taxonomic groups as delineated at the Laboratory of Bacterial Genetics (National Institute of Public Health, Prague) based on the comprehensive physiological/nutritional testing, *rpoB* and 16S rDNA phylogenies and on whole-cell MALDI-TOF profiling. Each of these taxa is, at the species level, clearly distinct from any of the known species with a valid name, genomic species, or species with effectively published names. All validly named species were represented by the respective type strains, whereas each genomic species included a strain used as a reference in previous DDH experiments. If more strains per species or taxon were included, these differed from respective type/reference strains in their microbiological and ecological characteristics. Organisms were grown, according to their physiological requirements, at 30–37 °C in brain-heart infusion broth and agar (Difco Laboratories, Detroit, MI).

#### Core-Genomes

We built core-genomes for the genus and for *A. baumannii*. Orthologs were identified as bidirectional best hits, using end-gap free global alignment, between the proteome of *A. baumannii* AYE as a pivot and each of the other proteomes (133 for the genus and 34 for the species). Hits with less than 40% (genus) or 80% (species) similarity in amino acid sequence or more than 20% difference in protein length were discarded. Genomes from the same species typically show low levels of genome rearrangements and this information can be used to identify orthologs more accurately (Dandekar et al. 1998; Rocha 2006). Therefore, the core-genome of the species was defined as the intersection of pairwise lists of strict positional orthologs (as in Touchon et al. 2009). The core-genomes consist in the genes present in all genomes of the two sets. They were defined as the intersection of the lists of orthologs between pairs of genomes.

#### Pan-Genomes

Pan-genomes were built by clustering homologous proteins into families. We determined the lists of putative homologs between pairs of genomes with BLASTp (Altschul et al. 1997) and used the *e* values ( $<10^{-4}$ ) to cluster them by similarity with Silix v1.2 (Miele et al. 2011). A protein is thus included in the family if it shares a relation of homology to a protein already in the family. Silix parameters were set such that a protein is homolog to another in a given family if the aligned part has at least 35% of identity and represents more than 80% of the smallest protein. The pan-genomes are the full complement of genes in the genus and in the species. The pan-genomes of the 133 *Acinetobacter* proteomes (470,582 proteins) and of the 34 *A. baumannii* proteomes (128,266 proteins) were determined independently. We used a more stringent criterion of protein similarity (60%) to compare the pan-genomes of different species of *Acinetobacter*.

#### Phylogenetic Analyses

Each of the 950 families of proteins of the *Acinetobacter* core-genome was used to produce a multiple alignment with muscle v3.8 (default parameters) (Edgar 2004). Poorly aligned regions were removed with BMGE (Criscuolo and Gribaldo 2010). The phylogenetic tree was inferred using the approximated maximum-likelihood method implemented in FastTree v1.4 with the Whelan and Goldman (WAG) matrix and a gamma correction for variable evolutionary rates (Price et al. 2009). We performed 100 bootstrap experiments on the concatenated sequences to assess the robustness of the topology. To root the genus phylogenetic tree, we used the genomes of species distant from each other and as close to *Acinetobacter* as possible. They were identified using a phylogenetic tree built with the 16S rDNA sequences of all the species of  $\gamma$ -proteobacteria with complete genomes in National Center for Biotechnology Information RefSeq (February 2013). The selection of a single strain and a single 16S copy per species resulted in 189 16S rDNA sequences that were aligned using MAFFT v7.1 (Katoh and Toh 2008). Poorly aligned regions were removed with Gblocks 0.91 b (default parameters) (Castresana 2000). The phylogenetic tree was inferred using PhyML v3.0 under the Hasegawa–Kishino–Yano model and a gamma correction for variable evolutionary rates with eight classes (Gascuel et al. 2010). This tree highlighted *Moraxella catarrhalis* (GenBank ID NC\_014147) and *Psychrobacter* species (GenBank ID NC\_009524) as the two species closest to the *Acinetobacter* genus in our data set. We therefore reconstructed a core-genome of the genus plus these two species (same method as above). This core-genome included 677 protein families (supplementary table S3, Supplementary Material online) and was used to build a tree of the 135 complete genomes using the method mentioned above for the genus.

The reference phylogenetic tree of *A. baumannii* was reconstructed from the concatenated alignments of the

1,590 protein families of the core-genome obtained with muscle v3.8 (default parameters). As 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. Poorly aligned positions were removed with BMGE. The tree was inferred with FastTree v1.4 under the general time reversible model and a gamma correction for variable evolutionary rates with eight classes. We performed 100 bootstrap experiments on the concatenated sequences to assess the robustness of the topology. To root the species tree, we used two closely related strains in the genus: *A. nosocomialis* NIPH 2119<sup>T</sup> and *A. pittii*-like ANC 4052, rebuilt the core-genome of the species including these two outgroups and performed a similar analysis on this data set.

#### Recombination and Population Genetic Analyses

We analyzed recombination in *A. baumannii* with RDP version 4.24 with default parameters except that RDP3 was used with the option "internal references only" and Geneconv with a G-scale mismatch penalty of 3 (Martin et al. 2010). The Phi test was computed using Phi Pack with default parameters, except that we made 10,000 permutations (Bruen et al. 2006). Analysis with or without permutations revealed highly correlated *P* values (Spearman's  $\rho=0.96$ ,  $P < 0.0001$ ). We used the results of the analysis with permutations and applied a sequential Bonferroni correction of the *P* values. The analyses with ClonalFrame (Didelot and Falush 2007) were carried out using default options (except for the options -x 50000 -y 50000) on 30 nonoverlapping sets of approximately 51 genes contiguous in terms of their position in the genome of *A. baumannii* AYE. Analysis of Tajima's *D* and Fay and Wu *H* was carried out with DnaSP 5.10.1 (Librado and Rozas 2009), using a sliding window of 1,000 positions and a step of 250. Statistical tests were performed on the average of the values on nonoverlapping windows. Analysis of *dN* and *dS* was carried out on the concatenate of the alignments of the genes of the core-genome using codeML from the PAML package version 4.4 (runmode=-2, model=2, CodonFreq=2) (Yang 2007).

#### Evolutionary Distances

For each pair of genomes, we computed a number of measures of similarity: 1) The phylogenetic distance was computed from the length of branches in the genus phylogenetic tree using the *cophenetic* function in APE package from R (Paradis et al. 2004), 2) the gene repertoire relatedness (GRR) was computed as the number of homologs shared by the two genomes divided by the number of genes in the smallest genome (Snel et al. 1999), 3) the ANI was computed using Jspecies (Richter and Rossello-Mora 2009), and 4) the spacer repertoire relatedness was computed as the number of similar spacers shared by the two genomes divided by the number of

spacers in the smallest clustered regularly interspaced short palindromic repeats (CRISPR)-arrays of the two genomes.

#### Identification of Specific Genes/Systems

CRISPR-arrays were identified with the CRISPR Recognition Tool using default parameters (Bland et al. 2007). The clusters of *cas* genes were identified as in Touchon and Rocha (2010) using the classification recently proposed (Makarova et al. 2011). Protospacers were identified using BLASTN to search for similarities between CRISPR spacer sequences and all the phage (831) and plasmid (3,861) genomes available in GenBank (default settings, *e* value  $< 10^{-5}$ ). We retained the matches showing  $\geq 90\%$  of identity and  $\leq 10\%$  difference in sequence length with the query. Prophages were detected with Phage Finder v.2.1 (Fouts 2006), discarding prophages less than 18 kb long. Loci encoding conjugative or mobilizable elements were identified with CONJscan (default parameters) (Guglielmini et al. 2011, 2014). Integrases were searched with the PFAM profile PF00589 for tyrosine recombinases and the pair PF00239 and PF07508 for Serine recombinases. The tyrosine recombinases of integrons were identified from the integrases using a protein profile for a region specific to these proteins that we built based on published data (Cambray et al. 2011). Error-prone polymerases were searched for with PFAM profiles: PF00136 (Pol2 like PolB), PF00817 (Y-Polymerases like UmuC, DinP, or ImuB), PF07733 (DnaE2), and PF00717 (UmuD). The cassettes *imuABC* were searched with the pair of profiles PF00817–PF07733 and *umuCD* using PF00817–PF00717 (Galhardo et al. 2005; Norton et al. 2013). All the protein profiles were searched with hmmer 3.0 with default parameters (Eddy 2011). We removed from further analysis the hits with low *e* values ( $e > 0.001$ ) or those for which the alignment matched less than half of the protein profile.

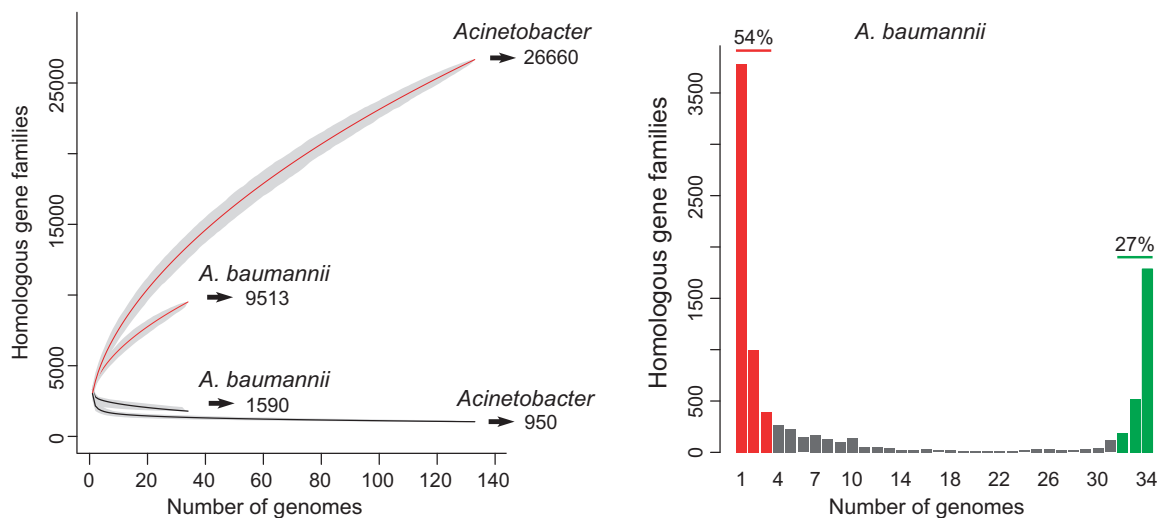
#### Identification of Integration/Deletion Hotspots

The *A. baumannii* core-genome was used to identify and locate large integration/deletion (indel) regions. All regions including more than ten genes between two consecutive core genes of the species were considered as large indel regions. The relative positions of these regions were defined by the order of the core genes in *A. baumannii* AYE. This strain was used as a reference to order *A. baumannii* genes, because it represents the most likely configuration of the chromosome in the ancestor of the species. Regions located between two nonsuccessive core genes, that is, with rearrangements in between them were removed.

## Results and Discussion

### Genetic Diversity of the Genus

We analyzed a panel of 133 genomes of *Acinetobacter* spp. covering the breadth of the known taxonomic diversity of the



**Fig. 1.**—Core- and pan-genomes of the genus and of *A. baumannii* (left) and spectrum of frequencies for *A. baumannii* gene repertoires (right). The pan- and core-genomes were used to perform gene accumulation curves using the statistical software R (R Core Team 2014). These curves describe the number of new genes (pan-genome) and genes in common (core-genome) obtained by adding a new genome to a previous set. The procedure was repeated 1,000 times by randomly modifying the order of integration of genomes in the analysis. The spectrum of frequencies (right) represents the number of genomes where the families of the pan-genome can be found, from 1 for strain-specific genes to 34 for core genes. Red indicates accessory genes and green the genes that are highly persistent in *A. baumannii*.

genus (supplementary table S1, Supplementary Material online). All validly named species were represented by the respective type strains. The provisionally designated genomic species or tentative novel species were represented by strains selected based on polyphasic analysis of multiple strains belonging to the given taxon. Additional strains of *A. baumannii* representing different genotypes were identified based on multilocus sequence typing (MLST) and other typing methods (Diancourt et al. 2010). In the case of clinically relevant species other than *A. baumannii*, one or two additional strains were included to study intraspecies variation. These strains were selected to differ from the type strains as much as possible in terms of relevant genotypic/phenotypic properties and origin (e.g., clinical vs. environmental). Some strains were added to study their taxonomic position. We sequenced 120 genomes with high coverage to which we added the complete sequences of the 13 genomes available from GenBank (see Materials and Methods and supplementary fig. S1, Supplementary Material online). The average genome in the data set had 13 scaffolds for 31 contigs and an average size of 3.87 Mb (range between 2.7 and 4.9 Mb). Genomic Guanine-Cytosine (GC) content showed little variation around the average value of 39.6%. The density of protein-coding sequences was found to be homogeneous between genomes at an average of 94%. To test the completeness of the incompletely assembled genomes, we carried out two tests. First, we merged the lists of essential genes reported in *Escherichia coli* (Baba et al. 2006) and *A. baylyi* (de Bernardinis et al. 2008)

(including “double band” mutants, which may not be essential). A total of 414 of the resulting 533 putatively essential genes had homologs in all *Acinetobacter* genomes (78%). Only 38 of these genes were present in less than 121 and more than 9 genomes (7%), among which only one deemed essential in both *A. baylyi* and *E. coli* (*ligA*, absent in two genomes) (supplementary table S2, Supplementary Material online). Second, we compared the fully and partially assembled *A. baumannii* genomes in terms of genome size and number of genes. The two groups of genomes were indistinguishable on both accounts (both  $P > 0.5$ , Wilcoxon tests), suggesting that not fully assembled genomes miss very few genes. Hence, our collection provides a unique and comprehensive reference data set of high quality genomes of the genus *Acinetobacter*.

To quantify the diversity of the gene repertoires, we computed the set of ubiquitous genes (core-genome) and the set of different homologous gene families (pan-genome) in the genus and in the 34 genomes of *A. baumannii* (fig. 1, supplementary tables S3 and S4, Supplementary Material online). The *Acinetobacter* core-genome contained 950 orthologous protein families corresponding to 37% of the size of the smallest proteome (*A. nectaris* CIP 110549<sup>T</sup>) and more than twice the number of essential genes in *A. baylyi*. The *A. baumannii* core-genome had 1,590 orthologous protein families corresponding to 44% of the size of the smallest proteome of the species (i.e., *A. baumannii* AB307 0294). Gene rarefaction analyses showed that core-genomes vary



**Table 1**

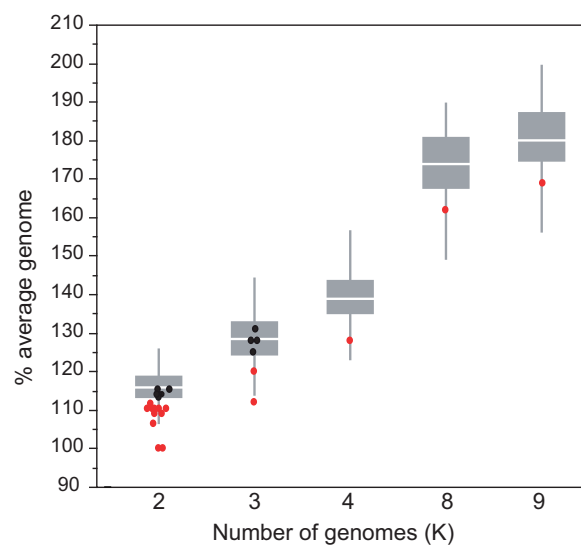
Core and Pan-Genomes for Species of the *Acinetobacter* Genus with At Least Three Sequenced Genomes (see [supplementary table S5, Supplementary Material](#) online, for all species)

| Species                            | #  | Core-Genome |    | Pan-Genome |     |
|------------------------------------|----|-------------|----|------------|-----|
|                                    |    | Size        | %  | Size       | %   |
| <i>Acinetobacter calcoaceticus</i> | 4  | 2,951       | 81 | 4,677      | 128 |
| <i>Acinetobacter indicus</i>       | 3  | 2,340       | 79 | 3,309      | 112 |
| <i>Acinetobacter lwoffii</i>       | 9  | 2,161       | 66 | 5,557      | 169 |
| <i>Acinetobacter parvus</i>        | 8  | 1,810       | 64 | 4,576      | 162 |
| <i>Acinetobacter pittii</i>        | 3  | 2,926       | 81 | 4,282      | 120 |
| <i>Acinetobacter schindleri</i>    | 3  | 2,391       | 76 | 3,929      | 125 |
| <i>Acinetobacter ursingii</i>      | 3  | 2,458       | 72 | 4,353      | 128 |
| <i>Acinetobacter junii</i>         | 3  | 2,293       | 70 | 4,292      | 131 |
| Genomic sp. 13BJ/14TU              | 3  | 2,782       | 73 | 4,839      | 128 |
| Genomic sp. 16                     | 3  | 3,222       | 78 | 5,210      | 125 |
| <i>Acinetobacter baumannii</i>     | 34 | 1,590       | 42 | 10,849     | 288 |

NOTE.—For each species, the number of genomes (#), size of the core- and pan-genome, and percentage of the two relative to the size of the average genome in the clade are indicated.

little with the addition of the last genomes (fig. 1), suggesting that our estimate of the core-genome is robust. Both the pan-genomes of the *Acinetobacter* genus and of *A. baumannii* were very large with, respectively, 26,660 and 9,513 gene families. Gene rarefaction analyses showed that in both cases the addition of new genomes to the analysis still significantly increases the size of the pan-genome. This was confirmed by the spectrum of gene frequencies for the *A. baumannii* pan-genome (fig. 1), which showed that the vast majority of gene families were either encoded in a few genomes (54% in three or less) or in most of them (27% in more than 31 genomes). Over a third of the pan-genome (40%) corresponded to gene families observed in a single genome, that is, strain-specific genes. These analyses confirm that the genus and *A. baumannii* have an extremely large pan-genome. Furthermore, the shape of the accumulation curves showed that we are yet very far from having sampled it enough. Further work will therefore be necessary to characterize the genetic diversity of the species in the genus.

We took advantage of the possibility offered by our data set to compute the genetic variability of all the other validly names species or genomic species in the genus (table 1 for clades with at least three genomes and [supplementary table S5, Supplementary Material](#) online, for all). With the exception of *A. baylyi*, *A. indicus*, and *A. brisouii*, for which the genomes are very similar, all genomes showed large variability of gene repertoires. The comparison of the size of the pan-genomes of these species with the distribution of random samplings of the pan-genome of *A. baumannii* with an equivalent number of genomes showed systematic larger values for the latter (fig. 2,  $P < 0.001$ , Binomial test). Hence, although

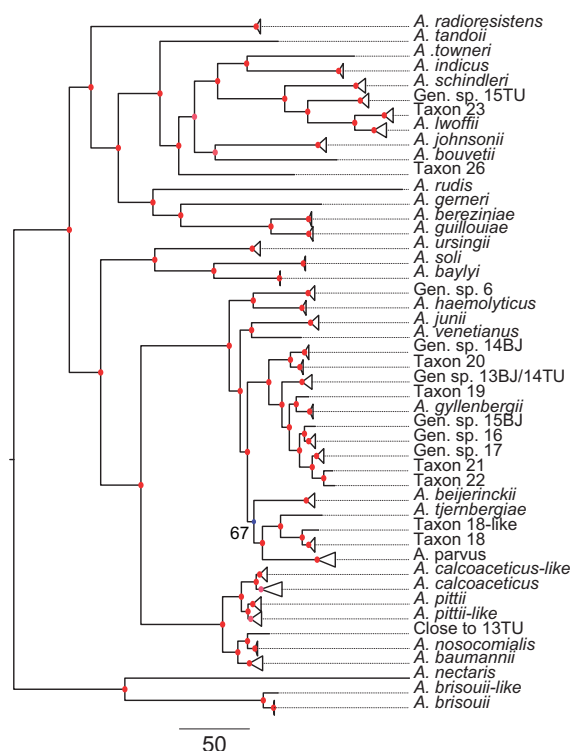


**Fig. 2.**—Comparisons of the pan-genome of *A. baumannii* computed with random samples of different size (boxplots) with those of the other species or genomic species. Each species, except *A. baumannii*, is only represented once, in the graph corresponding to the full number of available genomes for the taxa (e.g., nine genomes for *A. lwoffii*). The boxplots show the distribution of the size of the pan-genome of *A. baumannii* using random samples of  $K$  *A. baumannii* genomes ( $K = \{2, 3, 4, 8, 9\}$  genomes). Black dots correspond to pan-genomes of other species that are within the 25–75 percentiles of the distribution of the pan-genomes of *A. baumannii*, that is, these are pan-genomes approximately the size of *A. baumannii* given the same number of genomes. Red dots correspond to species with pan-genomes smaller than 75% of the *A. baumannii* pan-genomes (see [supplementary table S5, Supplementary Material](#) online, for full data).

we have chosen the genomes outside *A. baumannii* to maximize known biochemical and ecological diversity within these species, these were found to be less diverse than *A. baumannii*.

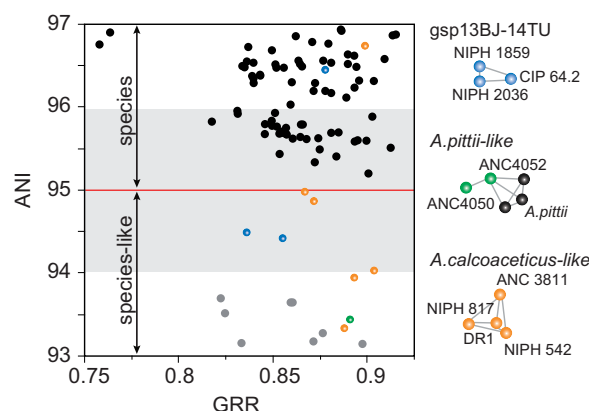
### Phylogeny and Systematics of the Genus

Despite recent progress, the understanding of the *Acinetobacter* evolution is incomplete at the phylogenetic and taxonomic levels. We used the 950 core protein families of the genus to build its phylogeny (see Materials and Methods). The resulting genus phylogenetic tree is extremely well supported from the statistical point of view, showing only one bifurcation with a bootstrap support lower than 95% (67%, see fig. 3). To root this tree and thus infer the order of evolutionary events in the genus, we used two genomes from the two most closely related genera for which complete genomes were available (*Moraxella* and *Psychrobacter*). This tree (topologically very similar to the previous one, see [supplementary fig. S2, Supplementary Material](#) online),



**Fig. 3.**—Phylogeny of the *Acinetobacter* genus based on the alignment of the protein families of the core-genome (see Materials and Methods). Triangles mark groups of taxa that are from the same species or have more than 95% ANI values and therefore might be regarded as coming from the same species. The nodes in red have bootstrap supports higher than 95%. The tree was rooted using two outgroup genomes (see main text).

positions *A. brisouii* and *A. nectaris* as the taxa branching deeper in the genus, that is, the taxa most distantly related with the remaining *Acinetobacter* spp. Relative to the rest of the genus, the small (~3.2 Mb) genomes of *A. brisouii* showed average G+C content (41.5%). The genome of *A. nectaris* was among the smallest in size (2.9 Mb) and lowest in GC content (36.6%). The small size and extreme GC content of the genome and the very long terminal branch of *A. nectaris* in the phylogenetic tree suggest rapid evolution for this species. This feature is typical of bacteria enduring strong ecological niche contractions (Ochman and Moran 2001). Following this split, the phylogeny separates two very large groups of taxa: One including species such as *A. baumannii*, *A. parvus*, and *A. baylyi*; the other including *A. lwoffii*, *Acinetobacter johnsonii*, and *A. guillouiae*. Among these taxa, two are more isolated in the phylogenetic tree. *Acinetobacter radioresistens* is believed to be highly resistant to gamma-ray irradiation and might be the origin of the OXA-23 carbapenem resistance determinant in *A. baumannii* (Poirel et al. 2008; Perichon et al. 2014). It branched deep in

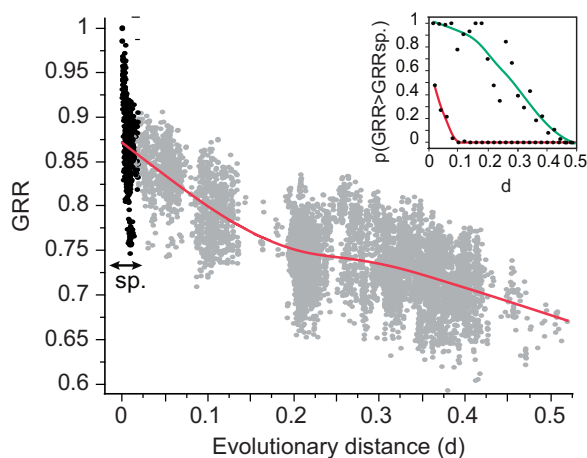


**Fig. 4.**—Analysis of the association between ANI and GRR (see Materials and Methods). The points in black correspond to the clades in triangles in figure 3. The points in gray correspond to comparisons between genomes that are closely related but not of the same species. We highlight three clades where some strains are closely related to the genomic species 13BJ-14TU, *A. pittii*, and *A. calcoaceticus*.

the tree and lacked closely related species (Nishimura et al. 1988; Poirel et al. 2008; Sahl et al. 2013). *Acinetobacter rudis* showed a long branch in the tree, even if its position is very well supported, suggesting higher evolutionary rates of this bacterium (isolated from raw milk and wastewater) (Vaz-Moreira et al. 2011).

To assess the age of the genus, we computed the average protein similarity of positional orthologs of the core-genome between the earlier branching species (*A. brisouii*) and *A. baumannii*. We did not use *A. nectaris* for this analysis because its long external branch would lead to an overestimate of the distances within the genus. The orthologs between *A. brisouii* and *A. baumannii* show an average sequence similarity of 80.1% (interval of confidence [IC], IC95%: 79.5–80.7%). As a matter of comparison, the same analysis between the orthologs of the core-genomes of *E. coli* and *Yersinia pestis*—placed in extreme opposites of the *Enterobacteriaceae* (after removing the fast-evolving *Buchnera* clade) (Williams et al. 2010)—shows an average protein similarity of 80% (IC95%: 79.3–80.7%). Hence, the genus of *Acinetobacter* is very ancient and its last common ancestor was close contemporary of the last common ancestor of *Enterobacteriaceae*. Accurate dating of bacterial genus is impossible given the lack of fossil records. Nevertheless, *Enterobacteriaceae* are thought to have diverged from the *Pasteurellaceae* over 500 Ma (Battistuzzi and Hedges 2009) and *Acinetobacter* might therefore be as ancient. The ancient history of the *Acinetobacter* genus contributes to explain its metabolic and ecological diversity.

The taxonomy of *Acinetobacter* still suffers from unclear taxonomic position and/or confusing nomenclature of some provisional species, high number of unidentifiable environmental strains (Nemec A and Krizova L, unpublished data),



**Fig. 5.**—Analysis of the association between GRR and the phylogenetic distance. Points in black indicate comparisons between pairs of genomes of the same species/genomic species (triangles in fig. 3) and points in gray indicate the other pairs. The red line is a spline fit of the data. The inset shows the relation between the evolutionary distance and the probability that comparing two genomes will result in a GRR value higher than the average within-species GRR (red) and higher than the minimal within-species GRR (green).

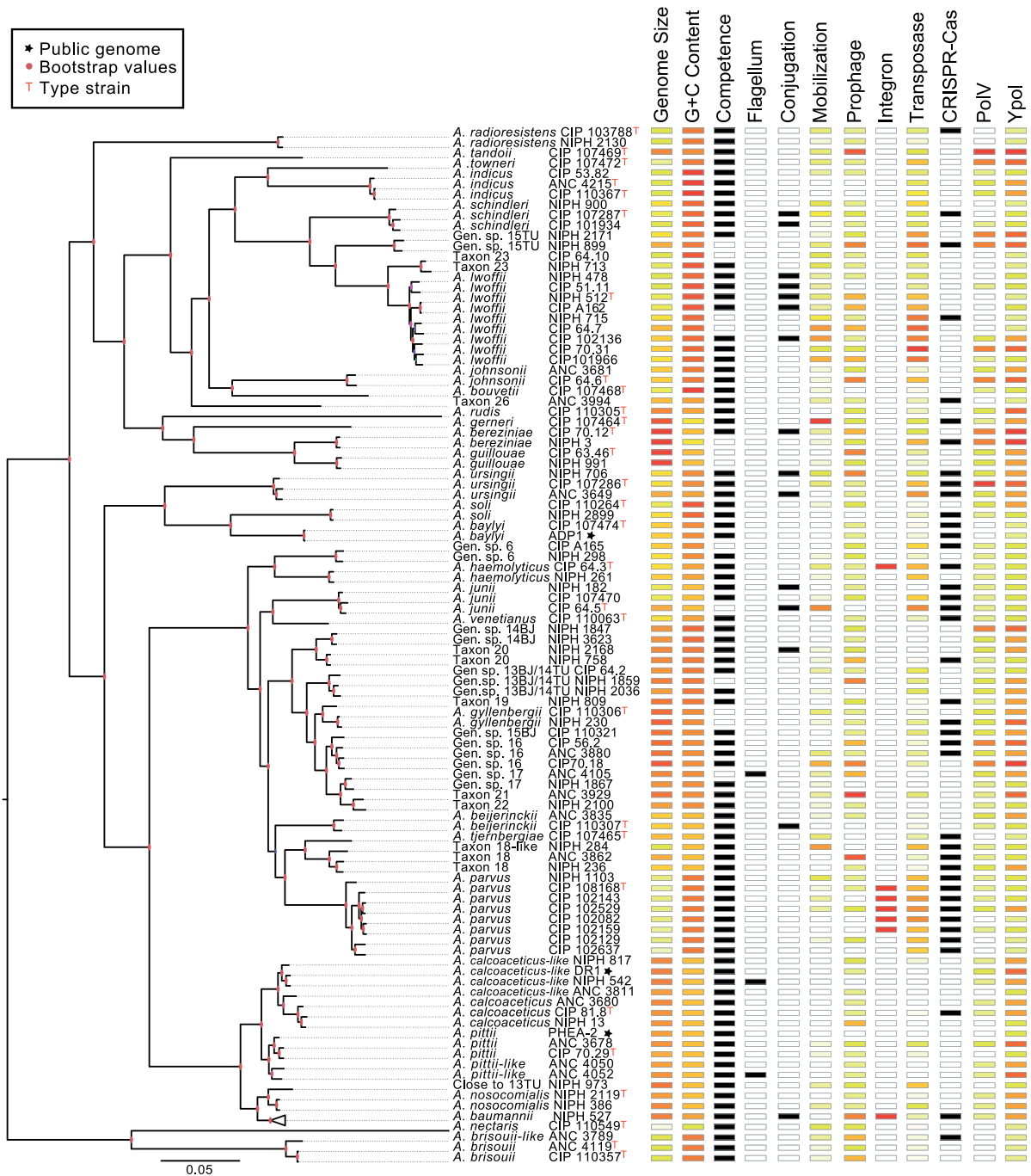
and a number of controversial interpretations of taxonomic data (Nemec et al. 2008, 2011; Vanechoutte et al. 2008). The availability of complete genomes and a robust phylogenetic history allow the identification of these taxonomical problems and the preliminary identification of taxa that might be regarded as good candidates for new species. We therefore put together the information on the core-genome tree and the ANI. Values of ANI between 94% and 96% have been proposed to be a good threshold for the definition of a bacterial species and to replace DDH measurements in preliminary identification of bacterial species from genome data (Konstantinidis et al. 2006; Richter and Rossello-Mora 2009). Most named species and genomic species in this analysis showed intraspecies ANI values higher than 95%. This is in close agreement with the idea that they represent bona fide species (fig. 4). ANI analysis also shed some light on the taxonomical status of several strains that were previously the subject of taxonomic controversies (see values in [supplementary table S1, Supplementary Material](#) online). First, high (>97%) ANI values unambiguously corroborated that *A. grimontii* CIP 107470<sup>T</sup> belongs to *A. junii* and “*Acinetobacter septicus*” ANC 3649 to *A. ursingii* as previously suggested based on DDH data (Nemec et al. 2008; Vanechoutte et al. 2008). On the other hand, strain CIP 64.10 which was believed to be derived from *A. lwoffii* NCTC 5866<sup>T</sup> (Bouvet and Grimont 1986) is clearly distinct from it (ANI of 88.3%). This finding explains previous controversial DDH results for these organisms (Tjernberg and Ursing

1989). Previous studies have also pointed out taxonomic problems with some closely related provisional species, notably among proteolytic and hemolytic strains (Bouvet and Jeanjean 1989). DDH values found by these authors for genomic sp. 15BJ and 16 are in agreement with the observed ANI reference strains of these species (92.6%), suggesting that the genetic distance between these taxa is lower but close to the thresholds underlying species definition. Genomic sp. 13BJ (Bouvet and Jeanjean 1989) and 14TU (Tjernberg and Ursing 1989) have also been considered as a single species based on DDH data, whereas their ANI values (~94.5%) are close to the threshold used to define a species. In such cases, rare in our data set, clear taxonomic conclusions will require analyses of biochemical and genetic data from more comprehensive sets of strains (to be published separately).

We studied the association between phylogenetic distance and the GRR (fig. 5). GRR was defined for each pair of genomes as the number of orthologs present in two genomes divided by the number of genes of the smallest genome (Snel et al. 1999) (see Materials and Methods). It is close to 100% if the gene repertoires are very similar (or one is a subset of the other) and lower otherwise. Consistent with the large pan-genome of most species in the genus, we observed highly variable gene repertoires for genomes within the same species (short phylogenetic distances). The most extreme differences were found when comparing the susceptible *A. baumannii* SDF strain with other strains of the same species. This strain endured a process of genome reduction concomitant with proliferation of insertion sequences (IS) (Vallenet et al. 2008). After its removal from the data set, the lowest within-species GRR (78%) was still found between *A. baumannii* strains (ATCC 17978 and ANC 4097), in line with our previous observation that this species is particularly diverse. As expected, pairs of genomes of the same species tended to have higher GRR than distantly related genomes. Some of the latter had only around 60% GRR. Yet, there were many exceptions to this average trend, and comparisons between distant genomes often showed higher GRR than comparisons between closely related strains of the same species (inset in fig. 5). For example, *A. baumannii* NIPH 146 and *Acinetobacter soli* CIP 110264 were very distant in the phylogenetic tree and have more than 82% GRR, which is more than many within-species comparisons. This shows the importance of sampling bacterial diversity using complete genome sequences and not just using MLST or core-genome-based analyses. In fact, given these patterns, some strains of distantly related *Acinetobacter* species might have more similar phenotypes than strains within the same species.

#### Mechanisms of Genomic Diversification

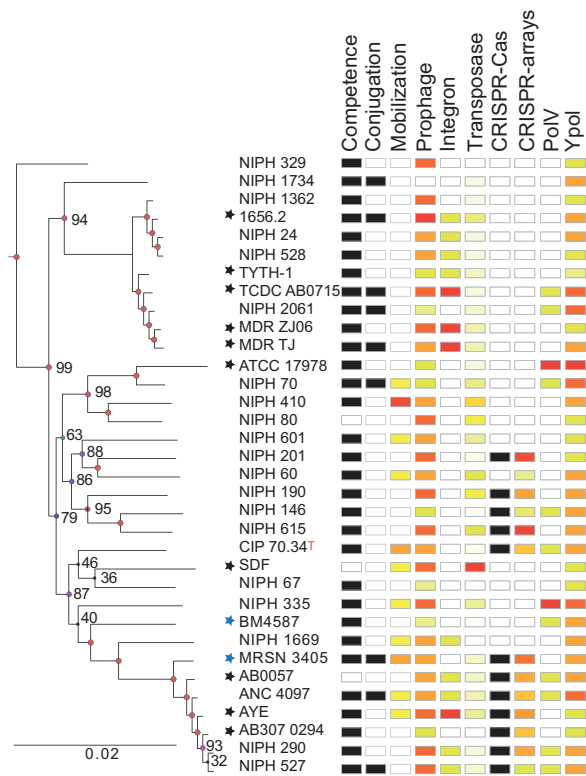
The emergence of antibiotic resistance genes in *Acinetobacter* is facilitated by conjugative elements (Goldstein et al. 1983),



**Fig. 6.**—Distribution of elements potentially related with genetic diversification in the genus. White indicates absence of the trait and black its presence. Genomes with many elements of a given type are indicated in red and those with few elements are indicated in yellow. Intermediate values are indicated in shades of orange. Black asterisks indicate complete genomes from GenBank.

integrons (Ploy et al. 2000; Hujer et al. 2006), IS (Turton et al. 2006), and natural transformation (Wright et al. 2014). Accordingly, we searched for the genes associated with horizontal gene transfer or its control (figs. 6 and 7,

supplementary table S6, Supplementary Material online). We have found 23 proteins matching the profiles of tyrosine recombinases and the specific profiles for integrases of integrons. Interestingly, 17 of these were of the type *intI1*



**Fig. 7.**—Distribution of elements potentially related with genetic diversification in *A. baumannii*. White indicates absence of the trait and black its presence. Genomes with many elements of a given type are indicated in red and those with few elements are indicated in yellow. Intermediate values are indicated in shades of orange. Black asterisks indicate complete genomes from GenBank, blue asterisks indicate genomes sequenced at Pasteur Institute and at Walter Reed.

(associated with a 3'-conserved segment) and were found in 13 strains of *A. baumannii* (four strains had two copies). The abundance of integrons in this species was much higher than would be expected if it were random in the genus ( $P < 0.0001$ ,  $\chi^2$  test). The six hits for integrases of integrons in other species do not match *Int1*, *Int2*, or *Int3* and require further functional study. Analysis of the integron cassette contents in *A. baumannii* showed an *Int0* structure, that is, no cassettes, in one strain (Bissonnette and Roy 1992), whereas the others contained two to five cassettes. We observed an atypical inverted organization where *sul1* was upstream from *int1* in *A. baumannii* 1656-2.

IS have also been implicated in antibiotic resistance. Most notably, *ISAbal* provides a promoter allowing expression of a downstream carbapenemase gene (Turton et al. 2006). IS are very diverse in type and abundance in the genus (from 0 to ~400 per genome), even when comparing closely related strains. The genomes of some species are particularly enriched in IS—*A. lwoffii*, *A. parvus*, *A. junii*, *A. ursingii*—and these

may have contributed to their reduced size. We found at least one copy of *ISAbal* (IS4 family, group IS10) in 36% of all genomes and at least ten copies in 10%, suggesting an important role of this type of IS in the genus.

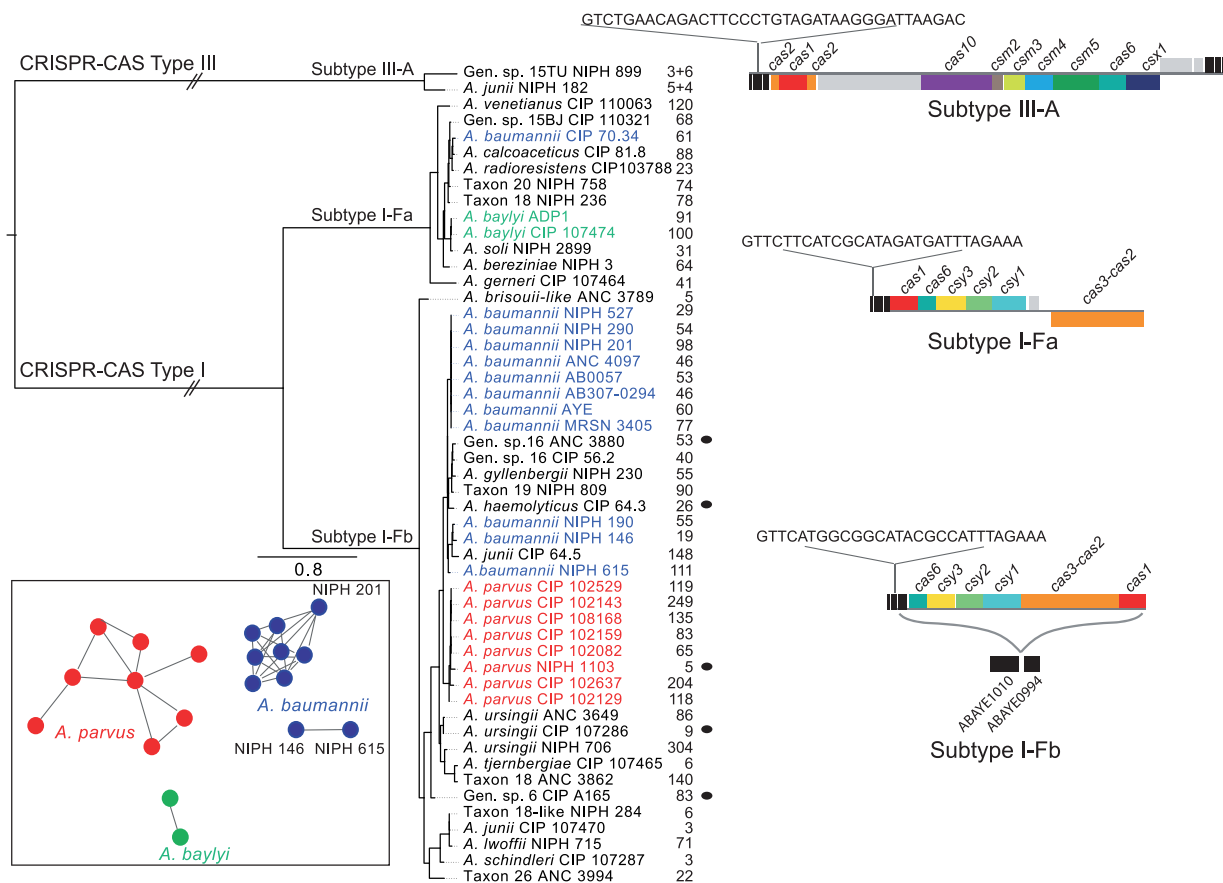
Although some *Acinetobacter* lytic phages have been studied for typing and phage therapy purposes (Bouvet et al. 1990; Ackermann et al. 1994; Shen et al. 2012), there is very little information in the literature on temperate phages infecting *Acinetobacter*. We found 260 prophages of dsDNA phages in the genomes of *Acinetobacter*. We have made a classification of these elements based on the available phages of gamma-proteobacteria (given the lack of information on *Acinetobacter* temperate phages) as in Bobay et al. (2013). More than 98% of the prophages were classified as Caudovirales, among which *Siphoviridae* (41%) and *Myoviridae* (37%) were by far the most abundant. The prophages accounted for a total of 10.4 Mb of genomic sequence in our data set, that is, an average of 2% of the genomes. Only 18 genomes lacked prophages. Hence, most *Acinetobacter* are lysogens. Among the 72 genomes with more than one prophage, *Acinetobacter* ANC 3929 stood out with six prophages (supplementary table S1, Supplementary Material online). Only a minority of these prophages (51) integrated next to tRNAs, as is common in other clades (Williams 2002), even if all identified phage integrases were tyrosine recombinases. The genes of these prophages will be studied in detail in a subsequent work. Nevertheless, as these data suggest an unsuspected role of transduction in driving horizontal gene transfer in the genus we have searched for putatively adaptive traits among prophages. Among other genes, we found one coding for a beta-lactamase in *A. baumannii* ANC 4097 and one coding for a chloramphenicol resistance protein in *A. baumannii* BM4587. Recent findings suggest that phages favor the horizontal transfer of antibiotic resistance determinants (Muniesa et al. 2013; Billard-Pomares et al. 2014). These results suggest that they may indeed contribute to antibiotic resistance in *Acinetobacter*.

Many conjugative elements have been described in association with the spread of antibiotic resistance genes between distant species (Doucet-Populaire et al. 1992; Juhas et al. 2008). We scanned genomes for genes encoding components of the conjugation machinery: Relaxases, coupling proteins, and type 4 secretion systems (T4SS) (Guglielmini et al. 2014). We identified 23 putative conjugation systems in the genus, of which 11 were classified as MPF<sub>F</sub> (family of the F plasmid), 4 MPF<sub>I</sub> (family of the R64 plasmid), and 8 MPF<sub>T</sub> (family of the Ti plasmid). As most genomes in our sample were not in a single contig, and breakpoints were typically found at mobile genetic elements, it is difficult to unambiguously distinguish integrative (ICE) from extrachromosomal (plasmids) conjugative elements. Yet, some information can be retrieved from the abundance of MPF<sub>F</sub> and MPF<sub>I</sub> types. These systems are much more frequently associated with plasmids than with ICE (Guglielmini et al. 2011) and they typically correspond to

narrow host-range mobile genetic elements (Encinas et al. 2014). Interestingly, the long flexible pili of these two families of elements endow them with the ability to engage in conjugation at high frequency in liquid (Bradley 1984). This suggests that liquid media may be relevant for the spread of genetic information in *Acinetobacter*. We identified 211 relaxases distant from any T4SS, mostly MOBQ (140) and MOBP1 (47) (Garcillan-Barcia et al. 2009), which presumably are part of elements mobilizable by conjugation in *trans*. Mobilizable elements are particularly abundant in the genome of *Acinetobacter gementii* (16) and in certain strains of *A. junii* (up to 10 in a genome) and *A. Iwoffii* (up to 9). They are less frequent in *A. baumannii* (between 0 and 3 per genome). Mobilizable plasmids are in general smaller than conjugative plasmids (Smillie et al. 2010) and it has been observed that *A. baumannii* plasmids tend to be small (Gerner-Smidt 1989; Fondi et al. 2010). Our observations suggest that mobilizable small elements may predominate over large

conjugative elements in the genus, a much stronger overrepresentation than in other prokaryotes (Smillie et al. 2010).

Competence for natural transformation has been described in a few strains of *A. baylyi* (Gerischer and Ornston 2001), *A. baumannii* (Harding et al. 2013; Wilharm et al. 2013), and *A. calcoaceticus* (Nielsen et al. 1997) but is thought to be rare in the genus (Towner 2006). In *A. baumannii*, competence and twitching motility are tightly linked and depend on the same type 4-pilus (T4P) (Harding et al. 2013; Wilharm et al. 2013). We searched for the 13 key T4P and competence-associated components and found most of them in all genomes. Only 16 genomes lacked one of the components, of which 13 lacked the *comP* gene that encodes a pilin (supplementary table S7, Supplementary Material online). These absences probably result from recent gene losses, as they are scattered in the phylogenetic tree of the genus. For example, *comP* is missing in only 3 of the 34 *A. baumannii* strains. It was suggested that *comP* was also absent from *A. baumannii*



**Fig. 8.**—Molecular phylogeny of the Cas1 protein across the genus. Phylogenetic tree for the Cas1 proteins was performed using PhyML with the WAG model and a Gamma correction. Cluster of cas genes organization, the most common repeat sequence, and the number of repeat sequences in each genome are indicated on the right part of the figure. Black circles indicate incomplete CRISPR-Cas systems. The left inset shows the genomes sharing spacers, each edge corresponds to the spacer repertoire relatedness (see Materials and Methods). Each color corresponds to a given species.

ATCC 17978 (Smith et al. 2007) but our reannotation procedure revealed a very good hit to the corresponding PFAM domain (profile coverage 99%,  $e$  value  $< 10^{-21}$ ). The very frequent specific deletion of *comP* is intriguing as it is one of the essential components of the natural transformation machinery in *A. baylyi* (Porstendorfer et al. 2000). It is tempting to speculate that the important antigenic potential of pilins (Nassif et al. 1993; Miller et al. 2014) might frequently favor selection for *comP* loss in host-associated bacteria. The conservation of the entire transformation machinery in the vast majority of the genomes suggests that most bacteria in the genus are naturally transformable under certain conditions. Further work will be required to understand the conditions leading to the expression of this trait.

CRISPR, together with associated sequences (*cas* genes and Cas proteins), form the CRISPR-Cas adaptive immune system against transmissible genetic elements such as plasmids and viruses (Sorek et al. 2013; Barrangou and Marraffini 2014). Fifty-one of the genomes encoded CRISPR-Cas systems (fig. 8). A type III-A system was associated with a cluster of 18 genes and a 37-bp repeat sequence and was present in only two distant strains (*A. junii* NIPH 182 and genomic sp. *Acinetobacter* 15TU NIPH 899). In both cases, the CRISPR-arrays located at each side of the *cas* gene clusters were very small (fig. 8). These traits, strain-specific small CRISPR-arrays found in few strains, suggest that the type III-A system has been recently acquired and/or accumulates few spacers. Most CRISPR-Cas systems in the genus were of type I-F. They included 37% of all genomes. The *cas* operon was composed of six to seven genes and the CRISPR repeat was 28 nt long. Based on the phylogenetic tree of Cas1 and the organization of the cluster of *cas* genes, we identified two I-F subtypes (I-Fa and I-Fb) that correspond to the subtypes identified in *A. baumannii* strains ADP and AYE, respectively (Hauck et al. 2012). Interestingly, the I-Fb subtype is probably very ancient as it was integrated at the same genomic locus in very distant species, for example, *A. parvus*, *A. junii*, *A. ursingii*. This CRISPR-Cas system contained some very large CRISPR-arrays in some genomes, up to 304 repeats, with highly conserved repeat sequences and highly variable spacers. This suggests the existence of a strong selective pressure on the activity of CRISPR-Cas systems of this type.

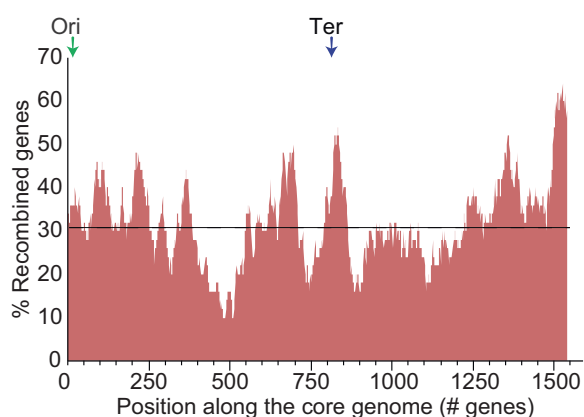
We identified approximately 3,000 spacers in CRISPR-arrays, most of which are unique (80%), that is, they are strain-specific. The vast majority of these spacers (88%) do not match other sequences in the *Acinetobacter* genomes. We found very few genomes having similar spacers and most of these cases corresponded to genomes of the same species among *A. baumannii*, *A. parvus*, or *A. baylyi* (fig. 8). Most spacers matched genes of unknown function, but some matched phage-related functions. As prophages and integrative elements are hard to delimit precisely, we searched for similarity between the spacers and the 831 complete phage and 3,861 complete plasmid genomes available in GenBank.

Only 2% of the spacers showed sequence similarity with elements of this data set. This is not surprising, given the paucity of *Acinetobacter* phages in GenBank. Notwithstanding, we identified ten spacers that match bacteriophages infecting *Acinetobacter* species (Bphi-B1251, AP22 phage), and 47 matching *Acinetobacter* plasmids (e.g., pABTJ1, pNDM-BJ0, pNDM-BJ02). Interestingly, among the few spacers matching known genes, we found homologs of VirB4, VirB5, VirB8, and resolvase proteins, which are all key components of the conjugation machinery. Nevertheless, we found no significant statistical association between the presence of the CRISPR-Cas system (or the number of repeat sequences) and the number of prophages, mobilizable, and conjugative elements in the genus (all  $P > 0.1$  Spearman's rho associations). The same negative results were obtained when the analysis was restricted to *A. baumannii*. The variability of CRISPR spacers might be used to type certain species, but only in combination with other markers, as many strains are devoid of such systems. Some of the CRISPR-arrays we have identified are among the largest ever found among bacteria. CRISPR-Cas systems are therefore likely to have an important role in the genome dynamics of the genus and in particular in controlling the transfer of conjugative elements.

Point mutations also account for the emergence of new traits in *Acinetobacter*, including antibiotic resistance (Yoon et al. 2013). The dynamics of adaptation by point mutations is accelerated when bacteria endure hypermutagenesis, for example following the implication of error-prone DNA polymerases in replicating damaged DNA (Tenailon et al. 2004). The SOS-response of *A. baumannii* does not involve LexA, the typical key regulator of this response (Robinson et al. 2010). Accordingly, we searched and found no ortholog of LexA in the genus. It has been suggested that error-prone polymerases, which have multiple homologs in certain genomes, facilitate the rapid emergence of antibiotic resistance in *Acinetobacter* spp. by stress-induced mutagenesis (Norton et al. 2013). We screened the genomes for homologs of PolB and Y-polymerases and found no homolog of PolB, nor of the *imuABC* operon, which is implicated in damage-induced mutagenesis (Galhardo et al. 2005). In contrast, we identified 345 Y-polymerases in the genus, that is, an average of almost three polymerases per genome (fig. 6). No single genome lacked Y-polymerases and certain harbored up to five copies of the gene. The pair *umuCD* (encoding PolV in *E. coli*) was present in nearly all genomes, often in multiple copies. The multiplicity of genes encoding Y-polymerases in these genomes is intriguing and suggests that they play important roles in *Acinetobacter*, for example, in acquiring tolerance to toxins and antibiotics and/or in their genetic diversification.

#### Origin and Diversification of *A. baumannii*

The large pan-genome of *A. baumannii* showed that this species has highly diverse gene repertoires suggestive of frequent



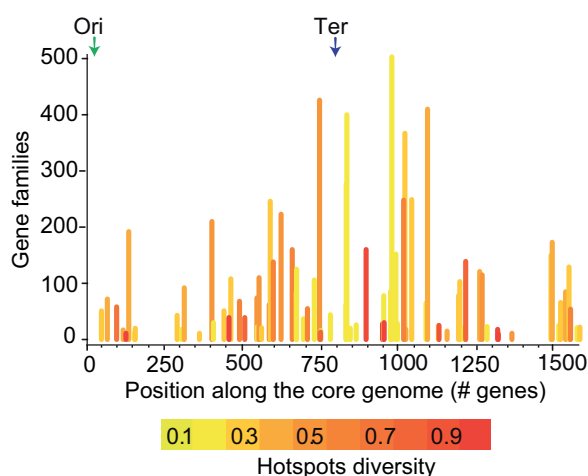
**Fig. 9.**—Distribution of genes of the core-genome of *A. baumannii* presenting significant evidence of recombination using Phi ( $P < 0.05$  after sequential Bonferroni correction) computed in sliding windows of 50 core genes. The dashed line indicates the average.

horizontal gene transfer (fig. 1). Genetic diversification can also result from allelic exchange by homologous recombination in the core-genome. We estimated the impact of this type of recombination in *A. baumannii* with Phi, a conservative and robust method to detect recombination (Bruen et al. 2006). We found that 32% of the core gene families are significantly affected by recombination ( $P < 0.05$ ) (see Materials and Methods). To quantify the number and size of recombination tracts, we concatenated the multiple alignments following the order of the *A. baumannii* ATCC 17978 strain and fetched 688 recombination events significantly highlighted by three procedures (RDP3, CH12, and GENECONV, see Materials and Methods). We were able to precisely delimit the tracts for 526 events of recombination. Their size averaged 2.1 kb (95% of the tracts were between 367 bp and 16 kb long). This size is an underestimate because of the presence of sequences separating core genes and because multiple events of recombination lead to shorter tracts. We also confirmed the presence of recombination using ClonalFrame (Didelot and Falush 2007). This program estimated that recombination contributed to the observed polymorphisms more than mutations (1.37 times). This value is very close to the one observed for MLST data (1.3) (Diancourt et al. 2010). Homologous recombination near the origin of replication was recently associated with the diversification of three outbreak strains of *A. baumannii* (Snitkin et al. 2011). We therefore quantified the distribution of recombination rates along the chromosome of *A. baumannii*. The highest density of recombining genes among the 34 genomes was indeed found close to the origin of replication, but only on the counterclockwise sense (end of the published sequence). Several other regions showed high frequency of recombination whereas others were nearly clonal (fig. 9). These results showed that a large fraction of the genes in *A. baumannii* are significantly affected by recombination,

that rates of recombination vary along the chromosome, and that recombination tracts tend to be small.

*Acinetobacter baumannii* has become a significant clinical problem in the 1970s (Bergogne-Berezin and Towner 1996), but whether this reflects adaptation of a small number of clones to hospital environments or population expansion is not known. The presence of short internal nodes close to the last common ancestor of the species and its large pan-genome have led to suggestions that *A. baumannii* might have endured one wave of population expansion during the diversification of the species and another very recently after the introduction of antibiotics at the hospital (Diancourt et al. 2010; Antunes et al. 2013). The assessment of the hypothesis for a recent population expansion will require a larger sample of closely related genomes. To test the hypothesis of an ancient population expansion, we computed Tajima's  $D$  in sliding windows along the genome of *A. baumannii* (see Materials and Methods) (Tajima 1989). We observed systematically negative values of  $D$  (average  $D = -0.50$ ,  $P < 0.001$ , Wilcoxon signed-rank test). Tajima's  $D$  is affected by recombination (Thornton 2005), but purging the alignments of genes for which Phi identified significant evidence of recombination resulted in even more negative values (average  $D = -0.9$ ,  $P < 0.001$ , same test). Negative  $D$  is consistent with population expansion and/or purifying selection. To separate between these two possibilities, we analyzed separately 4-fold degenerate synonymous ( $D_4$ ) and strictly nonsynonymous ( $D_0$ ) positions (supplementary fig. S3, Supplementary Material online). The two measures are equally affected by sampling biases, recombination, and population expansion. Differences between  $D_4$  and  $D_0$  pinpoint selective processes because nonsynonymous changes are much more deeply imprinted by natural selection than synonymous ones. The  $D_0$  values are significantly lower than those of  $D_4$  (resp. average  $D_0 = -1.5$  and  $D_4 = -0.4$ , difference significant  $P < 0.001$  Wilcoxon signed-rank test) even if both are significantly negative ( $P < 0.001$ ). This suggests that negative values of Tajima's  $D$  are driven by selection against nonsynonymous substitutions, a clear sign of purifying selection. To further test this conclusion, we measured the ratio of nonsynonymous and synonymous substitutions ( $dN/dS$ ). The average within-species  $dN/dS$  was only 0.05 ( $P < 0.001$ ) and even very closely related strains ( $dS < 0.001$ ) showed  $dN/dS$  lower than 0.2 (supplementary fig. S4, Supplementary Material online). This confirms that natural selection purges the vast majority of nonsynonymous mutations in the genome (Rocha et al. 2006). We then computed Fay and Wu  $H_4$  at 4-fold degenerate positions ( $H_4$ ) (Fay and Wu 2000). We found very low  $H_4$  average values ( $-51$ ,  $P < 0.001$ ). Negative  $H_4$  is an indication of selective sweeps or ancient population bottlenecks and negative  $D_4$  suggests population expansion. These results are thus consistent with the hypothesis of a population bottleneck in *A. baumannii* in the early stages of speciation





**FIG. 10.**—Distribution of integration/deletion hotspots along the core-genome of *A. baumannii* using gene orders of *A. baumannii* AYE strain as a reference (see Materials and Methods). The bars represent the number of different gene families in all the genomes found between two consecutive genes of the core-genome. The colors represent the diversity of these gene families, that is, the number of gene families divided by the number of genes found between two consecutive genes of the core-genome. If the number of genes is identical to the number of gene families (1, maximal diversity), then every genome has a different set of genes in the hotspot indicating many different insertions in the region. If the number of families equals the number of genes per genome (close to 1/33, minimal diversity), then most genomes have the same genes in the hotspot. This last scenario typically corresponds to strain-specific large deletions.

with subsequent population expansion under a regime dominated by purifying selection.

Acquisition of resistance often results from the transfer of a mobile element encoding several resistance genes. For example, the AYE multiresistant strain has a genomic island (AbaR1) containing 45 resistance genes including numerous determinants of antibiotic resistance (Fournier et al. 2006). This island was probably acquired in multiple steps of accretion and deletion of genetic material (Sahl et al. 2011). We studied the general patterns of integration of horizontally acquired genes in *A. baumannii* to quantify how many regions in the genome were integration/deletion hotspots. We identified 1,083 regions in the genomes that were flanked by two consecutive core genes and included more than ten genes inserted or deleted (indel) in at least one genome (fig. 10). These loci were not distributed randomly. Instead, the 1,083 regions with indels occurred at the same 78 hotspot regions (5% of all possible loci), that is, they were flanked by the same 78 pairs of core gene families. A third of these loci corresponded to a single indel in one single genome, typically a strain-specific deletion (indicated by light colors and a large number of families with low diversity in fig. 10). Other loci included many different protein families in different genomes. These corresponded to hotspots that endured multiple

integrations/deletions in different lineages. Hotspots tended to be concentrated closer to the terminus of replication and symmetrically distributed around this position. This tendency has previously been observed in other species (Bobay et al. 2013) and might result from a compromise between selection for genome plasticity and organization (Rocha 2004). Intriguingly, some large regions of the chromosome showed no signs of genome plasticity suggesting that they are less plastic (fig. 10). The 78 hotspots contain 5,203 families, that is, 5% of locations in the genome accumulated 66% of the accessory genome. Hence, most genetic diversification takes place at very few loci in the genome. Querying these regions might be an efficient means of typing *Acinetobacter* strains for specific genotypes. Understanding the mechanisms leading to hotspots should enlighten how new genetic information is accommodated in the genome of *A. baumannii*.

## Conclusions

We have proceeded to an extensive characterization of the molecular and evolutionary mechanisms driving the genetic diversification of *Acinetobacter*. Interestingly, we observed that temperate phages are much more abundant than conjugative elements, even though their role as vectors for horizontal transfer has been neglected in the past. Accordingly, we observed the presence of very complex and fast-evolving CRISPR-Cas systems in the genomes of *Acinetobacter*. Population genetic analyses are consistent with the notion that *A. baumannii* arose from an ancient population bottleneck. Nevertheless, this species is extremely diverse in terms of gene repertoires and shows strong effects of natural selection on protein evolution.

Our study sets a solid basis for the understanding of the evolution of the *Acinetobacter* genus. Further work will be necessary to understand how genetic diversification leads to the key features of the genus, notably high metabolic diversity, antibiotic resistance, and virulence. The confrontation between the genetic and the phenotypic data should facilitate predicting how multiple pathogens rise within a genus by virtue of their genetic backgrounds and genetic plasticity.

## Supplementary Material

Supplementary tables S1–S7 and figures S1–S4 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

This work was supported by an European Research Council grant to E.P.C.R. (EVOMOBILOME no. 281605), by grant 13-26693S from the Czech Science Foundation and by grant NT14466-3/2013 of the Internal Grant Agency of the Ministry of Health of the Czech Republic for A.N. and L.K., by an unrestricted grant from Reckitt-Benckiser to E.J.Y. This

project was funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institute of Health, Department of Health and Human Services, under Contract No.:HHSN272200900018C.

## Literature Cited

- Ackermann HW, Brochu G, Emadi Konjin HP. 1994. Classification of *Acinetobacter* phages. *Arch Virol.* 135:345–354.
- Adams MD, et al. 2008. Comparative genome sequence analysis of multi-drug-resistant *Acinetobacter baumannii*. *J Bacteriol.* 190:8053–8064.
- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
- Alvarez-Perez S, Lievens B, Jacquemyn H, Herrera CM. 2013. *Acinetobacter nectaris* sp. nov. and *Acinetobacter boissieri* sp. nov., isolated from floral nectar of wild Mediterranean insect-pollinated plants. *Int J Syst Evol Microbiol.* 63:1532–1539.
- Antunes LC, Imperi F, Carattoli A, Visca P. 2011. Deciphering the multifactorial nature of *Acinetobacter baumannii* pathogenicity. *PLoS One* 6:e22674.
- Antunes LC, Visca P, Towner KJ. 2014. *Acinetobacter baumannii*: evolution of a global pathogen. *Pathog Dis.* 71:292–301.
- Baba T, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol.* 2: 0008.
- Barbe V, et al. 2004. Unique features revealed by the genome sequence of *Acinetobacter* sp. ADP1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res.* 32:5766–5779.
- Barrangou R, Marraffini LA. 2014. CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Mol Cell.* 54:234–244.
- Battistuzzi FU, Hedges SB. 2009. Eubacteria. In: Hedges SB, Kumar S, editors. *The timetree of life*. New York: Oxford University Press. p. 106–115.
- Bergogne-Berezin E, Towner KJ. 1996. *Acinetobacter* spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clin Microbiol Rev.* 9:148–165.
- Billard-Pomares T, et al. 2014. Characterization of a P1-like bacteriophage encoding an SHV-2 extended-spectrum beta-lactamase from an *Escherichia coli* strain. *Antimicrob Agents Chemother.* 58: 6550–6557.
- Bissonnette L, Roy PH. 1992. Characterization of InO of *Pseudomonas aeruginosa* plasmid pV51, an ancestor of integrons of multiresistance plasmids and transposons of gram-negative bacteria. *J Bacteriol.* 174: 1248–1257.
- Bland C, et al. 2007. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209.
- Bobay LM, Rocha EP, Touchon M. 2013. The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol.* 30:737–751.
- Bouvet PJ, Jeanjean S. 1989. Delineation of new proteolytic genomic species in the genus *Acinetobacter*. *Res Microbiol.* 140:291–299.
- Bouvet PJ, Jeanjean S, Vieu JF, Dijkshoorn L. 1990. Species, biotype, and bacteriophage type determinations compared with cell envelope protein profiles for typing *Acinetobacter* strains. *J Clin Microbiol.* 28: 170–176.
- Bouvet PJM, Grimont PAD. 1986. Taxonomy of the genus *Acinetobacter* with the recognition of *Acinetobacter baumannii* sp-nov, *Acinetobacter haemolyticus* sp-nov, *Acinetobacter johnsonii* sp-nov, and *Acinetobacter junii* sp-nov and emended descriptions of *Acinetobacter calcoaceticus* and *Acinetobacter lwoffii*. *Int J Syst Bacteriol.* 36:228–240.
- Bradley DE. 1984. Characteristics and function of thick and thin conjugative pili determined by transfer-derepressed plasmids of incompatibility groups I1, I2, I5, B, K and Z. *J Gen Microbiol.* 130: 1489–1502.
- Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172: 2665–2681.
- Cambray G, et al. 2011. Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mob DNA.* 2:6.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.
- Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ. 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol.* 12:302.
- Crisuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol.* 10:210.
- Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci.* 23:324–328.
- de Berardinis V, et al. 2008. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Mol Syst Biol.* 4:174.
- Diancourt L, Passet V, Nemeč A, Dijkshoorn L, Brisse S. 2010. The population structure of *Acinetobacter baumannii*: expanding multiresistant clones from an ancestral susceptible genetic pool. *PLoS One* 5:e10034.
- Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175:1251–1266.
- Dijkshoorn L, Nemeč A, Seifert H. 2007. An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nat Rev Microbiol.* 5: 939–951.
- Doucet-Populaire F, Trieu-Cuot P, Andreumont A, Courvalin P. 1992. Conjugal transfer of plasmid DNA from *Enterococcus faecalis* to *Escherichia coli* in digestive tracts of gnotobiotic mice. *Antimicrob Agents Chemother.* 36:502–504.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7: e1002195.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Encinas D, et al. 2014. Plasmid conjugation from Proteobacteria as evidence for the origin of xenologous genes in Cyanobacteria. *J Bacteriol.* 196:1551–1559.
- Eveillard M, Kempf M, Belmonte O, Pailhories H, Joly-Guillou ML. 2013. Reservoirs of *Acinetobacter baumannii* outside the hospital and potential involvement in emerging human community-acquired infections. *Int J Infect Dis.* 17:e802–e805.
- Falagas ME, Karveli EA, Kelesidis I, Kelesidis T. 2007. Community-acquired *Acinetobacter* infections. *Eur J Clin Microbiol Infect Dis.* 26:857–868.
- Farrugia DN, et al. 2013. The complete genome and phenome of a community-acquired *Acinetobacter baumannii*. *PLoS One* 8:e58628.
- Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Fondi M, et al. 2010. Exploring the evolutionary dynamics of plasmids: the *Acinetobacter* pan-plasmidome. *BMC Evol Biol.* 10:59.
- Fondi M, et al. 2013. The genome sequence of the hydrocarbon-degrading *Acinetobacter venetianus* VE-C3. *Res Microbiol.* 164: 439–449.
- Fournier PE, et al. 2006. Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet.* 2:e7.
- Fouts DE. 2006. Phage\_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* 34:5839–5851.
- Galhardo RS, Rocha RP, Marques MV, Menck CF. 2005. An SOS-regulated operon involved in damage-inducible mutagenesis in *Caulobacter crescentus*. *Nucleic Acids Res.* 33:2603–2614.

- Garcillan-Barcia MP, Francia MV, de la Cruz F. 2009. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol Rev.* 33:657–687.
- Gascuel O, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.
- Gerischer U, Ornston LN. 2001. Dependence of linkage of alleles on their physical distance in natural transformation of *Acinetobacter* sp. strain ADP1. *Arch Microbiol.* 176:465–469.
- Gerner-Smidt P. 1989. Frequency of plasmids in strains of *Acinetobacter calcoaceticus*. *J Hosp Infect.* 14:23–28.
- Goldstein FV, et al. 1983. Transferable plasmid-mediated antibiotic resistance in *Acinetobacter*. *Plasmid* 10:138–147.
- Guglielmini J, et al. 2014. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* 42:5715–5727.
- Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP. 2011. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* 7: e1002222.
- Harding CM, et al. 2013. *Acinetobacter baumannii* strain M2 produces type IV pili which play a role in natural transformation and twitching motility but not surface-associated motility. *MBio* 4: e00360-00313.
- Harris SR, et al. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 327:469–474.
- Hauck Y, et al. 2012. Diversity of *Acinetobacter baumannii* in four French military hospitals, as assessed by multiple locus variable number of tandem repeats analysis. *PLoS One* 7:e44597.
- Hujer KM, et al. 2006. Analysis of antibiotic resistance genes in multidrug-resistant *Acinetobacter* sp. isolates from military and civilian patients treated at the Walter Reed Army Medical Center. *Antimicrob Agents Chemother.* 50:4114–4123.
- Imperi F, et al. 2011. The genomics of *Acinetobacter baumannii*: insights into genome plasticity, antimicrobial resistance and pathogenicity. *IJMB life* 63:1068–1074.
- Juhas M, Crook DW, Hood DW. 2008. Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cell Microbiol.* 10: 2377–2386.
- Karah N, et al. 2011. Species identification and molecular characterization of *Acinetobacter* spp. blood culture isolates from Norway. *J Antimicrob Chemother.* 66:738–744.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.
- Kennemann L, et al. 2011. *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A.* 108:5033–5038.
- Konstantinidis KT, Ramette A, Tiedje JM. 2006. The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci.* 361: 1929–1940.
- Krawczyk B, Lewandowski K, Kur J. 2002. Comparative studies of the *Acinetobacter* genus and the species identification method based on the *recA* sequences. *Mol Cell Probes.* 16:1–11.
- Krizova L, Maixnerova M, Sedo O, Nemeč A. 2014. *Acinetobacter bohemicus* sp. nov. widespread in natural soil and water ecosystems in the Czech Republic. *Syst Appl Microbiol.* 37:467–473.
- La Scola B, Gundi VA, Khamis A, Raoult D. 2006. Sequencing of the *rpoB* gene and flanking spacers for molecular identification of *Acinetobacter* species. *J Clin Microbiol.* 44:827–832.
- Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25:1451–1452.
- Makarova KS, et al. 2011. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol.* 9:467–477.
- Martin DP, et al. 2010. RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26:2462–2463.
- Mather AE, et al. 2013. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. *Science* 341: 1514–1517.
- McGann P, et al. 2014. Amplification of aminoglycoside resistance gene *aphA1* in *Acinetobacter baumannii* results in tobramycin therapy failure. *MBio* 5:e00915.
- Metzgar D, et al. 2004. *Acinetobacter* sp. ADP1: an ideal model organism for genetic analysis and genome engineering. *Nucleic Acids Res.* 32: 5780–5790.
- Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116.
- Miller F, et al. 2014. The hypervariable region of meningococcal major pilin PilE controls the host cell response via antigenic variation. *MBio* 5: e01024-01013.
- Muniesa M, Colomer-Lluch M, Jofre J. 2013. Potential impact of environmental bacteriophages in spreading antibiotic resistance genes. *Future Microbiol.* 8:739–751.
- Nassif X, et al. 1993. Antigenic variation of pilin regulates adhesion of *Neisseria meningitidis* to human epithelial cells. *Mol Microbiol.* 8: 719–725.
- Nemeč A, et al. 2001. *Acinetobacter ursingii* sp. nov. and *Acinetobacter schindleri* sp. nov., isolated from human clinical specimens. *Int J Syst Evol Microbiol.* 51:1891–1899.
- Nemeč A, et al. 2003. *Acinetobacter parvus* sp. nov., a small-colony-forming species isolated from human clinical specimens. *Int J Syst Evol Microbiol.* 53:1563–1567.
- Nemeč A, et al. 2009. *Acinetobacter beijerinckii* sp. nov. and *Acinetobacter gyllenbergii* sp. nov., haemolytic organisms isolated from humans. *Int J Syst Evol Microbiol.* 59:118–124.
- Nemeč A, et al. 2010. *Acinetobacter bereziniae* sp. nov. and *Acinetobacter guillouiae* sp. nov., to accommodate *Acinetobacter* genomic species 10 and 11, respectively. *Int J Syst Evol Microbiol.* 60:896–903.
- Nemeč A, et al. 2011. Genotypic and phenotypic characterization of the *Acinetobacter calcoaceticus-Acinetobacter baumannii* complex with the proposal of *Acinetobacter pittii* sp. nov. (formerly *Acinetobacter* genomic species 3) and *Acinetobacter nosocomialis* sp. nov. (formerly *Acinetobacter* genomic species 13TU). *Res Microbiol.* 162:393–404.
- Nemeč A, Musilek M, Vanechoute M, Falsen E, Dijkshoorn L. 2008. Lack of evidence for “*Acinetobacter septicus*” as a species different from *Acinetobacter ursingii*? *J Clin Microbiol.* 46:2826–2827; author reply: 2827.
- Nielsen KM, Bones AM, Van Elsas JD. 1997. Induced natural transformation of *Acinetobacter calcoaceticus* in soil microcosms. *Appl Environ Microbiol.* 63:3972–3977.
- Nishimura Y, Ino T, Iizuka H. 1988. *Acinetobacter radioresistens* sp. nov. isolated from cotton and soil. *Int J Syst Bacteriol.* 38:209–211.
- Norton MD, Spilki AJ, Godoy VG. 2013. Antibiotic resistance acquired through a DNA damage-inducible response in *Acinetobacter baumannii*. *J Bacteriol.* 195:1335–1345.
- Ochman H, Moran NA. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292:1096–1099.
- Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
- Peleg AY, et al. 2012. The success of *Acinetobacter* species; genetic, metabolic and virulence attributes. *PLoS One* 7:e46984.
- Peleg AY, Seifert H, Paterson DL. 2008. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clin Microbiol Rev.* 21:538–582.
- Perichon B, et al. 2014. Identification of 50 class D beta-lactamases and 65 *Acinetobacter*-derived cephalosporinases in *Acinetobacter* spp. *Antimicrob Agents Chemother.* 58:936–949.
- Ploy MC, Denis F, Courvalin P, Lambert T. 2000. Molecular characterization of integrons in *Acinetobacter baumannii*: description of

- a hybrid class 2 integron. *Antimicrob Agents Chemother.* 44: 2684–2688.
- Poirel L, Figueiredo S, Cattoir V, Carattoli A, Nordmann P. 2008. *Acinetobacter radioresistens* as a silent source of carbapenem resistance for *Acinetobacter* spp. *Antimicrob Agents Chemother.* 52:1252–1256.
- Porstendorfer D, Gohl O, Mayer F, Averhoff B. 2000. ComP, a pilin-like protein essential for natural competence in *Acinetobacter* sp. Strain BD413: regulation, modification, and cellular localization. *J Bacteriol.* 182:3673–3680.
- Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 26:1641–1650.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35:D61–D65.
- R Core Team. 2014. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
- Rainey FA, Lang E, Stackebrandt E. 1994. The phylogenetic structure of the genus *Acinetobacter*. *FEMS Microbiol Lett.* 124:349–353.
- Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A.* 106: 19126–19131.
- Robinson A, et al. 2010. Essential biological processes of an emerging pathogen: DNA replication, transcription, and cell division in *Acinetobacter* spp. *Microbiol Mol Biol Rev.* 74:273–297.
- Rocha EPC. 2004. Order and disorder in bacterial genomes. *Curr Opin Microbiol.* 7:519–527.
- Rocha EPC. 2006. Inference and analysis of the relative stability of bacterial chromosomes. *Mol Biol Evol.* 23:513–522.
- Rocha EPC, et al. 2006. Comparisons of dN/dS are time-dependent for closely related bacterial genomes. *J Theor Biol.* 239:226–235.
- Rodriguez-Bano J, et al. 2006. Nosocomial bacteremia due to an as yet unclassified *Acinetobacter* genomic species 17-like strain. *J Clin Microbiol.* 44:1587–1589.
- Sahl JW, et al. 2011. Genomic comparison of multi-drug resistant invasive and colonizing *Acinetobacter baumannii* isolated from diverse human body sites reveals genomic plasticity. *BMC Genomics* 12:291.
- Sahl JW, et al. 2013. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*. *PLoS One* 8:e54287.
- Shen GH, et al. 2012. Isolation and characterization of phikm18p, a novel lytic phage with therapeutic potential against extensively drug resistant *Acinetobacter baumannii*. *PLoS One* 7:e46537.
- Smet A, et al. 2012. OXA-23-producing *Acinetobacter* species from horses: a public health hazard? *J Antimicrob Chemother.* 67:3009–3010.
- Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. 2010. Mobility of plasmids. *Microbiol Mol Biol Rev.* 74:434–452.
- Smith MG, et al. 2007. New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes Dev.* 21:601–614.
- Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. *Nat Genet.* 21:108–110.
- Snitkin ES, et al. 2011. Genome-wide recombination drives diversification of epidemic strains of *Acinetobacter baumannii*. *Proc Natl Acad Sci U S A.* 108:13758–13763.
- Sorek R, Lawrence CM, Wiedenheft B. 2013. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu Rev Biochem.* 82: 237–266.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Tenaillon O, Denamur E, Matic I. 2004. Evolutionary significance of stress-induced mutagenesis in bacteria. *Trends Microbiol.* 12: 264–270.
- Thornton K. 2005. Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. *Genetics* 171: 2143–2148.
- Tjernberg I, Ursing J. 1989. Clinical strains of *Acinetobacter* classified by DNA-DNA hybridization. *APMIS* 97:595–605.
- Touchon M, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* 5: e1000344.
- Touchon M, Rocha EP. 2010. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One* 5: e11126.
- Towner K. 2006. The Genus *Acinetobacter*. In: Dworkin J, Falkow S, Rosenberg E, Schleifer KH, Stackebrandt E, editors. *The prokaryotes*. New York: Springer. p. 746–758.
- Turton JF, et al. 2006. The role of ISAbal in expression of OXA carbapenemase genes in *Acinetobacter baumannii*. *FEMS Microbiol Lett.* 258: 72–77.
- Turton JF, Shah J, Ozongwu C, Pike R. 2010. Incidence of *Acinetobacter* species other than *A. baumannii* among clinical isolates of *Acinetobacter*: evidence for emerging species. *J Clin Microbiol.* 48: 1445–1449.
- Vallenet D, et al. 2008. Comparative analysis of *Acinetobacter* species: three genomes for three lifestyles. *PLoS One* 3:e1805.
- Vanechoutte M, et al. 2008. Reclassification of *Acinetobacter grimontii* Carr et al. 2003 as a later synonym of *Acinetobacter junii* Bouvet and Grimont 1986. *Int J Syst Evol Microbiol.* 58: 937–940.
- Vaz-Moreira I, et al. 2011. *Acinetobacter rudis* sp. nov., isolated from raw milk and raw wastewater. *Int J Syst Evol Microbiol.* 61: 2837–2843.
- Wilharm G, Piesker J, Laue M, Skiebe E. 2013. DNA uptake by the nosocomial pathogen *Acinetobacter baumannii* occurs during movement along wet surfaces. *J Bacteriol.* 195:4146–4153.
- Williams KP. 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* 30:866–875.
- Williams KP, et al. 2010. Phylogeny of gammaproteobacteria. *J Bacteriol.* 192:2305–2314.
- Wright MS, et al. 2014. New insights into dissemination and variation of the health care-associated pathogen *Acinetobacter baumannii* from genomic analysis. *MBio* 5:e00963–00913.
- Yamahira K, et al. 2008. *Acinetobacter* sp. strain Ths, a novel psychrotolerant and alkalitolerant bacterium that utilizes hydrocarbon. *Extremophiles* 12:729–734.
- Yamamoto S, Bouvet PJ, Harayama S. 1999. Phylogenetic structures of the genus *Acinetobacter* based on gyrB sequences: comparison with the grouping by DNA-DNA hybridization. *Int J Syst Bacteriol.* 49(Pt 1): 87–95.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
- Yoon EJ, Courvalin P, Grillot-Courvalin C. 2013. RND-type efflux pumps in multidrug-resistant clinical isolates of *Acinetobacter baumannii*: major role for AdeABC overexpression and AdeRS mutations. *Antimicrob Agents Chemother.* 57:2989–2995.

Associate editor: John McCutcheon

## **Co-authored Manuscript 2: Abby et al, 2016**

In this work of Abby et al, I contributed by building the Macsyfinder model for the detection of the type 4 secretion system.

# SCIENTIFIC REPORTS

OPEN

## Identification of protein secretion systems in bacterial genomes

Sophie S. Abby<sup>1,2,†</sup>, Jean Cury<sup>1,2</sup>, Julien Guglielmini<sup>1,2,‡</sup>, Bertrand Néron<sup>3</sup>, Marie Touchon<sup>1,2</sup> & Eduardo P. C. Rocha<sup>1,2</sup>

Received: 15 October 2015

Accepted: 24 February 2016

Published: 16 March 2016

**Bacteria with two cell membranes (diderms) have evolved complex systems for protein secretion. These systems were extensively studied in some model bacteria, but the characterisation of their diversity has lagged behind due to lack of standard annotation tools. We built online and standalone computational tools to accurately predict protein secretion systems and related appendages in bacteria with LPS-containing outer membranes. They consist of models describing the systems' components and genetic organization to be used with MacSyFinder to search for T1SS-T6SS, T9SS, flagella, Type IV pili and Tad pili. We identified ~10,000 candidate systems in bacterial genomes, where T1SS and T5SS were by far the most abundant and widespread. All these data are made available in a public database. The recently described T6SS<sup>iii</sup> and T9SS were restricted to Bacteroidetes, and T6SS<sup>ii</sup> to *Francisella*. The T2SS, T3SS, and T4SS were frequently encoded in single-copy in one locus, whereas most T1SS were encoded in two loci. The secretion systems of diderm Firmicutes were similar to those found in other diderms. Novel systems may remain to be discovered, since some clades of environmental bacteria lacked all known protein secretion systems. Our models can be fully customized, which should facilitate the identification of novel systems.**

Proteins secreted by bacteria are involved in many important tasks such as detoxification, antibiotic resistance, and scavenging<sup>1</sup>. Secreted proteins also have key roles in both intra- and inter-specific antagonistic and mutualistic interactions<sup>2,3</sup>. For example, they account for many of the virulence factors of pathogens<sup>4,5</sup>. Bacteria with a Lipopolysaccharide-containing outer-membrane (abbreviated “diderm-LPS” in this article) require specific protein secretion systems. Six types of secretion systems, numbered Type I secretion system (T1SS) to Type VI secretion system (T6SS), were well characterised by numerous experimental studies (for some general reviews see<sup>6–8</sup>). The Type IX secretion system (T9SS or PorSS) was more recently uncovered in Bacteroidetes<sup>9,10</sup>. In this study, we focused on these diderm-LPS protein secretion systems. A few other systems have been described in diderm-LPS, such as the chaperone-usher pathway, sometimes named Type VII secretion system (T7SS), and the Type VIII secretion system (T8SS). They were not included in this study because they are only involved, respectively, in the export of type I pili and curli<sup>11</sup>. The ESAT-6 secretion system (ESX) system of *Mycobacteria*, named T7SS by some authors<sup>12</sup>, was also excluded from the analysis because it is absent from diderm-LPS bacteria.

The important role of secreted proteins has spurred interest in the production of ontologies and computational methods to categorise<sup>13</sup> and identify them (Table 1). These are difficult tasks. Firstly, protein secretion systems are large machineries with many different components, some of which are accessory and some interchangeable. Secondly, many of their key components are homologous between systems, which complicates their discrimination. For example, T2SS, T4SS and T6SS include distinct but homologous NTPases<sup>14</sup>. Some bacterial appendages require their own secretion systems to translocate their extracellular components<sup>15,16</sup>, and these are sometimes partly homologous to classical secretion systems. For example, several components of the Type IV pilus (T4P) and the Tight adherence (Tad) pilus are homologous to components of the T2SS from *Klebsiella oxytoca*<sup>17</sup>. Thirdly, the sequences of secreted proteins, including extracellular components of the secretion systems, evolve rapidly, thereby complicating the identification of homology by sequence similarity<sup>18</sup>. Fourthly, loci encoding secretion systems are frequently horizontally transferred and lost<sup>19,20</sup>, leading to the presence of partial (often inactive) systems in genomes<sup>21</sup>. Finally, experimental studies have focused on a small number of occurrences of each type

<sup>1</sup>Institut Pasteur, Microbial Evolutionary Genomics, Paris, 75015, France. <sup>2</sup>CNRS, UMR3525, Paris, 75015, France. <sup>3</sup>Institut Pasteur, C3BI, CIB, Paris, 75015, France. <sup>†</sup>Present address: Division of Archaea Biology and Ecogenomics, Department of Ecogenomics and Systems Biology, University of Vienna, A-1090 Vienna, Austria. <sup>‡</sup>Present address: Bioinformatics and Biostatistics HUB, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), Institut Pasteur, Paris, 75015, France. Correspondence and requests for materials should be addressed to S.S.A. (email: sophie.abby.univ@gmail.com)

| Name       | System   | Web | App | C (method)                          | CC  | S   | URL and Reference   |
|------------|--|-----|-----|-------------------------------------|-----|-----|---|
| AtlasT4SS  | T4   | Yes | No  | Yes (Blast)                         | No  | No  | <a href="http://www.t4ss.lncc.br">http://www.t4ss.lncc.br</a> <sup>119</sup>  |
| CONJscan   | T4   | Yes | No  | Yes (HMM)                           | No  | No  | <a href="http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::CONJscan-T4SSscan">http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::CONJscan-T4SSscan</a> <sup>67</sup> |
| SecReT4    | T4   | Yes | No  | Yes (Blast or HMM)                  | Yes | No  | <a href="http://db-mml.sjtu.edu.cn/SecReT4">http://db-mml.sjtu.edu.cn/SecReT4</a> <sup>120</sup>  |
| SecReT6    | T6 <sup>i-ii</sup>   | Yes | No  | Yes (Blast or HMM)                  | Yes | No  | <a href="http://db-mml.sjtu.edu.cn/SecReT6">http://db-mml.sjtu.edu.cn/SecReT6</a> <sup>121</sup>  |
| SSPred     | T1, T2, T3, T4   | Yes | No  | Yes (other: amino acid composition) | No  | No  | <a href="http://www.bioinformatics.org/sspred">http://www.bioinformatics.org/sspred</a> <sup>122</sup>  |
| T346Hunter | T3, T4, T6 <sup>i</sup>  | Yes | No  | Yes (Blast and HMM)                 | Yes | No  | <a href="http://bacterial-virulence-factors.cbcp.upm.es/T346Hunter">http://bacterial-virulence-factors.cbcp.upm.es/T346Hunter</a> <sup>43</sup>                     |
| T3DB       | T3   | Yes | No  | Yes (Blast)                         | No  | No  | <a href="http://biocomputer.bio.cuhk.edu.hk/T3DB/browse">http://biocomputer.bio.cuhk.edu.hk/T3DB/browse</a> <sup>123</sup>  |
| T3SSscan   | T3   | Yes | No  | Yes (HMM)                           | No  | No  | <a href="http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::T3SSscan-FLAGscan">http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::T3SSscan-FLAGscan</a> <sup>24</sup> |
| TXSScan    | T1, T2, T3, T4, T5 (a, b, c), T6 <sup>i-iii</sup> , T9, T4P, Tad, flagellum. | Yes | Yes | Yes (HMM)                           | Yes | Yes | <a href="http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::txsscan">http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::txsscan</a> This work                         |

**Table 1. Public webservers (Web) and downloadable applications (App) to identify components (C), clusters of components (CC), or complete (eventually scattered, S) bacterial protein secretion systems.**

of system, complicating the assessment of their genetic diversity. On the other hand, secretion systems are often encoded in one or a few neighbouring operons. This information can facilitate the identification of genes encoding secretion systems in genome data<sup>22,23</sup>.

Several programs were previously made available to identify components of some, but not all, protein secretion systems (Table 1). These programs are very useful to the biologist interested in browsing the known systems or in annotating a small set of sequences. However, they are web-based, and thus poorly adapted for the analysis of very large datasets. Few of these programs categorise systems as complete or incomplete, and none allows the definition of these parameters. These programs do not identify systems scattered in the chromosome, they only predict components or in some case clusters of components. This limits the detection power, because the ability to re-define the components and genetic organisation of secretion systems facilitates the search for their distant variants<sup>24</sup>.

We have used the vast body of knowledge accumulated from experimental studies of model protein secretion systems to build computational models describing their composition and genetic organization. The models can be plugged in MacSyFinder<sup>25</sup> to predict protein secretion systems using the standalone application. The pre-defined models can also be used on the webservice version available at <http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::txsscan>. The results can be visualized with MacSyView<sup>25</sup>. In the standalone application, the users can easily modify the models to change the composition and genetic organisation of the secretion systems. Some of these parameters can also be modified in the webservice version. The accuracy of the models was quantified against an independent dataset of experimentally validated systems. Importantly, we provide models to search for an unparalleled number of protein secretion systems (and some partly homologous systems): T1SS, T2SS (Tad and T4P), T3SS (flagellum), T4SS (conjugation system), T5SS, T6SS<sup>i-iii</sup>, and T9SS. We used the models to search for protein secretion systems in a large panel of bacterial genomes. Previous surveys, mostly dating from a time when few genomes were available, analysed the distribution of some specific protein secretion systems in genomes or metagenomes<sup>17,20,24,26-31</sup>. Due to space limitations, we will not attempt at re-assessing all these works. Instead, we describe our models, show their accuracy, and use them to provide a broad view of the distribution of the different protein secretion systems.

## Results and Discussion

**Overview of the approach.** We defined 22 customisable models for the protein secretion systems and related appendages (File S1, Figs 1–6, Figs S1–S5). This was done in four steps (see Materials and Methods for details): identification of reference and validation datasets, definition of the models, model validation, use of the models to identify the systems in bacterial genomes.

Firstly, we searched the primary literature, reviews and books for references of well-studied systems<sup>9,16,20,26,28,32-40</sup>. We used them to define two independent datasets (*reference* and *validation*) of experimentally studied secretion systems (Tables S1 and S2).

Secondly, the *reference dataset* was used to define the model and protein profiles for each type of system. The model includes information on the number of components that are *mandatory* (necessarily present in a system), *accessory* (not necessarily present in the system), and *forbidden* (never present in the system). The occurrences of the components were searched using specific hidden Markov model (HMM) protein profiles with HMMER<sup>41</sup>. Protein profiles are more sensitive and specific than Blast-based approaches<sup>42</sup>. Our models used 204 protein profiles (included in the package, File S1), of which 194 were built in our laboratory and the rest taken from public databases (Tables S4 and S5). We decided to build and use our own profiles instead of using those present in public databases, because they showed better specificity and sensitivity for our purpose: predicting and discriminating accurately secretion systems and related appendages. To quantify these trends we searched for profiles with sequence similarity to our profiles in TIGRFAM (the database providing the most specific profiles in our analyses). Only 102 of the 199 profiles not taken from TIGRFAM had significant hits in that database and nearly half of them (48) had multiple hits for the same profile. A table with TIGRFAM profiles matching ours is available (Table S6). The model indicates which genes are co-localised (at less than a given distance relative to contiguous genes in the cluster), and which genes might be encoded elsewhere in the genome (designated *loners*).

Thirdly, the models were validated both in the *reference* and in the independent *validation* (that was not used to design the models and protein profiles) datasets using MacSyFinder (Table 2)<sup>25</sup>. We also compared our results

| Model               | Systems detected in the validation dataset <sup>a</sup> (detected/total) | Systems detected in the reference dataset (detected/total) |
|---------------------|--|--|
| T1SS                | 6/6  | 8/8  |
| T2SS                | 14/18  | 9/9  |
| T3SS                | See <sup>24</sup>  | See <sup>24</sup>  |
| T4SS                | See <sup>67,107</sup>  | See <sup>67,107</sup>                                      |
| T5aSS               | 4/4  | NA (PFAM)  |
| T5bSS               | 2/3  | 6/6  |
| T5cSS               | 2/2  | NA (PFAM)  |
| T6SS <sup>i</sup>   | 8/8 <sup>e</sup>   | 9/9  |
| T6SS <sup>ii</sup>  | NA   | 1/1  |
| T6SS <sup>iii</sup> | NA   | 3/3  |
| T9SS                | NA   | 2/2  |

**Table 2. Summary of experimentally validated systems detected by TXSScan.** <sup>a</sup>The validation dataset was used to test the validity of the models and profiles built from the reference dataset. <sup>e</sup>Two contiguous T6SS were predicted as a single system (see main text) in *Escherichia coli* O42.

with those of T346Hunter<sup>43</sup> for T3SS and T6SS (see Supplementary TextS1 and Table S7). We could not make a direct comparison of our results and those of the remaining programs in Table 1 because they do not provide tables with results from the analysis of complete genomes (as T346Hunter does) and they cannot be used locally to analyze large datasets.

Finally, the models were used with MacSyFinder to identify occurrences of each system in 1,528 complete genomes of diderm-LPS species. This procedure retrieved automatically all validly predicted secretion systems (Table S3, and see <http://macsydb.web.pasteur.fr>). It also retrieved all hits to each component identified in the genomes whether they are part of a protein secretion system or not. In the following sections we describe the models of each type of protein secretion system and the occurrences of the system in bacterial genomes.

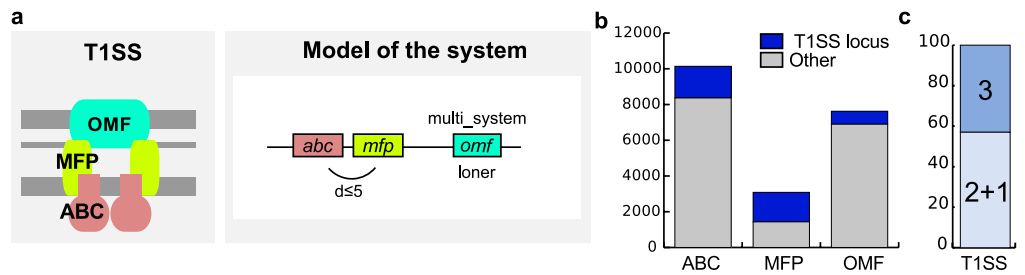
**T1SS.** We built protein profiles for the three essential components of T1SS<sup>32,44,45</sup>: the ABC-transporter (ATP-binding cassette transporter) providing an inner membrane channel, the porin (outer membrane factor, OMF) forming the outer-membrane channel, and the inner-membrane anchored adaptor protein (or membrane-fusion protein, MFP) that connects the OMF and the ABC components (Materials and Methods, Tables S4 and S5). T1SS can be difficult to identify because its components have homologs involved in other machineries, e.g., in ABC transporters for the ABC and in drug efflux systems for the MFP, or can themselves be involved in other machineries, in the case of the OMF<sup>46–50</sup>. The design of the model was facilitated by the previous observation that genes encoding the ABC and MFP components are always co-localised in T1SS loci<sup>32,44</sup>. The model is described in Fig. 1a, and its use resulted in the correct identification of all T1SS in the reference and in the validation datasets. Overall, 20,847 proteins matched the protein profiles of the T1SS components in the bacterial genomes (Fig. 1b). The vast majority of these were not part of T1SS because they did not fit the T1SS model. We found 1,637 occurrences of the T1SS model in 821 genomes (Fig. 1b). The remaining proteins were probably associated with the numerous other systems carrying components homologous to those of the T1SS.

We found T1SS in more than half of the genomes of diderm bacteria (54%). Some genomes contained many systems; e.g., *Bradyrhizobium oligotrophicum* S58 and *Nostoc sp.* PCC 7524 encoded a record number of 9 systems (Table S3). ABC and MFP were encoded together and OMF apart in more than half (57%) of the T1SS (Fig. 1c). We found 95 loci encoding ABC and MFP in replicons lacking OMF. Many of these systems may be functional, since 94 of these loci were found in genomes encoding at least one OMF in another replicon. Multi-replicon functional T1SS have been previously reported<sup>51</sup>.

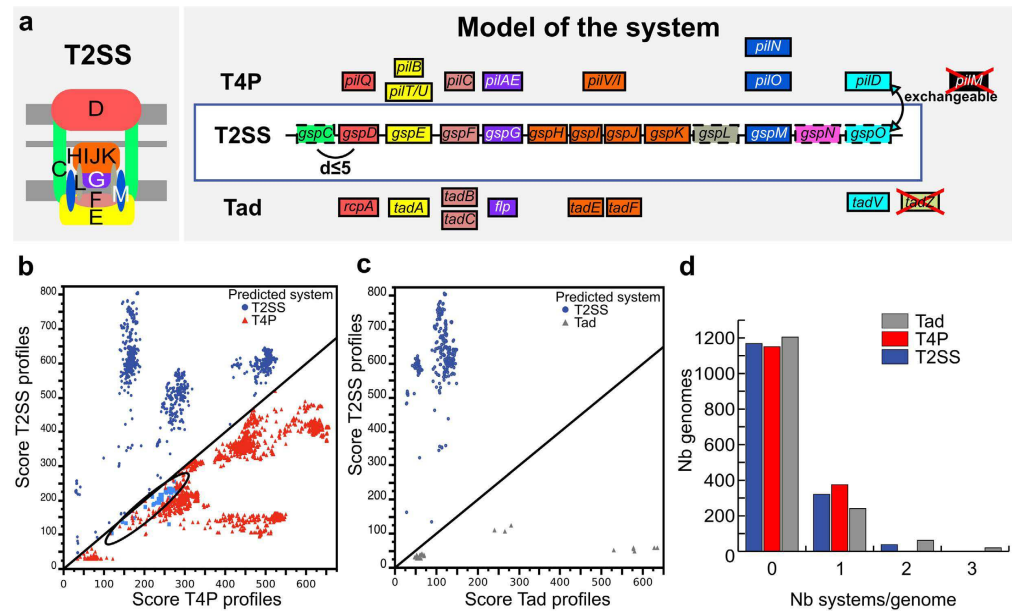
**T2SS, T4P and Tad pili.** T2SS are encoded by 12 to 16 genes, many of which are homologous to components of the T4P and the Tad pilus<sup>17,52,53</sup> (Fig. 2). We used the protein families conserved in the reference dataset to build 13 protein profiles for T2SS, 11 for type IV pili, and 10 for Tad pili (Materials and Methods, Tables S4 and S5). We did not build profiles for GspA and GspB because they were rarely identified in T2SS and their alignments were unreliable due to low sequence similarity. The most frequent components in the reference dataset were defined as mandatory in the models. The least conserved components were defined as accessory. Some profiles built for one type of system produced matches to (homologous) components of other types of systems. Discrimination between systems was facilitated by the definition of some specific components as forbidden (e.g., GspC was declared forbidden in Tad and T4P).

We identified all of the T2SS and Tad systems of the reference dataset using our models for these systems (Table S1). In the validation dataset we missed some components of four of the 18 T2SS (Table 2, Table S2), therefore failing to pass the threshold for a complete system in these four cases. For example, we missed the very atypical T2SS of *Legionella pneumophila* because it failed the co-localisation criterion (unusually, it is encoded in five distant loci, Table S2)<sup>54</sup>. The parameters we selected for our default models may be stringent, but MacSyFinder allows to easily modulate them according to the user's needs. We could for example retrieve three of the four missed T2SS by modifying the default T2SS model, e.g., the “Xps-type” system

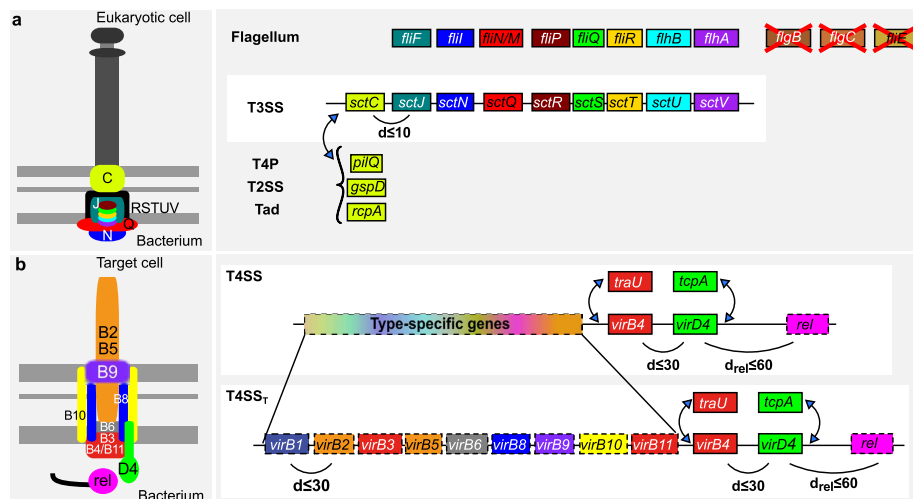




**Figure 1. Model and results for the T1SS.** (a) Schema of the structure (left panel) and model of the genetic organisation (right panel) of T1SS. We built protein profiles for the three components and modelled the two possible genetic architectures of the T1SS: one with the three components encoded in a single locus (*inter\_gene\_max\_space* parameter in MacSyFinder:  $d \leq 5$  genes), another with the ABC transporter and the MFP encoded in a locus while the OMF is further away (*loner* attribute). A single OMF can also be used by different T1SS<sup>46</sup> and this is noted by the attribute *multi\_system*. (b) Distribution of hits for the protein profiles of the T1SS components, separated in two groups: hits effectively part of a T1SS main locus (*i.e.*, containing at least ABC and MFP, blue) and hits found elsewhere (“Other”, grey). Even if encoded outside of “main loci” (grey area of the bar), OMF might be involved in T1SS (*loner* property), whereas it is not the case for ABC and MFP. (c) T1SS encoded in one single locus (ABC, MFP and OMF co-localise) (3) or in two (OMF encoded away from the other components) (2 + 1).



**Figure 2. Model of the T2SS for detection and discrimination from the T4 and Tad pili.** (a) Schema of the structure (left panel) and model of its genetic organisation (right panel), indicating components with homologies with T4P and Tad pilus. We built protein profiles for all these components (Tables S4 and S5). Protein families represented by the same colour are homologous, and their profiles often match proteins from the other systems (except for the Flp and TadE/F families that are less similar). Some prepilin peptidases of T2SS and T4P are defined as functionally interchangeable<sup>109–111</sup> (curved double-headed arrow, *exchangeable* attribute). Boxes represent components: *mandatory* (plain), *accessory* (dashed) and *forbidden* (red crosses). (b) Scores of proteins matched with the profiles of T2SS and T4P. The components of actual T2SS (dark blue) and actual T4P (in red) are well separated, indicating that in each case the best match corresponds to the profile of the correct model system. The exceptions (blue points surrounded by a black ellipse) concern the prepilin peptidases (light blue squares, circled in blue), which are effectively inter-changeable. (c) Representation similar to (b), but for the comparison between T2SS (blue) and Tad (grey) systems. In this case, the separation is perfect: the proteins always match better the protein profile of the correct system. (d) Number of detected systems per genome among the 1,528 genomes of diderm bacteria.



**Figure 3. Model of T3SS and T4SS.** (a) The models of T3SS and flagellum were built based on a previous study<sup>24</sup> (representation conventions as in Fig. 2). Of the nine *mandatory* components for the T3SS only the secretin is *forbidden* in the model of the flagellum. Conversely, three flagellum-specific components are *forbidden* in the T3SS model. Three different types of secretins are found in T3SS derived from different appendages, which are thus defined as *exchangeable* in the model. (b) Models of the T4SS were built based on a previous study<sup>67</sup>. Two different proteins have been described as type 4 coupling proteins (T4CP: VirD4 and TcpA) and two as the major ATPases (VirB4 and TraU, which are homologous). Some pT4SS lack a T4CP and secrete proteins from the periplasm<sup>65</sup>. The relaxase (*rel*), is necessary for conjugation but not for protein secretion, although some relaxase-encoding T4SS are found in both cT4SS and pT4SS<sup>112–114</sup>. Only two MPF types are associated with protein secretion - pT4SS<sub>i</sub> and pT4SS<sub>r</sub>, corresponding to MPF<sub>i</sub> and MPF<sub>r</sub> types. The specificity of type-specific profiles is assessed in Fig. S2.

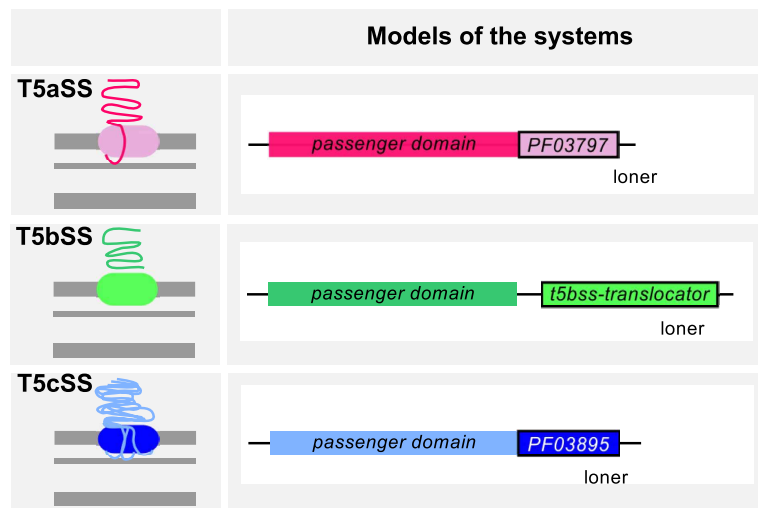
could be detected by decreasing the required number of components<sup>55</sup>. More relaxed parameters in terms of co-localisation and sequence similarity would have identified all T2SS, but at the cost of less correct discrimination from the two homologous systems, T4P and Tad.

The quality of the default T2SS model was confirmed by the analysis of genomic data. Proteins matched by the protein profiles of T2SS were typically either highly or poorly clustered (Fig. S1a). Clusters with many components were typically part of T2SS, whereas small clusters corresponded to other systems. The T2SS components co-localised much closer than the imposed distance threshold ( $d \leq 5$ , Fig. S1b). The vast majority (99%) of the T4P were encoded in multiple distant loci, which is accepted but not required by the model whereas most T2SS were encoded in one single locus (96.5%). To verify that the T2SS, T4P, and Tad loci were correctly classed, we compared the HMMER scores of proteins matched by protein profiles from different systems. Proteins matching profiles from two types of systems scored systematically higher for the system in which they were classed, *i.e.*, secretins of T2SS were systematically matched with a higher score with the profile for the T2SS (Fig. 2b,c).

We detected 400 T2SS in 360 genomes, 379 T4P in 377 genomes, and 425 Tad pili in 323 genomes. The high abundance of Tad pili is surprising given that they are much less studied than the other systems. Interestingly, we found one Tad pilus with the outer membrane channel (the secretin) in one of the rare Firmicutes with an outer membrane (Clostridia, *Acetohalobium arabaticum* DSM 5501)<sup>56</sup>, and also in Acidobacteria, Chlorobi, and Nitrospirae. T4P, T2SS, and to a lesser extent Tad pili, were usually found in a single copy per genome, but some genomes encoded up to three systems (Fig. 2d). The observed small number of T2SS per genome reinforces previous suggestions that many T2SS might secrete several different proteins<sup>57</sup>.

**T3SS and T4SS.** T3SS and T4SS secrete proteins directly into other cells. The T3SS, sometimes also termed non-flagellar T3SS or NF-T3SS, evolved from the flagellar T3SS (F-T3SS) and is encoded by 15 to 25 genes usually in a single locus<sup>24,58,59</sup> (Fig. 3a). Many of the core components of this system are homologous to the distinct F-T3SS that is part of the bacterial flagellum<sup>60–62</sup>. We have previously proposed models that accurately discriminate between the T3SS and the flagellum<sup>24</sup>. We used the same models in this work. We identified 434 NF-T3SS in 334 genomes and 837 flagella in 762 genomes. Some genomes encode many T3SS, *e.g.*, *Burkholderia thailandensis* MSMB121 encodes four T3SS. These results match experimental data showing that in *Burkholderia pseudomallei* the multiple T3SS target different types of cells<sup>63</sup>, and that in *Salmonella enterica* the two T3SS are expressed at different moments in the infection cycle (reviewed in<sup>64</sup>). Multiplicity of T3SS is therefore likely to be associated with complex lifestyles.

T4SS are involved in protein secretion, in conjugation and in some cases in DNA release to, or uptake from, the environment<sup>65</sup>. Here, we distinguished the protein secretion T4SS from the conjugation-related T4SS, which requires a relaxase<sup>66</sup>, by naming them respectively pT4SS and cT4SS. It should be noted that some cT4SS are also able to secrete proteins<sup>65</sup>. We have previously built and validated profiles and models for the pT4SS, and cT4SS<sup>67</sup> (Fig. 3). The latter can be divided in eight sub-types corresponding to different mating pair formation complexes



**Figure 4. Model of the T5aSS, T5bSS and T5cSS.** The left panel shows simplified schemas of the T5SS, and the right panel displays the respective genetic model (only one component that is classed as *loner*). The translocator, pore-forming domains were searched using PFAM domains for T5aSS and T5cSS (resp. PF03797 and PF03895), and a profile built for this work for the T5bSS (Tables S1, S4 and S5).

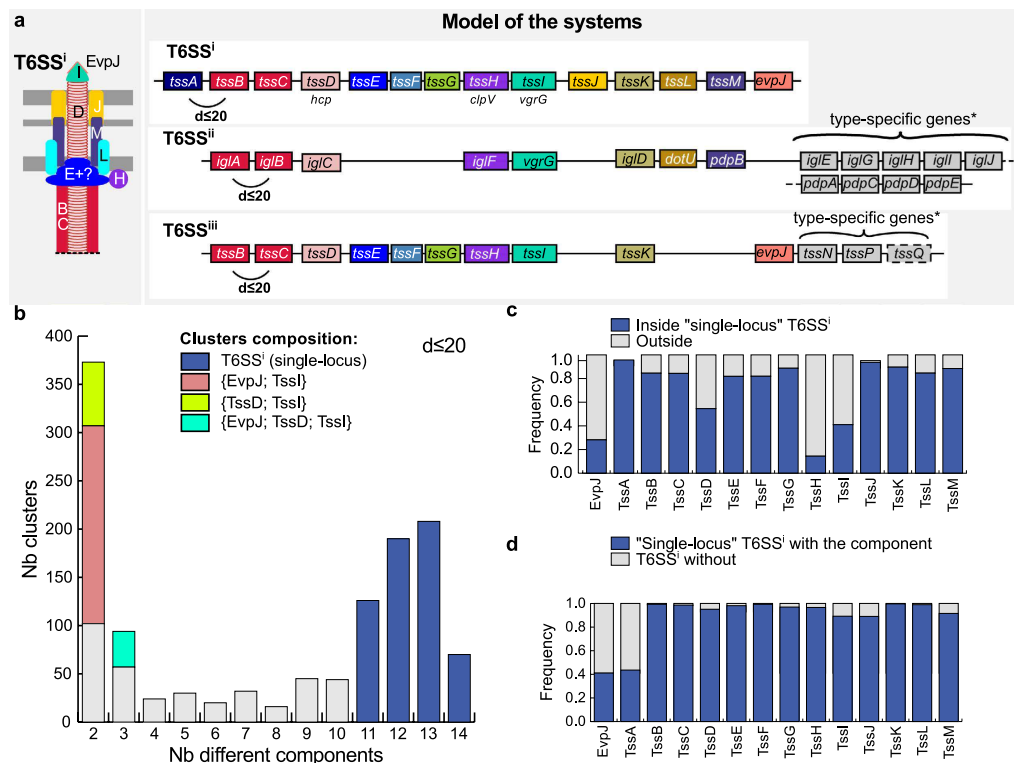
(MPF)<sup>30</sup>, of which six are found in diderm-LPS bacteria, and only two are known to include pT4SS (MPF<sub>I</sub> and MPF<sub>T</sub>). To test the specificity of the models of each T4SS sub-type, we studied the close co-occurrence of T4SS components. The results show that most protein profiles are highly specific to each T4SS sub-type (Fig. S2). Hence, our profiles are able to identify and distinguish between these different systems. We identified 156 pT4SS (among 990 T4SS) in 130 genomes of diderm bacteria (Table S3).

**T5SS.** T5SS are divided in five types (reviewed in<sup>68–71</sup>). Four types encode the translocator (pore-forming) and the passenger (secreted) domains in a single gene: the classical autotransporter (T5aSS), the trimeric autotransporter (T5cSS), the inverted autotransporter (T5eSS), and the fused two-partner system (T5dSS). In two-partner systems (T5bSS), the translocator and passenger are encoded in two separate (typically contiguous) genes. T5SS rely on the Sec machinery for inner-membrane translocation and require other cellular functions for biogenesis. Many of these functions are ubiquitous in diderm-LPS bacteria and do not facilitate the identification of T5SS. Hence, our models only included information on the conserved, mandatory translocator domain of T5SS (Fig. 4, Fig. S3). Two recently proposed families of T5SS - T5dSS and T5eSS<sup>72,73</sup> - were not matched by the T5SS profiles. We will build specific profiles for the detection of these sub-types when enough experimentally validated examples become available.

Our models were able to identify all T5SS in the *reference* and *validation* datasets, with the exception of an atypical T5bSS of *Pseudomonas aeruginosa* consisting of a translocator domain fused with a component of the chaperone usher pathway<sup>74</sup>. We found 3,829 T5aSS in the genomes of diderm bacteria, which makes them by far the most abundant secretion system in our dataset. Certain *Chlamydiae* genomes contain up to 21 T5aSS. We found 1,125 T5bSS (0–8 per genome) and 849 T5cSS (0–24 per genome). T5SS were encoded in 62% of the genomes of diderm bacteria.

**T6SS.** T6SS secrete effectors to bacterial or eukaryotic cells. They were recently divided in three sub-types<sup>40</sup>, among which T6SS<sup>I</sup> is by far the most studied<sup>75–81</sup>. This sub-type has more than a dozen components<sup>78,82</sup>. We built profiles for 14 conserved protein families (Fig. 5a, Materials and Methods, Tables S1, S4 and S5), of which 13 were previously described as the most conserved components of the T6SS<sup>I</sup><sup>20</sup>. The remaining profile corresponds to the PAAR-repeat-containing EvpJ protein family of the spike complex<sup>83</sup>, present in eight out of the nine T6SS<sup>I</sup> in the *reference* dataset. Using this model we identified all T6SS<sup>I</sup> of the *reference* and *validation* datasets. We only found an inaccuracy in *Escherichia coli* O42 where two systems adjacent in the genome were identified as a single system. Part of the T6SS<sup>I</sup> machinery is structurally homologous to the puncturing device of phages, from which it may have originated<sup>84</sup>. Yet, our model did not identify a T6SS<sup>I</sup> in any of the 998 phages present in GenBank, showing that it does not mistake puncturing devices for components of the T6SS.

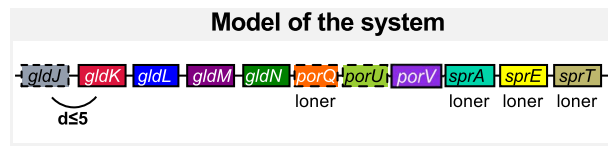
We identified 652 T6SS<sup>I</sup> in 409 bacterial genomes, with up to six T6SS<sup>I</sup> per genome in some *Burkholderia pseudomallei* strains. Around 9% of the T6SS<sup>I</sup> were encoded in multiple loci in the genome. Interestingly, 35% of the replicons encoding a T6SS<sup>I</sup> encoded TssI (VgrG) away from the main loci, with a PAAR-containing component (EvpJ) and/or the chaperone TssD (Hcp) (Fig. 5b–d). PAAR-motifs promote the physical interaction between VgrG and toxins, which are often encoded in the same locus<sup>81,83,85</sup>. It has recently been proposed that VgrG might also be involved in toxin export in a T6SS-independent way<sup>81</sup>. Genomes lacking T6SS<sup>I</sup> did carry some of these small *tssI*-associated clusters, although this corresponded to only 8% of the clusters. Hence, the study of the loci encoding TssI might uncover new T6SS<sup>I</sup> effectors.



**Figure 5. Model and results for the detection of T6SS.** (a) The left panel shows the schema of the structure of T6SS<sup>i</sup>, and the right panel displays the genetic model of the three sub-types of T6SS (representation conventions as in Fig. 2). For T6SS<sup>i</sup>, we built profiles for the 14 mandatory components, which were clustered if at a distance of  $d \leq 20$  (see Fig. S4). For T6SS<sup>ii</sup> and T6SS<sup>iii</sup>, we built 17 and 13 profiles respectively. All components were set as mandatory, except for TssQ, which is found in half of the T6SS<sup>ii</sup>. Homologies between components that are displayed by the mean of the same colours of boxes between the different sub-types are based on previous studies. \*Putative type-specific genes are displayed in grey boxes that do not represent homologies. However, several putative homologies were retrieved using Hhsearch ( $e$ -value  $< 1$  and  $p$ -value  $< 0.05$ ) on T6SS<sup>ii</sup> components: *iglC* (*tssG*), *iglG* (*tssF*), *iglH* (*tssE*), *iglI* (*tssH*) and *pdpD* (*tssH*). (b) Number of different components per cluster of T6SS<sup>i</sup>. Following this analysis, we set the quorum parameter of T6SS<sup>i</sup> to 11. (c) Frequency of hits for each type of T6SS<sup>i</sup> components in the genomes. Hits matching a single-locus T6SS<sup>i</sup> are in blue. The other hits match outside the T6SS<sup>i</sup> loci. (d) Frequency of each component within single-locus T6SS<sup>i</sup>. The components EvpJ and TssA were detected in less than 45% of the T6SS<sup>i</sup>, while the other components were found in most T6SS<sup>i</sup> loci ( $> 89\%$ ).

The T6SS<sup>ii</sup> sub-type described in *Francisella tularensis*, is involved in the subversion of the immune system (growth in macrophages)<sup>39,86–88</sup>. Three of the components of the T6SS<sup>ii</sup> were seldom matched by T6SS<sup>i</sup> profiles (*tssBCL*), complicating the detection of T6SS<sup>ii</sup> with the T6SS<sup>i</sup> model. We built 17 protein profiles and made a specific model for T6SS<sup>ii</sup> based on a *Francisella tularensis* system (see Fig. 5, Materials and Methods and Table S5)<sup>88,89</sup>. Using Hhsearch<sup>90</sup>, we confirmed the existence of weak sequence similarity between the proteins encoded by *tssBCIL* and T6SS<sup>i</sup> and/or T6SS<sup>iii</sup> components ( $p$ -value  $< 0.001$ ). The model detected 30 T6SS<sup>ii</sup> in bacterial genomes. All instances were identified exclusively within the 18 genomes of *Francisella*, and all genomes of the genus contained at least one system.

A recent report identified a new type of T6SS<sup>iii</sup> involved in bacterial competition in *Flavobacterium johnsoniae*<sup>40</sup>. This sub-type lacked homologs of the “trans-envelope subcomplex” and included nine homologs of the 13 described core components of T6SS<sup>i</sup> (Fig. 5). Furthermore, it had three specific components (TssN, TssO and TssP) that are absent in the other sub-types of T6SS. Only three loci were reported for this sub-type<sup>40</sup>. We used them to build 13 protein profiles, including the 12 abovementioned proteins and EvpJ. We could not build a protein profile for TssO because of the lack of sufficient representative conserved sequences (Table S5). The parameters of the model were inferred from the analysis of the clusters of hits for T6SS<sup>iii</sup> components’ profiles (Fig. S4), and from the three reference loci. We predicted 20 T6SS<sup>iii</sup> in 18 of the 97 Bacteroidetes genomes. TssQ, which was not previously recognized as conserved, was found in 50% of the systems’ occurrences. The family of TssQ proteins matched no PFAM profile, but using InterProScan we could predict the presence of one secretion signal and its cellular localisation at the outer-membrane<sup>91,92</sup>. Interestingly, we could find occurrences of EvpJ (harbouring a PAAR domain) within 6 of the 20 T6SS<sup>iii</sup> main loci, and outside of the main locus in 6 genomes with a T6SS<sup>iii</sup>.



**Figure 6. Genetic model of the T9SS.** The representation follows the conventions of Fig. 2. The model includes 11 components for which 13 protein profiles were obtained from PFAM (SprA and SprA-2), TIGRFAM (GldJ, GldK, GldL, GldM, GldN and SprA-3) or designed for this study (PorU, PorV, PorQ, SprE, SprT). Four components were declared as *loners*. The co-localisation distance for the others was set at  $d \leq 5$  (see Fig. S5). As several profiles were available for SprA, we included them all in the models, and declared them as exchangeable homologs in the model. GldJ is not part of the secretion system, but of the gliding motility system. It was included in the model as it facilitates the detection of T9SS components that co-localise with it.

This component co-localised with TssD (Hcp), TssE, or TssD and TssI (VgrG). This suggests it might have similar roles in T6SS<sup>i</sup> and in T6SS<sup>iii</sup>. T6SS<sup>iii</sup> was only identified among Bacteroidetes.

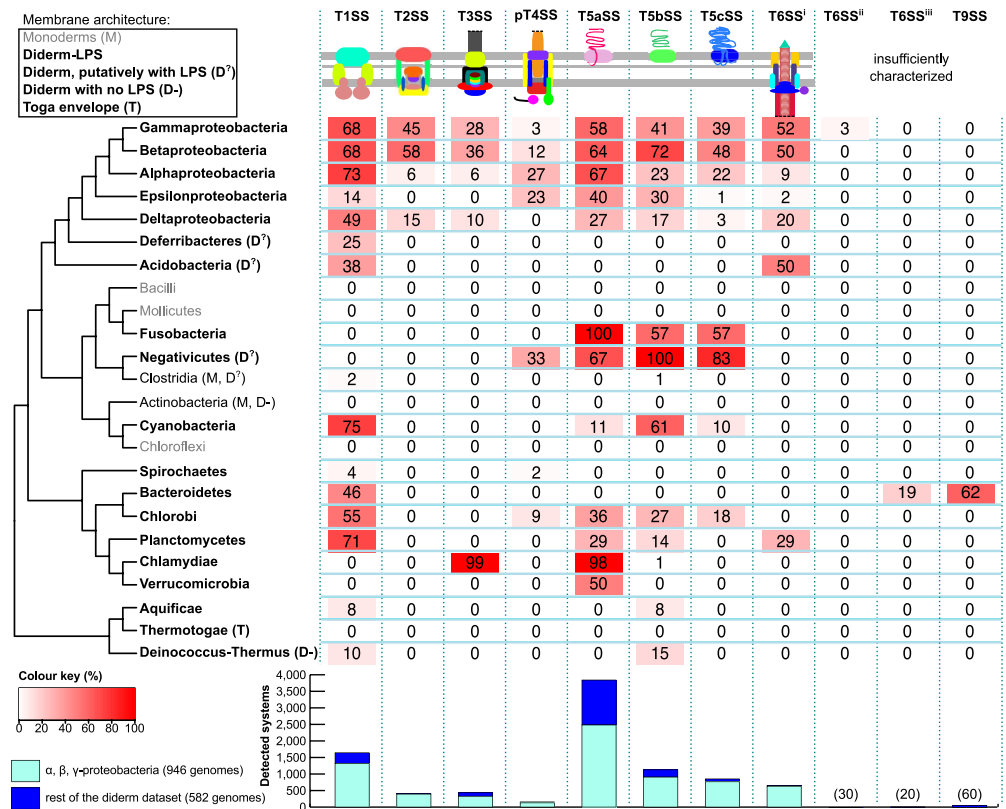
**T9SS.** A novel protein secretion system, T9SS or PorSS, has been described in *F. johnsoniae* and *Porphyromonas gingivalis*<sup>9,93</sup>. It is required for the secretion of components of the gliding motility apparatus, adhesins and various hydrolytic enzymes. We used eight protein profiles from TIGRFAM and PFAM for five components (some having several profiles), and built protein profiles for five other components (Tables S4 and S5, Materials and Methods). One of the profiles was not specifically associated with T9SS; it is part of the gliding motility machinery (GldJ). It was included in the model because it clusters with some of the T9SS components and thus facilitates their identification. Hence, our model for the T9SS includes 13 protein profiles for 10 core components<sup>93</sup> (Fig. 6, Fig. S5). To reflect the reference systems' genetic organization, four components of the T9SS were defined as scattered (*loners*) in the chromosome, whereas the others were defined as part of gene clusters (default behaviour). We detected 60 T9SS in 60 of the 97 genomes of Bacteroidetes, and none in the other bacterial clades, as previously shown<sup>10</sup>. T9SS were found in 62% of the species of Bacteroidetes.

**Distribution of secretion systems.** To the best of our knowledge, this is the first report comparing the frequency of all well-known protein secretion systems of diderm-LPS bacteria in bacterial genomes. Therefore, we analysed the distribution of these protein secretion systems in relation to bacterial phylogeny, including clades with more than four genomes and with reliable information on their phylogenetic position (Fig. 7). Only three clades, Alpha-, Beta- and Gamma-Proteobacteria, encoded all the six most-studied protein secretion systems (T1-T6SS<sup>i</sup>). Delta- and Epsilon-Proteobacteria showed fewer or no T2SS, T3SS and pT4SS. Most other clades encoded fewer types of systems. The distributions of T3SS, T4SS, T6SS, and T9SS have been described recently<sup>10,20,24,30,40</sup>, so we shall focus our analysis on the other systems.

T1SS and T5SS are the most widespread protein secretion systems (Fig. 7, see below). We predicted T1SS in phyla as diverse as Spirochaetes, Planctomycetes, Aquificae, Bacteroidetes, and Cyanobacteria. T1SS involved in the secretion of glycolipids for heterocysts formation were recently described in filamentous Cyanobacteria<sup>94,95</sup>. We found that T1SS were particularly abundant in this clade as 75% of the genomes harboured at least one T1SS. Overall, the three types of T5SS showed similar taxonomic distributions, even if T5cSS were less widespread (Fig. 7). Some phyla had only one type of T5SS: T5aSS in Thermodesulfobacteria and T5bSS in Aquificae and Deinococcus-Thermus. There were few genomes available for these clades. Further work will be needed to know if they lack the other T5SS.

We predicted very few T2SS outside Proteobacteria. T2SS were also absent from the 98 genomes of Epsilon-proteobacteria. We found a T2SS in a non-Proteobacterium, *Desulfurispirillum indicum* S5, a free-living spiral-shaped aquatic Chrysiogenetes (also encoding a T1SS). We could not find a description of the membrane architecture for this species, but our analysis reinforces previous suggestions that it is a diderm<sup>96</sup>. Putative T2SS were previously identified in clades where we failed to identify complete systems: *Synechococcus elongatus* (Cyanobacteria), *Chlamydia trachomatis* (Chlamydiae) and *Leptospira interrogans* (Spirochaetes)<sup>28,97-99</sup>. The cyanobacterial system, which has a role in protein secretion and biofilm formation, seems to be a typical T4P encoded in multiple loci. The role of T4P in secreting proteins that are not part of its structure has been described before<sup>100</sup>. To the best of our knowledge, the function of the *Leptospira* system was not experimentally tested. The Chlamydiae system was indeed associated with protein secretion<sup>98</sup>. From the point of view of our models the putative T2SS from these two last clades form incomplete systems (although they could be retrieved by lowering the minimum required number of components for a valid T2SS in the model). Preliminary phylogenetic analyses did not allow conclusive assignment of these systems to T2SS or to T4P. Further experimental and computational work will be necessary for their precise characterisation.

**Secretion systems and the cell envelope.** The distribution of secretion systems is linked with the structure of the cell envelope. Expectedly, all genomes of monoderms lacked loci encoding diderm-like protein secretion system. Several clades of diderm bacteria lacked many types of protein secretion systems, but only one lacked them all: the Thermotogae. These bacteria are thermophilic, and one could speculate that high temperatures could be incompatible with the protein secretion systems that we searched for. Yet, life under high temperatures is also typical of the sister-clade Aquificae, where we found T1SS and T5SS. The lack of typical protein secretion systems in Thermotogae might be caused by the peculiar sheath-like structure present in their outer cell envelope,



**Figure 7. Phylogenetic distribution of protein secretion systems in bacteria.** Within each clade, the proportion of genomes harbouring each system is indicated in boxes whose colours follow a gradient from full red (100%) to white (0%) (see legend). Clades were classed as monoderms (grey or “M” symbol), diderms with Lipopolysaccharide-containing outer membranes (Diderm-LPS in bold, no symbol), diderms with homologs of LPS pathway that putatively have LPS (D<sup>2</sup>) and diderms with no LPS (D-). The peculiar envelope of the Thermotogae is indicated (T). The Firmicutes are typically monoderms, but some of their members are diderms (the Negativicutes, some Clostridia, Mycobacteria). The bar plot shows the number of detected systems. Bars are split in two categories to separate on one side Alpha- Beta- and Gamma-proteobacteria, and on the other genomes from other bacteria. We display the number of occurrences of systems occurring rarely in our dataset on top of the bars. Clades with less than 4 genomes and/or with unreported phylogenetic position are not shown (*i.e.*, Chrysiogenetes, Gemmatimonadetes, Nitrospirae and Thermodesulfobacteria). This sketch tree was drawn from the compilation of different published phylogenetic analyses<sup>115–118</sup>.

the “toga”<sup>101</sup>. This may have led to the evolution of secretion systems specifically adapted to this structure. Accordingly, only a few porins have been identified so far in Thermotogae<sup>102</sup>. In an analogous way, *Mycobacteria* (Actinobacteria), which have a peculiar mycolate outer membrane, have specific secretion systems<sup>12</sup>.

The cell envelope of recipient cells is also a key determinant of the evolution of systems secreting effectors directly into other cells. The extracellular structures of T3SS are tightly linked with the type of eukaryote cell (plant vs. animal) with which the bacterium interacts<sup>24</sup>.

Interestingly, diderm bacteria in taxa dominated by monoderms have protein secretion systems homologous to those of Proteobacteria (including Clostridia, Cyanobacteria, Fusobacteria and Negativicutes). For example, we predicted in Negativicutes (a clade of Firmicutes) putative pT4SS and the three types of T5SS. Some genomes of Halanaerobiales (a sub-clade of Clostridia, Firmicutes) encode T1SS and T5bSS. Similarities in the cell envelope may thus lead to the presence of similar systems in very distant bacteria.

## Conclusion

We were able to identify nearly all protein secretion systems in both the *reference* and the *validation* datasets. The few missed systems were either very atypical (such as the scattered T2SS of *Legionella*) or included components very divergent in sequence (several T2SS). In the latter case, the relaxation of the parameters of the T2SS model allowed their identification. We emphasize that our models are publicly available and can be modified by the user to increase their sensitivity. Relaxing the parameters for the detection of the components (HMMER i-value and profile coverage), or for the genetic organization (required quorum of components, co-localisation criterion) often allowed retrieving more putative systems. We emphasize that we have not modified the default models in function of the validation procedure because that would have made our validation procedure inaccurate. Yet, the

user is free to take the default models and make them less strict. Nevertheless, this might lead to an increased number of false positives. Complementary analyses can also facilitate the identification of systems. For example, when multiple profiles match a given protein, the one of the system usually provides the highest score (Fig. 2b,c). This is one of the advantages of using specifically designed protein profiles, instead of generic profiles as can be found in PFAM: the system-specific profiles distinguish between homologs components in different types of molecular systems.

We may have under-estimated the presence of protein secretion systems in poorly sampled phyla because of the rapid evolution of extracellular components and the paucity of experimental data. Yet, several pieces of evidence suggest that we may have identified most systems. 1) We identified almost all known systems in the *reference* and *validation* datasets. 2) We identified at least one type of secretion system in almost all clades of diderm bacteria. 3) We identified components of T4P and Tad (homologous to T2SS), F-T3SS (homologous to NF-T3SS), and cT4SS (homologous to pT4SS) with profiles for the protein secretion systems in many clades, including monoderms (Table S3). Most of these systems are monophyletic. If our protein profiles match homologs in outgroup systems, then they probably match all occurrences of the system. Given these arguments, it is tempting to speculate that currently unknown protein secretion systems remain to be discovered in clades where few or no secretion systems could be identified. Interestingly, the recently discovered T6SS<sup>iii</sup> and T9SS are restricted to Bacteroidetes<sup>9,40</sup>, while T6SS<sup>ii</sup> are only found in *Francisella*. The search for protein secretion systems and other cellular appendages with relaxed criteria may help in identifying novel unknown systems.

## Materials and Methods

**Data.** The genomes of bacteria (2,484) and archaea (159) were downloaded from NCBI RefSeq (<ftp://ftp.ncbi.nih.gov/genomes/>, November 2013). We took from this dataset the 1,528 genomes of bacteria classed as diderm in the literature<sup>103–105</sup> (Table S3). A total of 998 genomes of phages were downloaded from Genbank (last access, February 2013). The sequences of the reference protein secretion systems were retrieved from Genbank or from complete genomes (Tables S1 and S2).

**Systems definition and identification.** We built a dataset of experimentally studied secretion systems (T1SS–T6SS, T9SS) and related appendages (Tad, T4P and the bacterial flagellum) from the analysis of published data. We selected these systems in order to maximise sequence diversity. They form our *reference* set of systems (Table S1). This *reference* dataset was used to build the models and the corresponding HMM protein profiles (see below) of each system using MacSyFinder. This software is publicly available<sup>25</sup>. A detailed explanation of this program can be found in its original publication<sup>25</sup>. Here, we focus on the features that are pertinent for this work. A model in MacSyFinder defines the components of the secretion system, the minimal acceptable number of components, and their genetic organisation. Among other things (see <http://macsyfinder.readthedocs.org> for full documentation), one can specify the following relevant information. 1) Systems are encoded in a single locus (*single-locus* system) or in several loci (*multi-loci* system). 2) Core components (ubiquitous and essential) are defined as *mandatory*. 3) Components that are accessory or poorly conserved in sequence are defined as *accessory*. These components are accessory for the computational model, but their function may be essential. This happens when different proteins have analogous functions or when proteins evolve so fast that distant homologs are not recognisable by sequence analysis. 4) Some genes are ubiquitous and specific to a system and can be defined as *forbidden* in models of other systems. This facilitates the discrimination between systems with homologous components. For example, the NF-T3SS-specific secretin may be declared as *forbidden* in the F-T3SS. 5) An occurrence of a system is validated when a pre-defined number (*quorum*) of mandatory components and/or sum of mandatory and accessory components is found<sup>25</sup>. 6) Components can be defined as reciprocally *exchangeable* in the quorum (which prevents them from being counted twice). 7) Two components are co-localised when they are separated by less than a given number of genes (parameter  $d = \text{inter\_gene\_max\_space}$ ). 8) A component defined with the *loner* attribute does not need to be co-localised with other components to be part of a system (e.g., OMF in T1SS). 9) A component that can participate in several instances of a system (e.g., OMF in T1SS) receives the *multi\_system* attribute. These different properties can be combined when necessary.

The models for the different protein secretion systems were described using a dedicated Extensible Markup Language (XML) grammar<sup>25</sup>. The files with the models were named after the system (e.g., T1SS.xml, File S1). Models can be easily modified on the standalone version of MacSyFinder. The webserver allows the use of the pre-defined models and the modification of the most important search parameters.

MacSyFinder was used to identify protein secretion systems in bacterial genomes in three steps (for corresponding command-lines see the README file in File S1, and for a full description of the software see<sup>25</sup>). Firstly, components were identified using protein profile searches with HMMER<sup>41</sup>. Hits with alignments covering more than 50% of the protein profile and with an *i*-value  $< 10^{-3}$  were kept for further analysis (default parameters). Secondly, the components were clustered according to their proximity in the genome using the parameter *d*. Finally, the clusters were validated if they passed the criteria specified in the model.

**Definition of protein profiles.** The models include 204 protein profiles. The two profiles for T5aSS and T5cSS were extracted from PFAM<sup>68,91</sup>. Eight profiles for T9SS were extracted from PFAM or TIGRFAM<sup>91,106</sup>. The remaining 194 profiles were the result of our previous work<sup>24,67,107</sup> or this study (84 protein profiles for T1SS, T2SS, Tad, type IV pilus, T5bSS, T6SS<sup>i</sup>, T6SS<sup>ii</sup>, T6SS<sup>iii</sup> and T9SS, listed in Table S4). To build the new profiles, we sampled the experimentally studied systems from our *reference* set of systems for proteins representative of each component of each system. Protein families were constructed by clustering homologous proteins using sequence similarity. The details of the methods and parameters used to build each protein profile are described in Table S5. In the case of the T9SS, where only two systems were experimentally characterised, we used components from the well-studied system of *F. johnsoniae* (or *P. gingivalis* when the gene was absent from *F. johnsoniae*) for

Blastp searches against our database of complete genomes, and retained the best sequence hits (e-value < 10<sup>-20</sup>) to constitute protein families. A similar approach was taken to build protein profiles for the T6SS<sup>si</sup>, based on the *Francisella tularensis* subsp. *tularensis* SCHU S4 FPI system displayed in Table 1 of<sup>39</sup>. The largest families were aligned and manually curated to produce hidden Markov model profiles with HMMER 3.0<sup>41</sup>.

**Availability.** Detection and visualization of all systems described in this paper can be performed online on the Mobyle-based<sup>108</sup> webserver TXSScan: <http://mobyle.pasteur.fr/cgi-bin/portal.py#forms:txsscan>. Detection can also be performed locally using the standalone program MacSyFinder<sup>25</sup>, and the sets of models and profiles described here. MacSyFinder is freely available for all platforms at <https://github.com/gem-pasteur/macsyfinder>. Models and required protein profiles are available as supplemental material (File S1) at <https://research.pasteur.fr/en/tool/txsscan-models-and-profiles-for-protein-secretion-systems>. The models are provided as simple text (XML) files, so they can be easily modified and extended by the user. The results of MacSyFinder can be visualized with MacSyView, available online at <http://macsyview.web.pasteur.fr> or for download at <https://github.com/gem-pasteur/macsyview> (also included in the release of MacSyFinder). The systems detected in this study are available on the form of a database at <http://macsydb.web.pasteur.fr>.

## References

- Wandersman, C. & Delepelaire, P. Bacterial iron sources: from siderophores to hemophores. *Annu. Rev. Microbiol.* **58**, 611 (2004).
- Ruhe, Z. C., Low, D. A. & Hayes, C. S. Bacterial contact-dependent growth inhibition. *Trends Microbiol.* **21**, 230 (2013).
- Viprey, V., Del Greco, A., Golinowski, W., Broughton, W. J. & Perret, X. Symbiotic implications of type III protein secretion machinery in *Rhizobium*. *Mol. Microbiol.* **28**, 1381 (1998).
- Ma, W. & Guttman, D. S. Evolution of prokaryotic and eukaryotic virulence effectors. *Curr. Opin. Plant Biol.* **11**, 412 (2008).
- Raymond, B. *et al.* Subversion of trafficking, apoptosis, and innate immunity by type III secretion system effectors. *Trends Microbiol.* **21**, 430 (2013).
- Bleves, S. *et al.* Protein secretion systems in *Pseudomonas aeruginosa*: A wealth of pathogenic weapons. *Int. J. Med. Microbiol.* **300**, 534 (2010).
- Dalbey, R. E. & Kuhn, A. Protein traffic in Gram-negative bacteria—how exported and secreted proteins find their way. *FEMS Microbiol. Rev.* **36**, 1023 (2012).
- Chang, J. H., Desveaux, D. & Creason, A. L. The ABCs and 123s of bacterial secretion systems in plant pathogenesis. *Annu Rev Phytopathol* **52**, 317 (2014).
- Sato, K. *et al.* A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc. Natl. Acad. Sci. USA* **107**, 276 (2010).
- McBride, M. J. & Zhu, Y. Gliding motility and Por secretion system genes are widespread among members of the phylum bacteroidetes. *J. Bacteriol.* **195**, 270 (2013).
- Desvaux, M., Hebraud, M., Talon, R. & Henderson, I. R. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* **17**, 139 (2009).
- Abdallah, A. M. *et al.* Type VII secretion—mycobacteria show the way. *Nat. Rev. Microbiol.* **5**, 883 (2007).
- Tseng, T. T., Tyler, B. M. & Setubal, J. C. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol.* **9** Suppl 1, S2 (2009).
- Planet, P. J., Kachlany, S. C., DeSalle, R. & Figurski, D. H. Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. *Proc. Natl. Acad. Sci. USA* **98**, 2503 (2001).
- Minamino, T. & Namba, K. Self-assembly and type III protein export of the bacterial flagellum. *J. Mol. Microbiol. Biotechnol.* **7**, 5 (2004).
- Pellicic, V. Type IV pili: e pluribus unum? *Mol. Microbiol.* **68**, 827 (2008).
- Peabody, C. R. *et al.* Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149**, 3051 (2003).
- Nogueira, T., Touchon, M. & Rocha, E. P. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. *PLoS ONE* **7**, e49403 (2012).
- Gophna, U., Ron, E. Z. & Graur, D. Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events. *Gene* **312**, 151 (2003).
- Boyer, F., Fichant, G., Berthod, J., Vandenbrouck, Y. & Attree, I. Dissecting the bacterial type VI secretion system by a genome wide in silico analysis: what can be learned from available microbial genomic resources? *BMC Genomics* **10**, 104 (2009).
- Ren, C. P. *et al.* The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J. Bacteriol.* **186**, 3547 (2004).
- Huynen, M., Snel, B., Lathe, W., 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204 (2000).
- Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* **11**, 356 (2001).
- Abby, S. S. & Rocha, E. P. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genet.* **8**, e1002983 (2012).
- Abby, S. S., Neron, B., Menager, H., Touchon, M. & Rocha, E. P. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS ONE* **9**, e110726 (2014).
- Yen, M. R. *et al.* Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochim. Biophys. Acta* **1562**, 6 (2002).
- Pallen, M. J., Beatson, S. A. & Bailey, C. M. Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. *FEMS Microbiol. Rev.* **29**, 201 (2005).
- Cianciotto, N. P. Type II secretion: a protein secretion system for all seasons. *Trends Microbiol.* **13**, 581 (2005).
- Saier, M. H., Ma, C. H., Rodgers, L., Tamang, D. G. & Yen, M. R. Protein secretion and membrane insertion systems in bacteria and eukaryotic organelles. *Adv. Appl. Microbiol.* **65**, 141 (2008).
- Guglielmini, J., de la Cruz, F. & Rocha, E. P. Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.* **30**, 315 (2013).
- Barret, M., Egan, F. & O'Gara, F. Distribution and diversity of bacterial secretion systems across metagenomic datasets. *Environmental microbiology reports* **5**, 117 (2013).
- Delepelaire, P. Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta* **1694**, 149 (2004).
- Bouige, P., Laurent, D., Piloyan, L. & Dassa, E. Phylogenetic and functional classification of ATP-binding cassette (ABC) systems. *Curr. Protein Pept. Sci.* **3**, 541 (2002).
- Dassa, E. & Bouige, P. The ABC of ABCs: a phylogenetic and functional classification of ABC systems in living organisms. *Res. Microbiol.* **152**, 211 (2001).



35. Jacob-Dubuisson, F., Fernandez, R. & Coutte, L. Protein secretion through autotransporter and two-partner pathways. *Biochim. Biophys. Acta* **1694**, 235 (2004).
36. Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C. & Ala'Aldeen, D. Type V protein secretion pathway: the autotransporter story. *Microbiol. Mol. Biol. Rev.* **68**, 692 (2004).
37. Linke, D., Riess, T., Autenrieth, I. B., Lupas, A. & Kempf, V. A. Trimeric autotransporter adhesins: variable structure, common function. *Trends Microbiol.* **14**, 264 (2006).
38. Tomich, M., Planet, P. J. & Figurski, D. H. The tad locus: postcards from the widespread colonization island. *Nat. Rev. Microbiol.* **5**, 363 (2007).
39. Broms, J. E., Sjostedt, A. & Lavander, M. The Role of the Francisella Tularensis Pathogenicity Island in Type VI Secretion, Intracellular Survival, and Modulation of Host Cell Signaling. *Front. Microbiol.* **1**, 136 (2010).
40. Russell, A. B. *et al.* A Type VI Secretion-Related Pathway in Bacteroidetes Mediates Interbacterial Antagonism. *Cell Host Microbe* **16**, 227 (2014).
41. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
42. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
43. Martinez-Garcia, P. M., Ramos, C. & Rodriguez-Palenzuela, P. T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS ONE* **10**, e0119317 (2015).
44. Holland, I. B., Schmitt, L. & Young, J. Type I protein secretion in bacteria, the ABC-transporter dependent pathway. *Mol. Membr. Biol.* **22**, 29 (2005).
45. Kanonenberg, K., Schwarz, C. K. & Schmitt, L. Type I secretion systems - a story of appendices. *Res. Microbiol.* **164**, 596 (2013).
46. Paulsen, I. T., Park, J. H., Choi, P. S. & Saier, M. H. Jr. A family of gram-negative bacterial outer membrane factors that function in the export of proteins, carbohydrates, drugs and heavy metals from gram-negative bacteria. *FEMS Microbiol. Lett.* **156**, 1 (1997).
47. Dinh, T., Paulsen, I. T. & Saier, M. H. Jr. A family of extracytoplasmic proteins that allow transport of large molecules across the outer membranes of gram-negative bacteria. *J. Bacteriol.* **176**, 3825 (1994).
48. Dassa, E. Natural history of ABC systems: not only transporters. *Essays Biochem.* **50**, 19 (2011).
49. Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* **72**, 317 (2008).
50. Koronakis, V., Eswaran, J. & Hughes, C. Structure and function of TolC: the bacterial exit duct for proteins and drugs. *Annu. Rev. Biochem.* **73**, 467 (2004).
51. Burland, V. *et al.* The complete DNA sequence and analysis of the large virulence plasmid of Escherichia coli O157:H7. *Nucleic Acids Res.* **26**, 4196 (1998).
52. Korotkov, K. V., Sandkvist, M. & Hol, W. G. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10**, 336 (2012).
53. Nivaskumar, M. & Francetic, O. Type II secretion system: a magic beanstalk or a protein escalator. *Biochim. Biophys. Acta* **1843**, 1568 (2014).
54. Cianciotto, N. P. Many substrates and functions of type II secretion: lessons learned from Legionella pneumophila. *Future Microbiol.* **4**, 797 (2009).
55. Karaba, S. M., White, R. C. & Cianciotto, N. P. Stenotrophomonas maltophilia Encodes a Type II Protein Secretion System That Promotes Detrimental Effects on Lung Epithelial Cells. *Infect. Immun.* **81**, 3210 (2013).
56. Zhilina, T. N. & Zavarzin, G. A. Extremely halophilic, methylotrophic, anaerobic bacteria. *FEMS Microbiol. Lett.* **87**, 315 (1990).
57. Rondelet, A. & Condemine, G. Type II secretion: the substrates that won't go away. *Res. Microbiol.* **164**, 556 (2013).
58. Galan, J. E. & Wolf-Watz, H. Protein delivery into eukaryotic cells by type III secretion machines. *Nature* **444**, 567 (2006).
59. Cornelis, G. R. The type III secretion injectisome. *Nat. Rev. Microbiol.* **4**, 811 (2006).
60. Ginocchio, C. C., Olmsted, S. B., Wells, C. L. & Galan, J. E. Contact with epithelial cells induces the formation of surface appendages on Salmonella typhimurium. *Cell* **76**, 717 (1994).
61. Van Gijsegem, F. *et al.* The hrp gene locus of Pseudomonas solanacearum, which controls the production of a type III secretion system, encodes eight proteins related to components of the bacterial flagellar biogenesis complex. *Mol. Microbiol.* **15**, 1095 (1995).
62. Young, G. M., Schmiel, D. H. & Miller, V. L. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. *Proc. Natl. Acad. Sci. USA* **96**, 6456 (1999).
63. Sun, G. W. & Gan, Y. H. Unraveling type III secretion systems in the highly versatile Burkholderia pseudomallei. *Trends Microbiol.* **18**, 561 (2010).
64. Hansen-Wester, I. & Hensel, M. Salmonella pathogenicity islands encoding type III secretion systems. *Microbes Infect.* **3**, 549 (2001).
65. Alvarez-Martinez, C. E. & Christie, P. J. Biological diversity of prokaryotic type IV secretion systems. *Microbiol. Mol. Biol. Rev.* **73**, 775 (2009).
66. de la Cruz, F., Frost, L. S., Meyer, R. J. & Zechner, E. Conjugative DNA Metabolism in Gram-negative Bacteria. *FEMS Microbiol. Rev.* **34**, 18 (2010).
67. Guglielmini, J. *et al.* Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.* (2014).
68. Dautin, N. & Bernstein, H. D. Protein Secretion in Gram-Negative Bacteria via the Autotransporter Pathway. *Annu. Rev. Microbiol.* **61**, 89 (2007).
69. Mazar, J. & Cotter, P. A. New insight into the molecular mechanisms of two-partner secretion. *Trends Microbiol.* **15**, 508 (2007).
70. Leyton, D. L., Rossiter, A. E. & Henderson, I. R. From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis. *Nat. Rev. Microbiol.* **10**, 213 (2012).
71. Leo, J. C., Grin, I. & Linke, D. Type V secretion: mechanism(s) of autotransport through the bacterial outer membrane. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 1088 (2012).
72. Salacha, R. *et al.* The Pseudomonas aeruginosa patatin-like protein PlpD is the archetype of a novel Type V secretion system. *Environ. Microbiol.* **12**, 1498 (2010).
73. Oberhettinger, P. *et al.* Intimin and invasin export their C-terminus to the bacterial cell surface using an inverse mechanism compared to classical autotransport. *PLoS ONE* **7**, e47069 (2012).
74. Ruer, S., Ball, G., Filloux, A. & de Bentzmann, S. The 'P-usher', a novel protein transporter involved in fimbrial assembly and TpsA secretion. *EMBO J.* **27**, 2669 (2008).
75. Mougous, J. D. *et al.* A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. *Science* **312**, 1526 (2006).
76. Hood, R. D. *et al.* A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria. *Cell Host Microbe* **7**, 25 (2010).
77. Schwarz, S. *et al.* Burkholderia type VI secretion systems have distinct roles in eukaryotic and bacterial cell interactions. *PLoS Pathog.* **6**, e1001068 (2010).
78. Silverman, J. M., Brunet, Y. R., Cascales, E. & Mougous, J. D. Structure and regulation of the type VI secretion system. *Annu. Rev. Microbiol.* **66**, 453 (2012).
79. Basler, M., Ho, B. T. & Mekalanos, J. J. Tit-for-tat: type VI secretion system counterattack during bacterial cell-cell interactions. *Cell* **152**, 884 (2013).

80. Brunet, Y. R., Espinosa, L., Harchouni, S., Mignot, T. & Cascales, E. Imaging type VI secretion-mediated bacterial killing. *Cell reports* **3**, 36 (2013).
81. Hachani, A., Allsopp, L. P., Oduko, Y. & Filloux, A. The VgrG proteins are “A la carte” delivery systems for bacterial type VI effectors. *J. Biol. Chem.* **289**, 17872 (2014).
82. Cascales, E. & Cambillau, C. Structural biology of type VI secretion systems. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367**, 1102 (2012).
83. Shneider, M. M. *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **500**, 350 (2013).
84. Pukatzki, S., Ma, A. T., Revel, A. T., Sturtevant, D. & Mekalanos, J. J. Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. *Proc. Natl. Acad. Sci. USA* **104**, 15508 (2007).
85. Whitney, J. C. *et al.* Genetically distinct pathways guide effector export through the type VI secretion system. *Mol. Microbiol.* **92**, 529 (2014).
86. Nano, F. E. *et al.* A Francisella tularensis pathogenicity island required for intramacrophage growth. *J. Bacteriol.* **186**, 6430 (2004).
87. Ludu, J. S. *et al.* The Francisella pathogenicity island protein PdpD is required for full virulence and associates with homologues of the type VI secretion system. *J. Bacteriol.* **190**, 4584 (2008).
88. Barker, J. R. *et al.* The Francisella tularensis pathogenicity island encodes a secretion system that is required for phagosome escape and virulence. *Mol. Microbiol.* **74**, 1459 (2009).
89. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
90. Soding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951 (2005).
91. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281 (2008).
92. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116 (2005).
93. Kharade, S. S. & McBride, M. J. Flavobacterium johnsoniae PorV is required for secretion of a subset of proteins targeted to the type IX secretion system. *J. Bacteriol.* **197**, 147 (2015).
94. Moslavac, S. *et al.* A TolC-like protein is required for heterocyst development in Anabaena sp. strain PCC 7120. *J. Bacteriol.* **189**, 7887 (2007).
95. Staron, P., Forchhammer, K. & Maldener, I. Structure-function analysis of the ATP-driven glycolipid efflux pump DevBCA reveals complex organization with TolC/HgdD. *FEBS Lett.* **588**, 395 (2014).
96. Rauschenbach, I., Yee, N., Haggblom, M. M. & Bini, E. Energy metabolism and multiple respiratory pathways revealed by genome sequencing of Desulfurispirillum indicum strain S5. *Environ. Microbiol.* **13**, 1611 (2011).
97. Zeng, L. *et al.* Extracellular proteome analysis of Leptospira interrogans serovar Lai. *Omic: a journal of integrative biology* **17**, 527 (2013).
98. Nguyen, B. D. & Valdivia, R. H. Virulence determinants in the obligate intracellular pathogen Chlamydia trachomatis revealed by forward genetic approaches. *Proc. Natl. Acad. Sci. USA* **109**, 1263 (2012).
99. Schatz, D. *et al.* Self-suppression of biofilm formation in the cyanobacterium Synechococcus elongatus. *Environ. Microbiol.* **15**, 1786 (2013).
100. Hager, A. J. *et al.* Type IV pili-mediated secretion modulates Francisella virulence. *Mol. Microbiol.* **62**, 227 (2006).
101. Huber, R. *et al.* Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 C. *Arch. Microbiol.* **144**, 324 (1986).
102. Petrus, A. K. *et al.* Genes for the major structural components of Thermotoga species’ togas revealed by proteomic and evolutionary analyses of OmpA and OmpB homologs. *PLoS ONE* **7**, e40236 (2012).
103. Sutcliffe, I. C. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* **18**, 464 (2010).
104. Francke, C. *et al.* Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. *BMC Genomics* **12**, 385 (2011).
105. Vesth, T. *et al.* Veillonella, Firmicutes: Microbes disguised as Gram negatives. *Stand Genomic Sci* **9**, 431 (2013).
106. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**, D387 (2013).
107. Guglielmini, J., Quintais, L., Garcillan-Barcia, M. P., de la Cruz, F. & Rocha, E. P. The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.* **7**, e1002222 (2011).
108. Neron, B. *et al.* Mobyle: a new full web bioinformatics framework. *Bioinformatics* **25**, 3005 (2009).
109. Nunn, D. N. & Lory, S. Product of the Pseudomonas aeruginosa gene pilD is a prepilin leader peptidase. *Proc. Natl. Acad. Sci. USA* **88**, 3281 (1991).
110. Pepe, C. M., Eklund, M. W. & Strom, M. S. Cloning of an Aeromonas hydrophila type IV pilus biogenesis gene cluster: complementation of pilus assembly functions and characterization of a type IV leader peptidase/N-methyltransferase required for extracellular protein secretion. *Mol. Microbiol.* **19**, 857 (1996).
111. Marsh, J. W. & Taylor, R. K. Identification of the Vibrio cholerae type 4 prepilin peptidase required for cholera toxin secretion and pilus formation. *Mol. Microbiol.* **29**, 1481 (1998).
112. Christie, P. J. Type IV secretion: the Agrobacterium VirB/D4 and related conjugation systems. *Biochim. Biophys. Acta* **1694**, 219 (2004).
113. Nagai, H. & Kubori, T. Type IVB Secretion Systems of Legionella and Other Gram-Negative Bacteria. *Front. Microbiol.* **2**, 136 (2011).
114. Schroder, G., Schuelein, R., Quebatte, M. & Dehio, C. Conjugative DNA transfer into human cells by the VirB/VirD4 type IV secretion system of the bacterial pathogen Bartonella henselae. *Proc. Natl. Acad. Sci. USA* **108**, 14643 (2011).
115. Abby, S. S., Tannier, E., Gouy, M. & Daubin, V. Lateral gene transfer as a support for the tree of life. *Proc. Natl. Acad. Sci. USA* **109**, 4962 (2012).
116. Yutin, N., Puigbo, P., Koonin, E. V. & Wolf, Y. I. Phylogenomics of prokaryotic ribosomal proteins. *PLoS ONE* **7**, e36972 (2012).
117. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**, 1056 (2009).
118. Boussau, B., Gueguen, L. & Gouy, M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. *BMC Evol. Biol.* **8**, 272 (2008).
119. Souza, R. C. *et al.* AtlasT4SS: a curated database for type IV secretion systems. *BMC Microbiol.* **12**, 172 (2012).
120. Bi, D. *et al.* SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.* **41**, D660 (2013).
121. Li, J. *et al.* SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environ. Microbiol.* **17**, 2196 (2015).
122. Pundhir, S. & Kumar, A. SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems. *Bioinformatics* **6**, 380 (2011).
123. Wang, Y., Huang, H., Sun, M., Zhang, Q. & Guo, D. T3DB: an integrated database for bacterial type III secretion system. *BMC Bioinformatics* **13**, 66 (2012).

## Acknowledgements

We are grateful to Elie Dassa, Olivera Francetic, and Marc Garcia-Garcerà for fruitful discussions, and Hervé Ménager for its contribution to MacSyView. We thank Eric Duclaud for discussions and comments on T9SS. This work was supported by the CNRS, the Institut Pasteur and the European Research Council (grant EVOMOBILOME, number 281605). JC is a member of the French doctoral school “Ecole Doctorale Interdisciplinaire Européenne Frontières du Vivant ED474”.

### Author Contributions

S.S.A and E.P.C.R. designed the analyses. S.S.A. designed the secretion systems models and profiles, and performed the analyses. J.C. and J.G. contributed to the T4SS models, the T4SS protein profiles and the corresponding analyses. B.N. contributed to the MacSyFinder and MacSyView software, and created the online interface for TXSScan, MacSyView. J.G. and B.N. created the online interface for the TXSSdb database. MT contributed to the T9SS models and the T9SS protein profiles. S.S.A. and E.P.C.R. wrote the manuscript with the help of the other authors. All authors read and approved the full manuscript.

### Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Abby, S. S. *et al.* Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080; doi: 10.1038/srep23080 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

## **Co-authored Manuscript 3: Perrin et al, 2017**

This work led to the identification of an ICE inserted in a DNA repair gene. I identified the insertion sequence manually and proposed a model of integration and excision for this ICE (Figure 4).

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

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.

<sup>1</sup>Institut Pasteur, Microbial Evolutionary Genomics, F-75724 Paris, France. <sup>2</sup>CNRS, UMR 3525, F-75724 Paris, France. <sup>3</sup>Institut Pasteur, Hub Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, F-75724 Paris, France. <sup>4</sup>CNRS, UMS 3601 IFR-Core, F-91198 Gif-sur-Yvette, France. <sup>5</sup>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. <sup>6</sup>Centre for Systems Genomics, University of Melbourne, Parkville, Victoria 3010, Australia. <sup>7</sup>Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia. <sup>8</sup>Wisconsin State Laboratory of Hygiene, Madison, Wisconsin 53718, USA. <sup>9</sup>CIP—Collection de l'Institut Pasteur, Institut Pasteur, F-75724 Paris, France. <sup>10</sup>Institut Pasteur, Pasteur International Bioresources network (PIBnet), Plateforme de Microbiologie Mutualisée (P2M), F-75724 Paris, France. <sup>11</sup>Division of Scientific Resources, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. <sup>12</sup>Epidemic Intelligence Service, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. <sup>13</sup>Division of Public Health, Wisconsin Department of Health Services, Madison, Wisconsin 53701, USA. <sup>14</sup>Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia 30329, USA. <sup>15</sup>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<sup>1</sup>. Despite recent genomic and experimental work<sup>2–6</sup>, 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<sup>7–9</sup> although sporadic, community-acquired cases have been occasionally reported<sup>10</sup>, as has a single instance of transmission of *E. anophelis* from mother to infant at birth<sup>11</sup>. Human infections have varied presentations, including meningitis and septicemia<sup>12–15</sup>. Strains have been isolated from diverse environments such as hospital sinks (*E. meningoseptica* and *E. anophelis*)<sup>6,7</sup>, the mosquito mid-gut (*E. anophelis*)<sup>1</sup> and the space station Mir (*E. miricola*)<sup>16</sup>. 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<sup>1</sup>, 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<sup>2,4,6,17,18</sup>. *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<sup>15</sup>.

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 (ICEEa1) 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<sup>19</sup>: 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<sup>T</sup> of *E. endophytica* was placed within the *E. anophelis* branch, consistent with a recent report<sup>20</sup>. 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<sup>20</sup>.

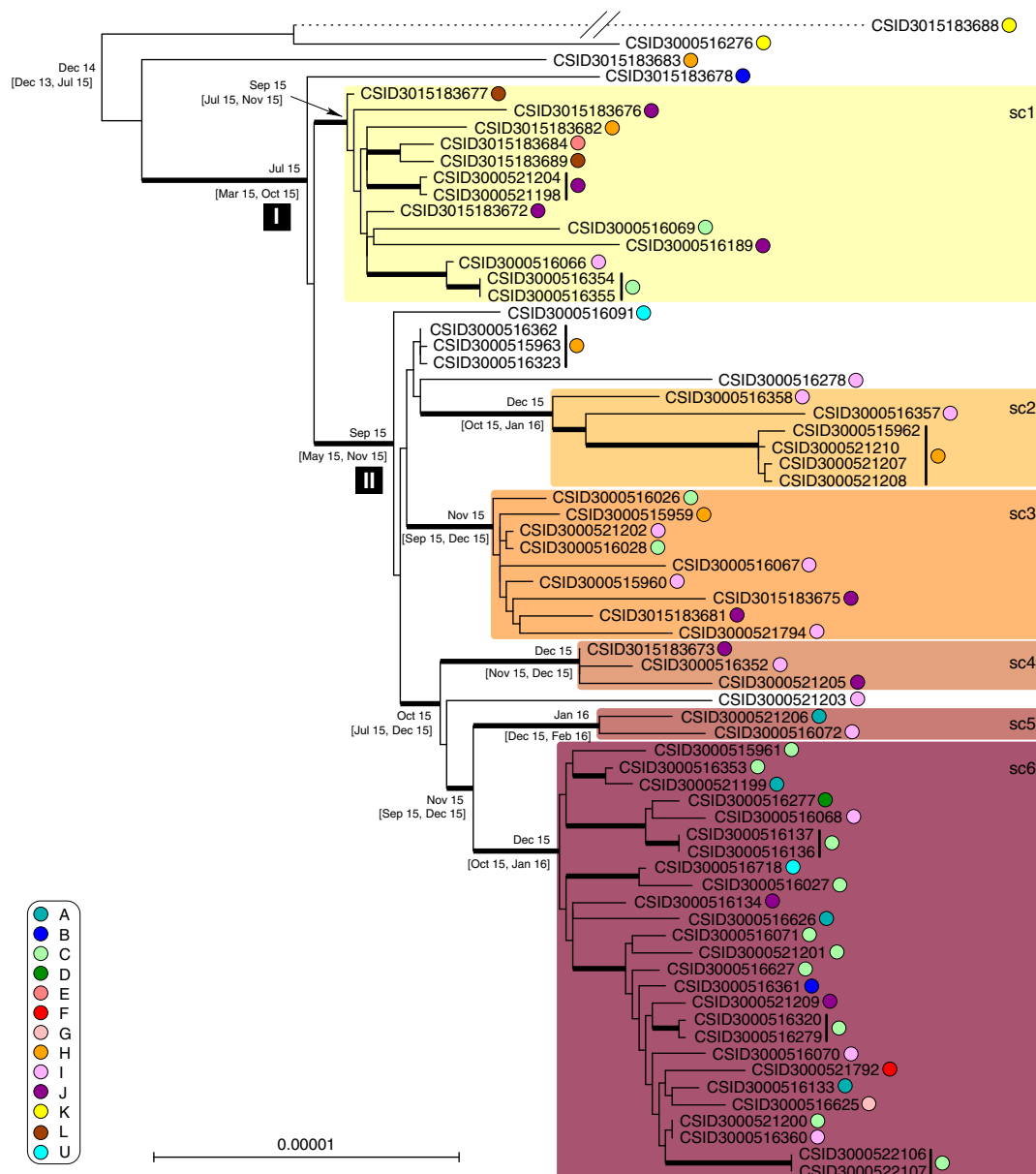
The outbreak isolates made up a compact phylogenetic group within *E. anophelis* (sublineage 15 in Supplementary Fig. 1b), indicating that the outbreak was caused by a single ancestral strain. The long branch that separated the outbreak strain from all other sequenced *E. anophelis* strains showed that 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 ~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  $5.98 \times 10^{-6}$  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

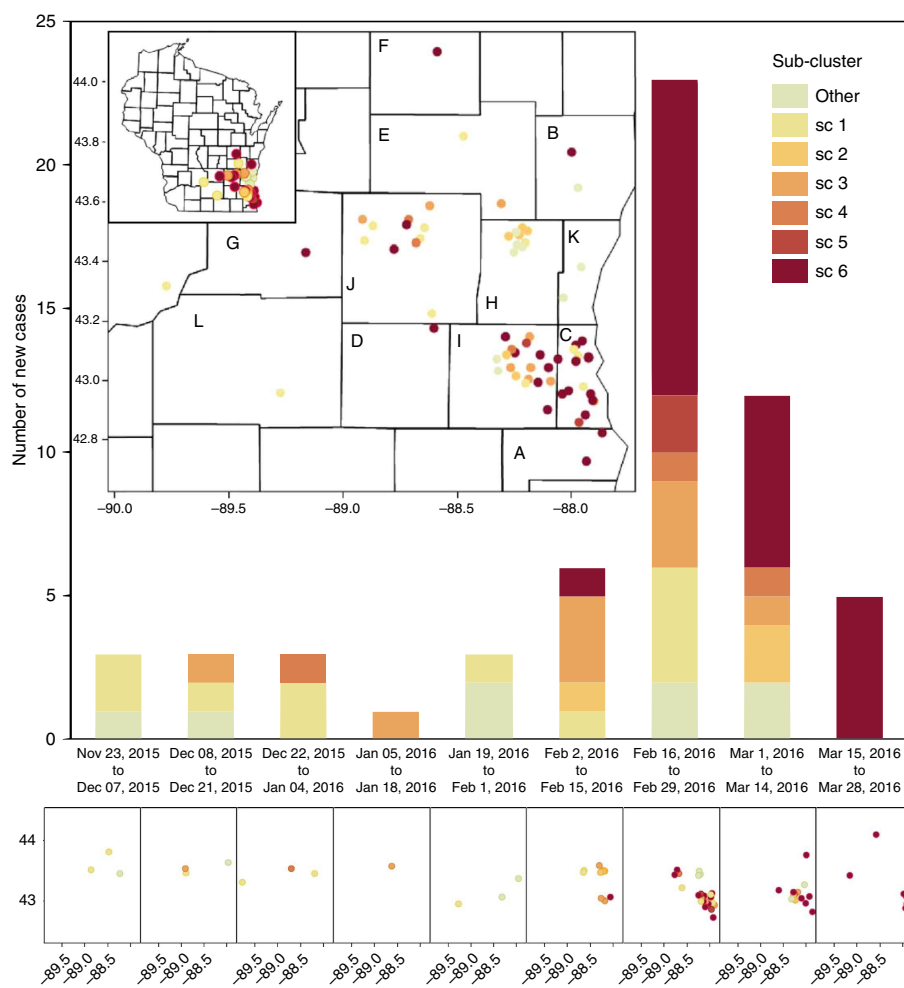


**Figure 1 | Phylogenetic tree of the outbreak isolates.** Maximum likelihood phylogenetic tree inferred from 3,411,033 aligned nucleotide characters (1,137,011 codons) based on cgMLST data. The tree was rooted based on phylogenetic analyses using epidemiologically unrelated *E. anophelis* strains as 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.

$6.35 \times 10^{-6}$  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.

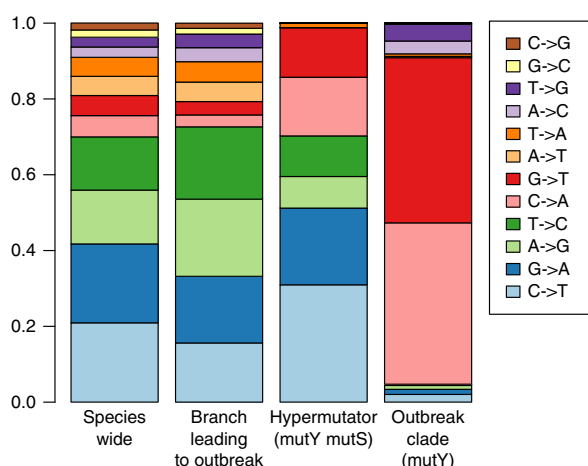
corner of the 12 county outbreak region during the outbreak peak and followed by a wider geographic spread of sc6 after March 1. This is consistent with the relative branching order and estimated ages of sc1 and sc6 inferred from the phylogenetic analysis of genomic sequences (Fig. 1). The fit between 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 ( $5.98 \times 10^{-6}$  substitutions per site per year within core genes, and  $6.35 \times 10^{-6}$  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





**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.

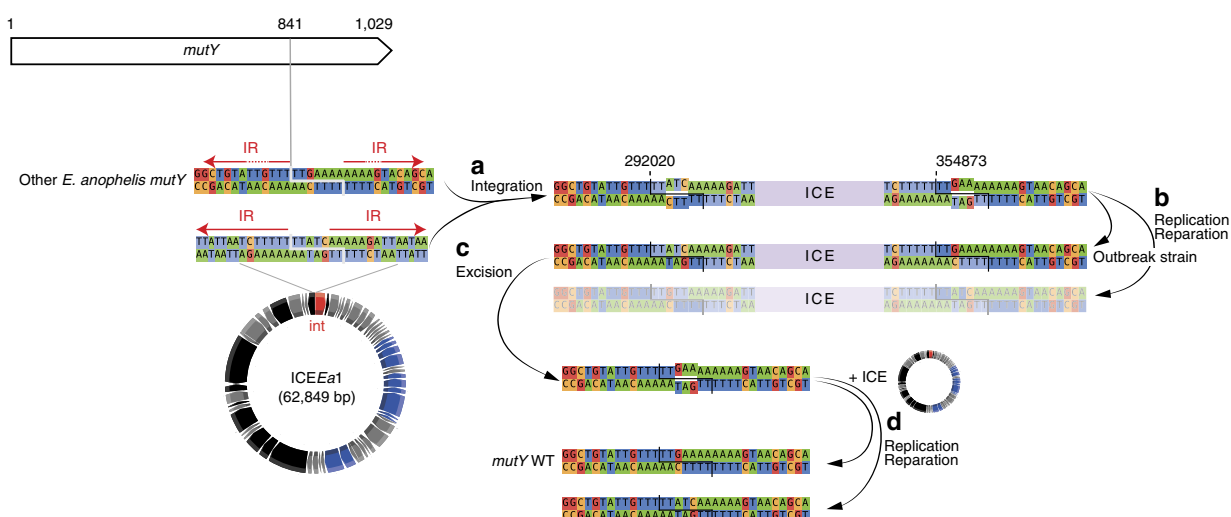
these lesions<sup>21–24</sup>. We, therefore, inspected the genes that pertain to the GO system, and found that *mutY* was interrupted at position 841 in all outbreak isolates by the insertion of the 62,849 bp Integrative and Conjugative Element ICEEa1 (for integrative and conjugative element 1 of *E. 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<sup>25</sup>. 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 ICEEa1 occurred shortly before the last common ancestor of the outbreak isolates.

ICEEa1’s integrase is 64% similar to the integrase of CTnDOT, a well-studied ICE<sup>26,27</sup>. We identified a potential integration site (TTT<sup>^</sup>TT) at position 841 of the *mutY* gene, flanked by inverted repeats in the ICEEa1 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 ICEEa1 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<sup>28</sup>.

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 ICEEa1 can lead to *mutY* WT in outbreak strains.** Here the insertion site is TTT<sup>^</sup>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 scenarios.

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<sup>29,30</sup>. 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*<sup>2,4,6,17</sup>, 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).

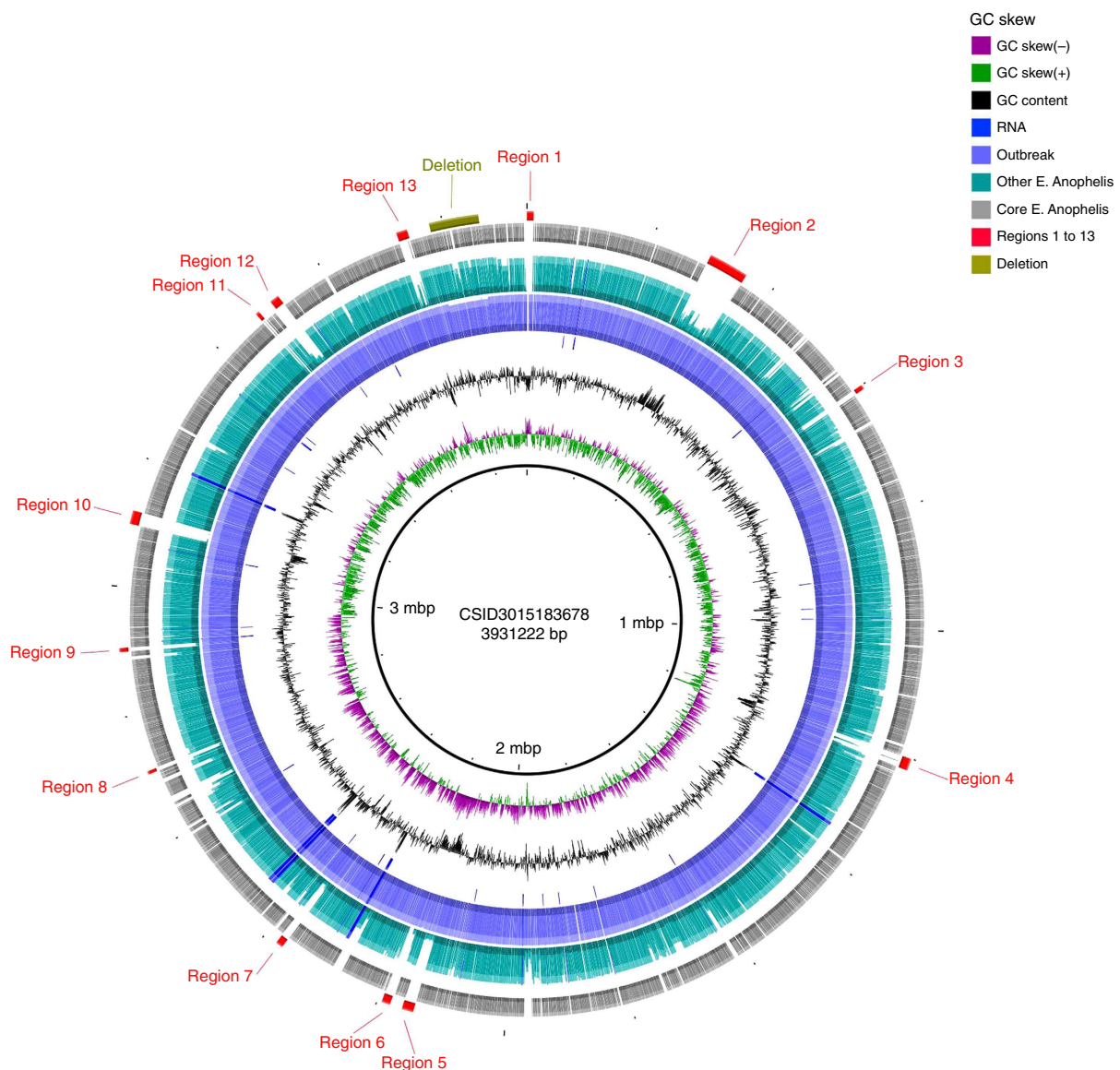
A search for putative virulence genes led to the identification of 67 genes (Supplementary Data 5). Among these, genes that were highly associated to the outbreak strain as compared with other *E. anophelis* isolates, included a CobQ/CobB/MinD/ParA nucleotide-binding domain protein located on the ICEEa1 element (see below) and five genes involved in capsular polysaccharide synthesis. Capsules are important virulence factors of bacterial pathogens<sup>31</sup>. 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<sup>2</sup>, 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<sup>2,6</sup>. 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<sup>2,6</sup> 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 ICEEa1 was present in all outbreak isolates but was absent in most other *E. anophelis* strains (region 2 in Fig. 5 and Supplementary Fig. 6). ICEEa1 belongs to the *Bacteroidetes* type 4 secretion system (T4SS-B) class<sup>32</sup>. 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), ICEEa1 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 ICEEa1 element to detoxification of divalent cations and to acquire iron from the host during infection. Within the wider *E. anophelis* genome set, the ICEEa1 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 ICEEa1 with virulence deserves further functional investigation. In the six other strains, the ICEEa1 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)<sup>11</sup>.



**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)<sup>73</sup>. 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<sup>6,14,15,33</sup>. 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         | ICEEa1; 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<sup>34,35</sup>. 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<sup>15</sup>. This further indicates that the clinical importance of *E. anophelis* was previously underestimated, in agreement with results of a recent study<sup>15</sup>. 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<sup>36,37</sup>.

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<sup>37,38</sup>. 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 *Wz* 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<sup>39</sup>. Therefore, the integration of the *ICEEa1* 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<sup>40,41</sup>. Based on the structure of the integration site, we hypothesize that the outbreak strain could revert to a functional *mutY* sequence by losing the *ICEEa1* 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*<sup>20</sup>, 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 *NcoI* optical map to generate a complete circularized genome, which was confirmed based on a PacBio genome sequence<sup>42</sup>. 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<sup>43</sup>. 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<sup>2</sup> 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 Wisconsin 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<sup>49</sup>. Allele sequences and their corresponding numerical designations are publicly accessible at <http://bigfdb.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<sup>51</sup> with the codon evolutionary model being selected to minimize the BIC criterion, that is, GY + F +  $\Gamma_4$  and GY + F<sup>52</sup> 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>)<sup>53</sup>. 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<sup>54</sup> 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 Gng<sup>56</sup>. A ~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<sup>51</sup> (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<sup>59–61</sup> 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<sup>2</sup>. 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<sup>64</sup> (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<sup>67</sup>. CRISPR-Cas systems were searched using MacSyFinder v.1.0.2 with Cas-Finder profiles<sup>66</sup> and CRISPR-Finder<sup>68</sup>, with default parameters. Integrons were searched using IntegronFinder v.1.4 with -local\_max option<sup>69</sup>, and prophages using VirSorter v.1.0.3 on RefSeqDB only<sup>70</sup> 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](http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_6.0_Breakpoint_table.pdf))<sup>72</sup>. 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, cephalixin, cefuroxime, cefoxitin, cefepime, ceftoperazone-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, fosfomicin, 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://bigfdb.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.e.3674146.v5>). We also created a project on microreact, available at this link: <https://microreact.org/project/SyaeGCjvg>.

## References

- 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).
- Breurec, S. *et al.* Genomic epidemiology and global diversity of the emerging bacterial pathogen *Elizabethkingia anophelis*. *Sci. Rep.* **6**, 30379 (2016).
- Chen, S., Bagdasarian, M. & Walker, E. D. *Elizabethkingia anophelis*: molecular manipulation and interactions with mosquito hosts. *Appl. Environ. Microbiol.* **81**, 2233–2243 (2015).
- Kukutla, P. *et al.* Insights from the genome annotation of *Elizabethkingia anophelis* from the malaria vector *Anopheles gambiae*. *PLoS ONE* **9**, e97715 (2014).
- 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).
- Teo, J. *et al.* Comparative genomic analysis of malaria mosquito vector-associated novel pathogen *Elizabethkingia anophelis*. *Genome Biol. Evol.* **6**, 1158–1165 (2014).
- Moore, L. S. P. *et al.* Waterborne *Elizabethkingia meningoseptica* in adult critical care. *Emerg. Infect. Dis.* **22**, 9–17 (2016).
- 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).
- 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).
- Hayek, S. S. *et al.* Rare *Elizabethkingia meningosepticum* meningitis case in an immunocompetent adult. *Emerg. Microbes Infect.* **2**, e17 (2013).
- Lau, S. K. P. *et al.* Evidence for *Elizabethkingia anophelis* transmission from mother to infant, Hong Kong. *Emerg. Infect. Dis.* **21**, 232–241 (2015).
- King, E. O. Studies on a group of previously unclassified bacteria associated with meningitis in infants. *Am. J. Clin. Pathol.* **31**, 241–247 (1959).
- 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).
- Frank, T. *et al.* First case of *Elizabethkingia anophelis* meningitis in the Central African Republic. *Lancet (London, England)* **381**, 1876 (2013).
- Lau, S. K. P. *et al.* *Elizabethkingia anophelis* bacteremia is associated with clinically significant infections and high mortality. *Sci. Rep.* **6**, 26045 (2016).
- 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).
- 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).
- González, L. J. & Vila, A. J. Carbapenem resistance in *Elizabethkingia meningoseptica* is mediated by metallo- $\beta$ -lactamase BlaB. *Antimicrob. Agents Chemother.* **56**, 1686–1692 (2012).
- 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).
- 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).
- 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).
- Boiteux, S. & Radicella, J. P. Base excision repair of 8-hydroxyguanine protects DNA from endogenous oxidative stress. *Biochimie* **81**, 59–67 (1999).
- 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).
- David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447**, 941–950 (2007).
- 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).
- Malanowska, K., Salyers, A. A. & Gardner, J. F. Characterization of a conjugative transposon integrase, IntDOT. *Mol. Microbiol.* **60**, 1228–1240 (2006).
- 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).
- 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).
- Mackenzie, A. K. *et al.* Two SusD-like proteins encoded within a polysaccharide utilization locus of an uncultured ruminant bacteroidetes phylogroup bind strongly to cellulose. *Appl. Environ. Microbiol.* **78**, 5935–5937 (2012).
- 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).
- Moxon, E. R. & Kroll, J. S. The role of bacterial polysaccharide capsules as virulence factors. *Curr. Top. Microbiol. Immunol.* **150**, 65–85 (1990).
- 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).
- 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).
- Bialek-Davenet, S. *et al.* Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. *Emerg. Infect. Dis.* **20**, 1812–1820 (2014).
- 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).
- Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
- 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).
- Zhou, Z. *et al.* Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* serovar Agona. *PLoS Genet.* **9**, e1003471 (2013).
- Oliver, A. & Mena, A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin. Microbiol. Infect.* **16**, 798–808 (2010).
- Denamur, E. & Matic, I. Evolution of mutation rates in bacteria. *Mol. Microbiol.* **60**, 820–827 (2006).
- Söderberg, R. J. & Berg, O. G. Kick-starting the ratchet: the fate of mutators in an asexual population. *Genetics* **187**, 1129–1137 (2011).
- Nicholson, A. C. *et al.* Complete genome sequences of four strains from the 2015–2016 *Elizabethkingia anophelis* outbreak. *Genome Announc.* **4**, e00563–16 (2016).
- 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).
- Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
- Crisuolo, 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).
- Liu, Y., Schroder, J. & Schmidt, B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**, 308–315 (2013).
- Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res.* **4**, 900 (2015).
- Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
- Jolley, K. A. & Maiden, M. C. J. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 1–11 (2010).
- Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 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).
- Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
- Schultz, M. B. *et al.* Repeated local emergence of carbapenem-resistant *Acinetobacter baumannii* in a single hospital ward. *Microb. Genom.* **2**, e000050 (2016).
- Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
- Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
- 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](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-IBED), 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 ICEEa1 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.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017



## **Co-authored Manuscript 4: Oliveira et al, 2017**



I provided integrons and conjugative elements data for this work.

ARTICLE

DOI: 10.1038/s41467-017-00808-w

OPEN

# The chromosomal organization of horizontal gene transfer in bacteria

Pedro H. Oliveira <sup>1,2</sup>, Marie Touchon<sup>1,2</sup>, Jean Cury <sup>1,2</sup> & Eduardo P.C. Rocha<sup>1,2</sup>

Bacterial adaptation is accelerated by the acquisition of novel traits through horizontal gene transfer, but the integration of these genes affects genome organization. We found that transferred genes are concentrated in only ~1% of the chromosomal regions (hotspots) in 80 bacterial species. This concentration increases with genome size and with the rate of transfer. Hotspots diversify by rapid gene turnover; their chromosomal distribution depends on local contexts (neighboring core genes), and content in mobile genetic elements. Hotspots concentrate most changes in gene repertoires, reduce the trade-off between genome diversification and organization, and should be treasure troves of strain-specific adaptive genes. Most mobile genetic elements and antibiotic resistance genes are in hotspots, but many hotspots lack recognizable mobile genetic elements and exhibit frequent homologous recombination at flanking core genes. Overrepresentation of hotspots with fewer mobile genetic elements in naturally transformable bacteria suggests that homologous recombination and horizontal gene transfer are tightly linked in genome evolution.

<sup>1</sup> Microbial Evolutionary Genomics, Institut Pasteur, 25–28 rue du Docteur Roux, Paris 75015, France. <sup>2</sup> CNRS, UMR3525, 25–28 rue du Docteur Roux, Paris 75015, France. Pedro H. Oliveira and Marie Touchon contributed equally to this work. Correspondence and requests for materials should be addressed to P.H.O. (email: [pcphco@gmail.com](mailto:pcphco@gmail.com)) or to M.T. (email: [mtouchon@pasteur.fr](mailto:mtouchon@pasteur.fr))

The gene repertoires of bacterial species are often very diverse, which is central to bacterial adaptation to changing environments, new ecological niches, and co-evolving eukaryotic hosts<sup>1</sup>. Novel genes arise in bacterial genomes mostly by horizontal gene transfer (HGT)<sup>2</sup>, a pervasive evolutionary process that spreads genes between, eventually very distant, bacterial lineages<sup>3</sup>. It is commonly thought that the majority of genes acquired by HGT are neutral or deleterious and thus rapidly lost<sup>4</sup>. Yet, HGT is also responsible for the acquisition of many adaptive traits, including antibiotic resistance in nosocomials<sup>5</sup>. Hence, genome diversification is shaped by the balancing processes of gene acquisition and loss<sup>6</sup>, moderated by positive selection on some genes, and purifying selection on many others<sup>7</sup>.

Chromosomes are organized to favor the interactions of DNA with the cellular machinery<sup>8</sup>. For example, most bacterial genes are co-transcribed in operons, leading to strong and highly conserved genetic linkage between neighboring genes<sup>9</sup>. At a more global level, early-replicating regions are enriched in highly expressed genes in fast-growing bacteria to enjoy replication-associated gene dosage, creating a negative gradient of expression along the axis from the origin (ori) to the terminus (ter) of replication (ori->ter)<sup>10, 11</sup>. These organizational traits can be disrupted by the integration of novel genetic information. At a local level, new genes rarely integrate within an operon and, instead, they tend to be incorporated at its edges, where they are less likely to affect gene expression<sup>12</sup>. At the genome level, the frequency of integration of prophages in the genome of *Escherichia coli* increases along the ori->ter axis<sup>13</sup>. The results of these studies suggest that the fitness effects of HGT in terms of chromosome organization depend on the specific site of integration.

In prokaryotes, HGT takes place by three main mechanisms: natural transformation, conjugation, and transduction. Mobile genetic elements (MGEs) play a key role in HGT because they are responsible for the latter two processes, respectively by the activity of conjugative elements and phages<sup>14</sup>. Integrative conjugative elements (ICEs) and prophages are large genetic elements that may account for a significant fraction of the bacterial genome<sup>15, 16</sup>, and bring to the chromosome many genes in a single event of integration. For example, some strains of *E. coli* have up to 18 prophages<sup>17</sup>, and *Mesorhizobium loti*

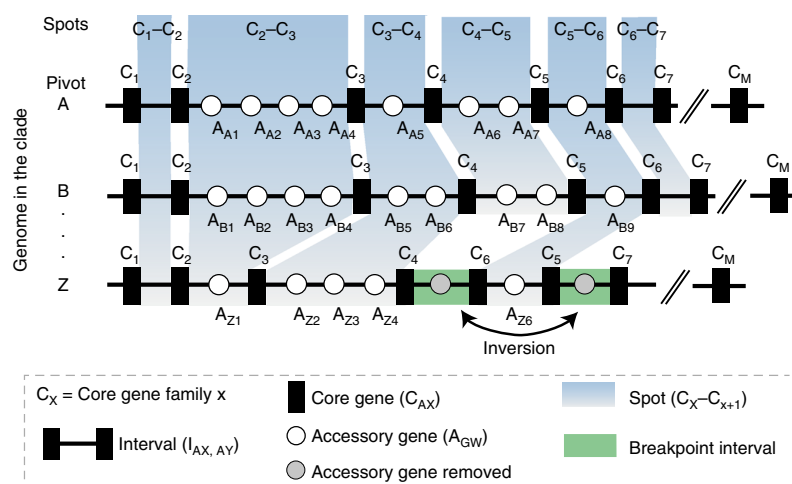
encodes one ~500 kb ICE<sup>18</sup>. The integration of these large MGEs changes the chromosome size and may split adaptive genetic structures such as operons. This might contribute to explain why most integrative MGEs use site-specific recombinases (integrases) that target very specific sites in the chromosome<sup>19</sup>. Integrases and MGEs have co-evolved with the host genome to decrease the fitness cost of their integration<sup>13</sup>.

MGEs carrying similar integrases tend to integrate at the same sites in the chromosome, leading to regions with unexpectedly high frequency of MGEs at homologous regions. This concentration of MGEs in few sites has been frequently described<sup>20, 21</sup>, especially in relation to the presence of neighboring tRNA and tmRNA genes<sup>22</sup>. Yet, a previous work described the existence of regions with high rates of diversification in *E. coli* (hotspots), some of which lacked recognizable integrases<sup>23</sup>. In particular, the genes flanking two hotspots were associated with high rates of homologous recombination (*rfa* and *leuX*). In *Streptococcus pneumoniae*, the chromosomal genes flanking MGEs also showed higher rates of homologous recombination<sup>24, 25</sup>. In this species, it was suggested that integration of MGEs close to core genes under selection for diversification could be adaptive by facilitating the transfer and subsequent recombination of the latter<sup>26</sup>.

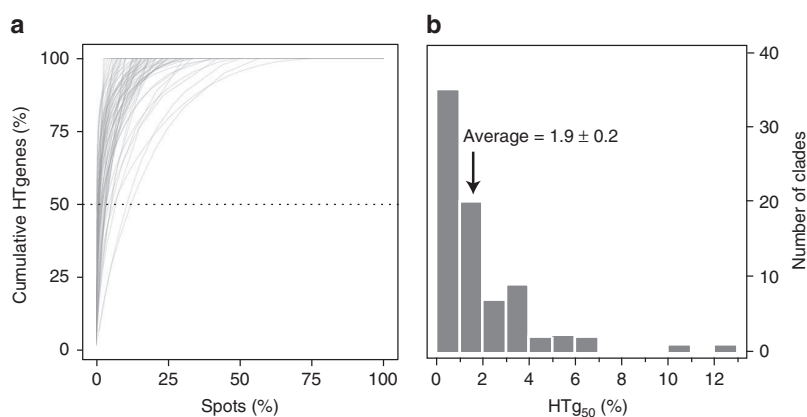
Here, we define and identify hotspots in a large and diverse panel of bacterial species and show how they reflect the mechanisms driving genome diversification by HGT.

## Results

**Quantification of HGT and definition of hotspots.** To study the distribution of gene families in bacterial chromosomes, we analyzed 932 complete genomes of 80 bacterial species (Supplementary Data set 1). We inferred the core genome, the pan-genome, the accessory genome (genes from the pan-genome absent from the core), and the phylogeny of each species, as before<sup>27</sup> (Methods, Supplementary Figs. 1 and 2). We partitioned the genomes into an array of core genes and intervals (Fig. 1, Table 1). The latter were defined as the positions between consecutive core genes. We defined a spot as the set of intervals delimited by members of the same two families of core genes in the genomes of the clade (see Methods for rigorous definitions, Supplementary Fig. 2a, b). We observed that 99.4% of the intervals were part of the species' spots and only 0.6% were in



**Fig. 1** Scheme depicting key concepts used in this study. Intervals flanked by the same core gene families ( $C_x$ ,  $C_y$ ) as those from pivot genome A were defined as syntenic intervals (i.e., the members of the core gene families X and Y were also contiguous in the pivot). The intervals that do not satisfy this constraint were classed as breakpoint intervals (green-shaded regions) and excluded from our analysis. For every interval in the pivot genome, we defined spot as the set of intervals flanked by members of the same core gene families (blue-shaded regions)



**Fig. 2** Cumulative frequency of HTgenes. **a** Cumulative distribution of horizontally transferred genes (HTgenes, %) in spots for the 80 bacterial clades. **b** Histogram of the minimum number of spots needed to attain 50% of the total number of HTgenes (HT<sub>g50</sub> index). The average HT<sub>g50</sub> was only 1.9% ( $\pm 0.2$ ; standard deviation)

**Table 1** Acronyms used in this study

|                   |  |
|-------------------|--|
| MGE               | Mobile genetic element (i.e., prophage, ICE, IME and integron)     |
| ICE               | Integrative conjugative element                                    |
| IME               | Integrative mobilizable element                                    |
| MAP               | Mobility-associated protein (i.e., integrase and transposase (IS)) |
| ARG               | Antibiotic resistance gene   |
| HGT               | Horizontal gene transfer   |
| HTgenes           | Genes having been horizontally transferred                         |
| HT <sub>g50</sub> | Number of spots required to include 50% of HTgenes                 |
| T <sub>95%</sub>  | Minimal number of HTgenes required to define a hotspot             |

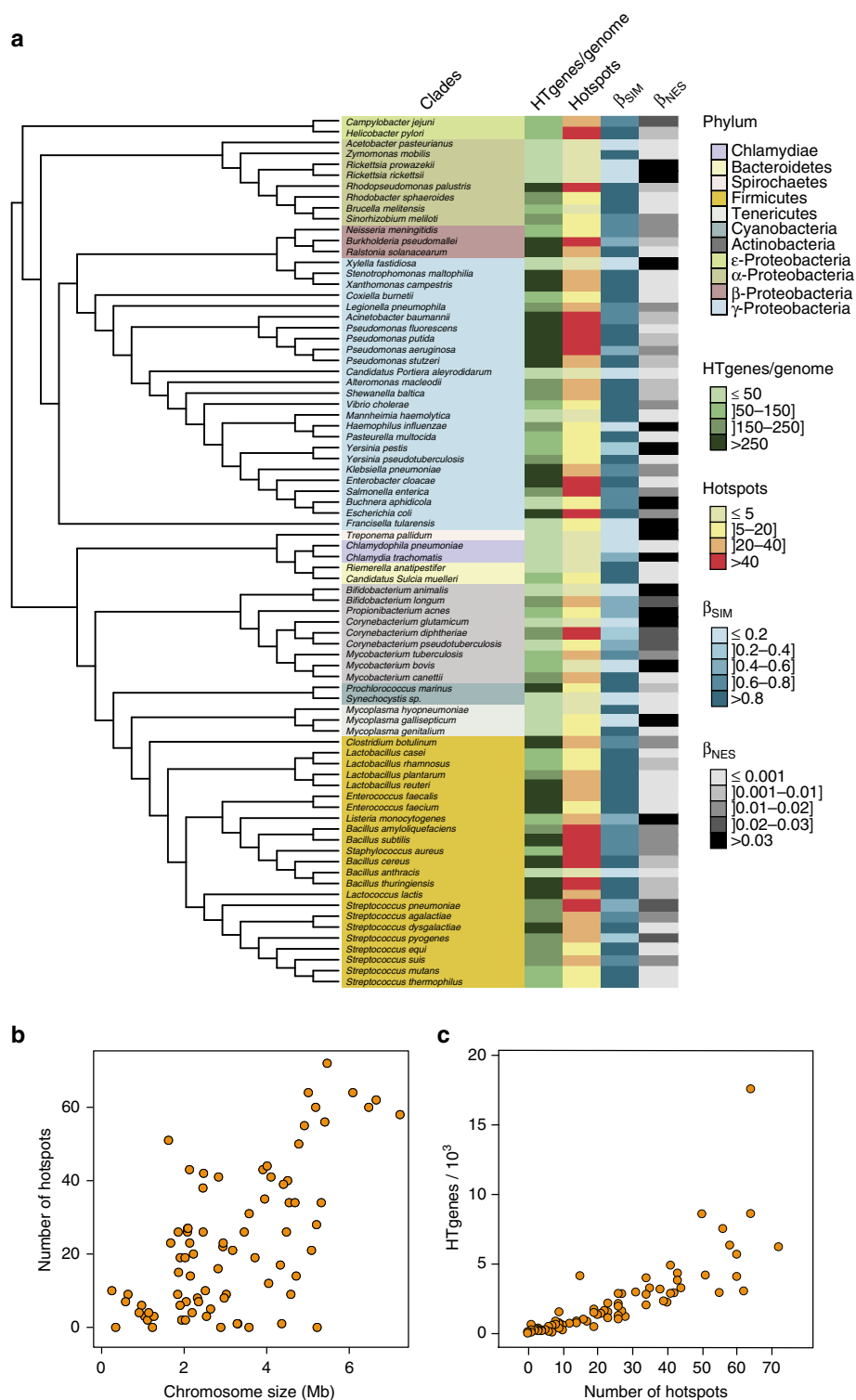
breakpoint intervals. Since 99.8% of spots are flanked by the same two families of core genes in at least half of the genomes of each clade (and 99% in all genomes), it is most parsimonious to consider that the two core genes were already contiguous in the last common ancestor of the clade. Hence, we split the pan-genomes in spot pan-genomes, i.e., sets of gene families located in each spot (Methods, Supplementary Fig. 2). The genes outside spots, i.e., in intervals that were split by events of rearrangement, accounted for < 2% of the total number of genes and were discarded from further analysis.

We used birth-and-death models to identify HGT events in the clade's phylogenetic trees from the patterns of presence/absence of each gene family (Methods). Note that HGT events are defined gene per gene (which will be called HTgenes for Horizontally Transferred Genes), not as blocks, because there are no tools available for the latter and because the goal of our work was to study the clustering of genes acquired by HGT without using a priori models. Spots contained 170,041 HTgenes (15.5% of the total number of accessory genes). We quantified the clustering of these genes by counting the minimal number of spots required to accumulate at least 50% of the HTgenes (HT<sub>g50</sub>) (Fig. 2a). The distribution of these values was skewed toward small values (Fig. 2b). Hence, < 2% of the largest hotspots accumulate >50% of all HTgenes. Conversely, 72.6% of the spots were on average empty, i.e., had no accessory gene in any genome. Similar qualitative conclusions were obtained in the analysis of the distribution of all accessory genes, despite the latter being slightly less clustered (Supplementary Fig. 3). These results show that most HTgenes are integrated in a very small number of sites in the genome.

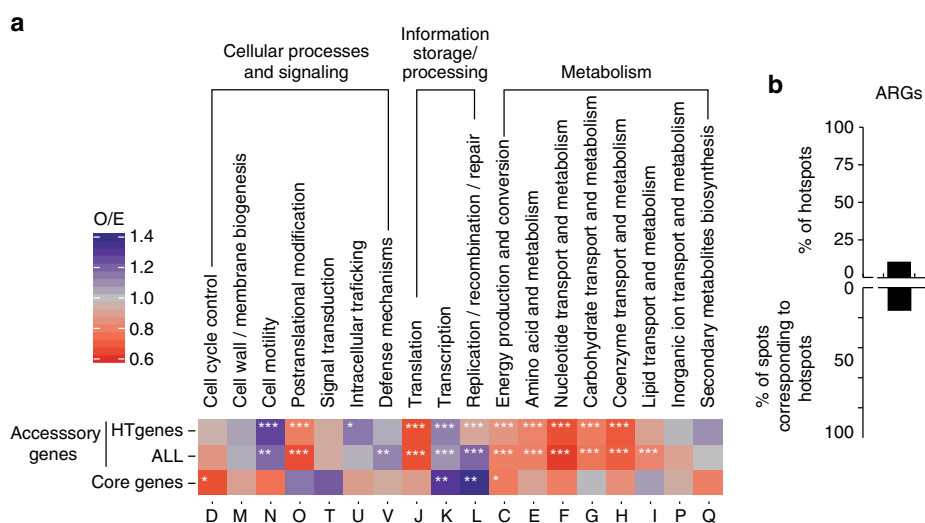
We used simulations to infer the statistical thresholds for the degree of clustering of HTgenes in each clade (Methods, Supplementary Fig. 4). We made the null hypothesis that these genes are organized in operons like the other genes, and are uniformly distributed among spots. We identified the spot with the highest number of HTgenes in each simulation ( $\text{Max}_{\text{HTg},i}$ ), and computed the 95th percentile of the distribution of these maximal values ( $T_{95\%}$ , Supplementary Data set 1). Simulations disregarding the existence of operons produced lower values of  $T_{95\%}$  showing the importance of incorporating information about genetic organization in the model (Supplementary Fig. 5). Spots with more than  $T_{95\%}$  HTgenes were called hotspots, spots lacking accessory genes were called empty, and the others were called coldspots. We found a total of 1841 hotspots in the 80 clades (Supplementary Data set 1). They represent only 1.2% of the spots, but they concentrate 47% of the accessory gene families and 60% of the HTgenes.

The number of hotspots differed widely among clades, from none or very few in *Acetobacter pasteurianus*, *Bacillus anthracis*, and the obligatory endosymbionts, to more than 60 in *Bacillus thuringiensis*, *E. coli*, and *Pseudomonas putida* (Fig. 3a). This variance was partly a function of chromosome size (Fig. 3b), but was especially associated with the number of HTgenes (Fig. 3c). Increases in the latter resulted in a less-than-linearly increase in the number of hotspots and in a linear increase in hotspot density per Mb (Supplementary Fig. 6). Hence, a few hotspots aggregate most of the genes acquired by horizontal transfer and this trend is more pronounced when the rates of transfer are high.

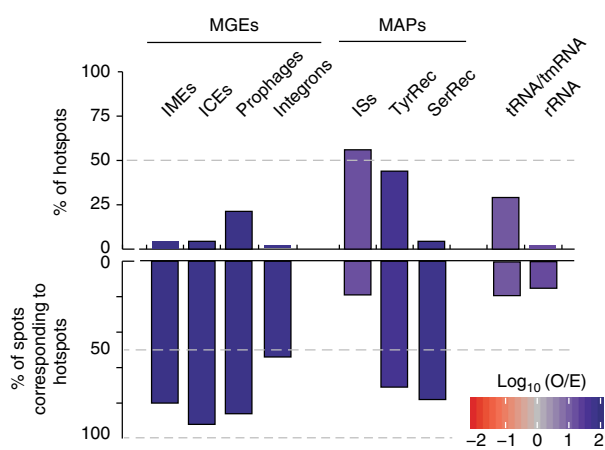
**Functional and genetic characterization of hotspots.** We investigated the function of the genes in the spots, using the eggNOG categories, to assess if hotspots were enriched in particular traits (Methods, Fig. 4a). Genes classified as poorly characterized or as having an unknown function were not considered in the subsequent functional analyses (they were 13.1% of the total). We then compared the distribution of the functions of all accessory genes and that of HTgenes in hotspots relative to coldspots. Both analyses showed an underrepresentation of translation and post-translational modification genes in hotspots. These genes tend to be essential and are less frequently transferred horizontally<sup>28</sup>. In contrast, hotspots overrepresented genes associated with cell motility, defense mechanisms, transcription, and replication and repair. Moreover, around 9% of the hotspots encoded antibiotic



**Fig. 3** Analysis of HTgenes and the abundance and distribution of hotspots. **a** 16S rRNA phylogenetic tree of the 80 bacterial clades. The tree was drawn using the iTOL server ([itol.embl.de/index.shtml](http://itol.embl.de/index.shtml))<sup>70</sup>. The first *column* indicates the clade and is colored by phylum. The four subsequent *columns* correspond respectively to: the average number of HTgenes per genome computed using Count, the number of hotspots, the average Simpson dissimilarity index ( $\beta_{SIM}$ , accounting for turnover), and the average multiple-site dissimilarity index accounting only for nestedness ( $\beta_{NES}$ ). These values are given in Supplementary Data set 1. **b** Distribution of the average number of hotspots per clade according to the average genome size ( $G_S$ ). **c** Association between the number of hotspots and the number of HTgenes in the clade



**Fig. 4** Functional characterization of hotspots. **a** Observed/expected (O/E) ratios of non-supervised orthologous groups (NOGs, shown as *capitalized letters*). The first two lines represent the values of HTgenes and accessory genes observed in hotspots when the null model was computed from the distribution of the same type of genes in coldspots. The last line shows the same type of analysis for the core genes flanking hotspots when the null model is computed using the core genes not flanking hotspots. Expected values were obtained by multiplying the number of HTgenes, accessory, or core genes in hotspots by the fraction of genes assigned to each NOG. \* $P < 0.05$ ; \*\* $P < 10^{-2}$ ; \*\*\* $P < 10^{-3}$ ,  $\chi^2$ -test. **b** Percentage of hotspots with antibiotic resistance genes (ARGs, *top*), and percentage of spots with ARGs that are hotspots (*bottom*). Note that hotspots are only 1.2% of all the spots



**Fig. 5** Genetic mobility of hotspots. We represent the percentage of hotspots containing the different genetic elements (*top*) and the percentage of spots containing such elements that are hotspots (*bottom*). Note that hotspots are only 1.2% of all the spots. The analysis includes MGEs (IMEs, ICEs, prophages, integrations), mobility-associated proteins (MAPs) (ISs, TyrRec, SerRec), and tRNA/tmRNA, rRNA. Also, shown in *colored bins* are the observed/expected ( $\log_{10}O/E$ ) number of hotspots that contain the abovementioned elements, when the null model was computed from the distribution of coldspots containing the same type of elements. Expected values were obtained by multiplying the number of hotspots by the fraction of spots containing each type of element

resistance genes (ARGs), which is much more than expected by chance (0.8%) (Fig. 4b).

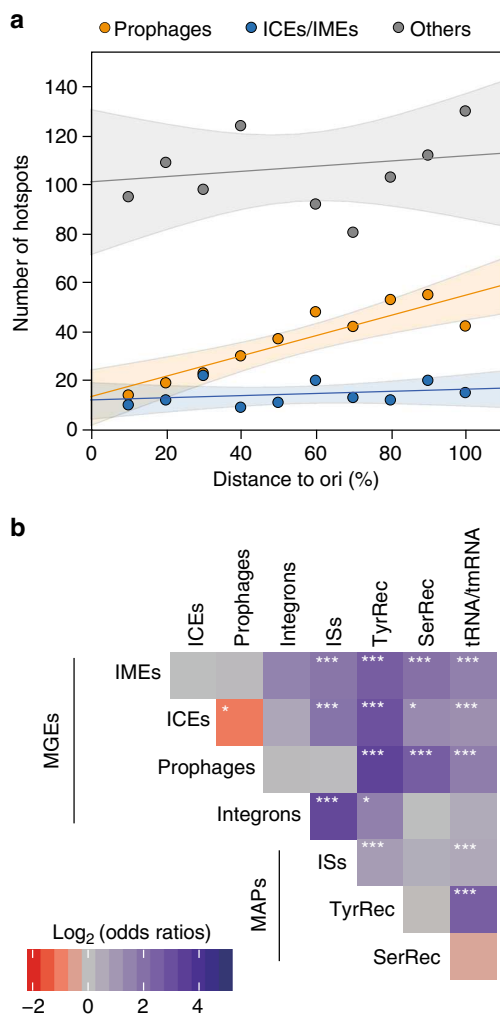
Some of the functions overrepresented in hotspots—defense, replication, repair—are typically found in MGEs, which concentrate in specific loci targeted by integrases (often at tRNAs). Accordingly, the vast majority of self-mobilizable MGEs—89% of the prophages and 90% of ICEs—were identified in hotspots (Supplementary Data set 3 and Supplementary Fig. 7). On the

other hand, only around 9% of the hotspots encoded ICEs or integrative mobilizable elements (IMEs), and only 23% encoded prophages (Fig. 5). Integrations were even rarer (present in 1% of the hotspots). Non-self-transferable MGEs lack conjugation or virion structural genes, but usually encode integrases. The vast majority of integrases was identified in hotspots, but less than half (45%) of the hotspots encoded an integrase and only 29% encoded tRNA or tmRNA genes (Fig. 5). Hence, although most self-mobilizable MGEs are in hotspots, most hotspots lack them (Supplementary Fig. 8).

Insertion sequences (ISs) encoding DDE recombinases (transposases) are frequent within MGEs, and we found them in many hotspots (56%). The integration of these elements has low-sequence specificity, which explains why hotspots accounted for a small fraction of the locations with ISs (19%), unlike what we observed for self-mobilizable MGEs and integrases. Altogether, half of the hotspots lacked evidence for the presence of MGEs and 27% lacked any of the mobility-associated proteins (MAPs, integrases and transposases) that we searched for. These results confirm that hotspots concentrate most MGEs and integrases, but not the majority of ISs. They also show that regions with high concentration of HTgenes often lack recognizable MGEs, suggesting that other mechanisms are implicated in their genesis and turnover.

**The chromosomal context of hotspots.** We then searched to identify the preferential genetic contexts of hotspots, since they might illuminate constraints associated with the chromosomal organization of HGT. We analyzed whether the distribution of hotspots was random relative to the function of the neighboring core genes. Interestingly, these core genes showed an overrepresentation of several functions, notably replication, recombination/repair, and transcription (Fig. 4a). In contrast, cell cycle control genes were underrepresented. Hence, hotspots are preferentially associated with specific functions of neighboring core genes.

We then tested whether hotspots were randomly distributed in genomes. Since replication drives much of the large-scale



**Fig. 6** Chromosomal context of hotspots. **a** Number of hotspots containing prophages, ICES/IMEs, and none of the above along the origin-terminus axis of replication. Linear regression and the confidence limits (*shaded area*) for the expected value (mean) were indicated for each category. The number of hotspots including prophages increases linearly with the distance to the origin of replication (Spearman's  $\rho = 0.87$ ,  $P < 10^{-3}$ ), but this is not the case for the other two categories (both  $P > 0.05$ ). **b** Heatmap of odds ratios of co-localizations in hotspots of MGEs, mobility-associated proteins (MAPs) and RNAs. \*\*\* $P < 10^{-3}$ ; \*\* $P < 10^{-2}$ ; \* $P < 0.05$ ; Fisher's exact test

organization of bacterial genomes<sup>8</sup>, we analyzed the position of hotspots relative to the distance to the origin of replication along the replicore. These results showed that the frequency of hotspots including prophages, as previously shown in *E. coli*<sup>13</sup>, increases linearly along the ori- $\rightarrow$  ter replication axis (Fig. 6a). Interestingly, this does not seem to be the case for ICES and IMEs, nor for the very large category of hotspots that lack ICES, IMEs, and prophages.

As these results show that prophages and ICES have different distribution patterns, we quantified the frequency of co-occurrence of different MGEs and MAPs in the same hotspots (but not necessarily in the same intervals, Fig. 6b). In line with expectations, most MGEs significantly co-occurred with integrases, integrons, ISs, and tRNAs. The most notable exceptions concerned the prophages, that did not significantly

co-occurred with ISs, presumably because ISs are rare in phages<sup>29</sup>, or integrons, and they were found less frequently than expected in spots with conjugative elements. This is in line with the analysis showing that they have specifically different distributions along the chromosome replication axis.

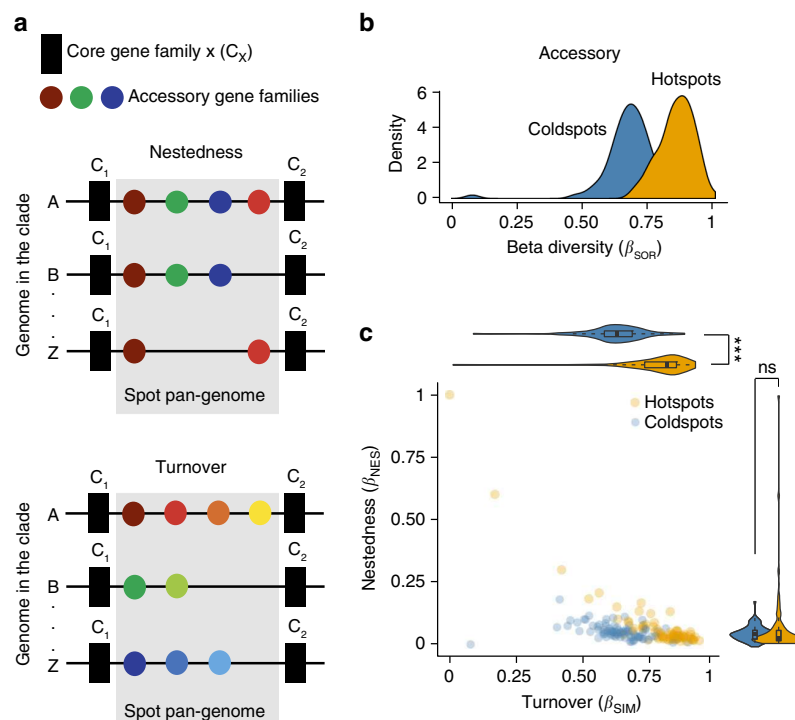
**Genetic diversity of hotspots.** The integration of a MGE in the chromosome adds a large number of genes in one single location, potentially creating a hotspot on itself. Such events result in a concentration of HTgenes in a genome (strain-specific integration), or in several genomes (when the integration took place at the last common ancestor of several strains). The distribution of the number of genomes with orthologous HTgenes in hotspots suggests that these cases are relatively rare (Supplementary Fig. 9a). Only 8% of the hotspots had all accessory gene families represented in one genome (Supplementary Fig. 9b, c). Hence, few of these regions seem to have been created by the integration of a single MGE.

To assess whether genetic diversity in hotspots was compatible with one single ancient integration event, we introduced measures derived from the analysis of beta diversity in Ecology, where it is used to measure the differences in species composition between different locations<sup>30</sup> (Methods). Here we used it to measure the difference in gene repertoires among a set of intervals from the same spot. We measured the Sørensen index ( $\beta_{SOR}$ ) for hotspots and coldspots of each species using the binary matrix of gene presence/absence. Diversity results from a mixture of independent gene acquisitions and replacements (turnover) and differential gene loss (nestedness), and  $\beta_{SOR}$  can be partitioned into the two related additive terms: turnover ( $\beta_{SIM}$ ) and nestedness ( $\beta_{NES}$ ) ( $\beta_{SOR} = \beta_{NES} + \beta_{SIM}$ , Fig. 7a).

Beta diversity of accessory genes was higher in hotspots than in coldspots (Fig. 7b). This difference was caused by turnover, since only  $\beta_{SIM}$  was significantly higher in hotspots than in coldspots (Fig. 7c). The values of  $\beta_{NES}$  were very low in both cases; confirming that most hotspots are not caused by singular events of integration of MGEs. We obtained similar results when the analysis of diversity was restricted to HTgenes (Supplementary Fig. 10). While genetic diversity is high in hotspots and coldspots, these results show faster diversification in hotspots because they endure higher genetic turnover.

Finally, we wished to test whether hotspots lacking MAPs had such a high genetic turnover that MGEs would be rapidly removed. We split the hotspots into two categories: hotspots containing and lacking MAPs. Both categories showed values of genetic diversity close to one that were caused by high turnover. Nevertheless, hotspots lacking MAPs showed slightly lower values for these variables (Supplementary Fig. 11). Hence, the absence of MAPs in these hotspots is not due to an excess of genetic turnover.

**Hotspots of homologous recombination.** Many hotspots lack identifiable MGEs or even integrases. Yet, they show high genetic diversity, suggesting that other mechanisms may drive their evolution. We tested the possibility that these regions could integrate HTgenes by homologous recombination at the flanking core genes, as suggested for certain hotspots of *E. coli*<sup>23</sup> and *S. pneumoniae*<sup>24</sup> (Fig. 8a). Our hypothesis predicts higher levels of homologous recombination in core genes flanking hotspots than in the rest of the core genome. We tested this prediction in two complementary ways. Firstly, we detected homologous recombination events in the core genes using ClonalFrameML (Methods). We found 50% more recombination events in core genes flanking hotspots than in the other core genes (Fig. 8b). Secondly, we searched for evidence of phylogenetic incongruence between each



**Fig. 7** Genetic diversity of the accessory genes present in hotspots and coldspots. **a** Examples of gene nestedness and turnover in a spot. Turnover measures the segregation between intervals in terms of gene families, i.e., it accounts for the replacement of some genes by others. Nestedness accounts for differential gene loss and measures how the gene repertoires of some intervals are a subset of the repertoires of the others. It typically reflects a non-random process of gene loss. **b** Distributions of  $\beta$  diversity ( $\beta_{SOR}$ ) in hotspots and coldspots. **c** Partition of  $\beta_{SOR}$  in its components of nestedness ( $\beta_{NES}$ ) and turnover ( $\beta_{SIM}$ ) for hotspots and coldspots ( $\beta_{SOR} = \beta_{NES} + \beta_{SIM}$ ). \*\*\*\* $P < 10^{-3}$ ; Mann-Whitney-Wilcoxon test; ns: not significant

core gene family and the whole core genome tree of the clade using the Shimodaira–Hasegawa (SH) test (Methods). The number of genes with significant phylogenetic incongruence was 30% higher among core genes flanking hotspots than among the others (Fig. 8b). In line with these observations, core genes flanking hotspots also had higher nucleotide diversity (Fig. 8c). We found qualitatively similar results when the analysis was performed on a per species or per genus basis (Supplementary Data set 5). Hence, core genes flanking hotspots are more targeted by recombination processes than the others.

Naturally transformable bacteria have the ability to acquire genetic material independently of MGEs. In these species, transfer of chromosomal material mediated by homologous recombination at the flanking core genes might be particularly frequent. To test this hypothesis, we put apart the 19 bacterial species that are known to be naturally transformable in our dataset<sup>31</sup> (Supplementary Data set 1). We observed that these species had more hotspots than the others ( $P < 0.05$ , Mann–Whitney–Wilcoxon test). We searched for MAPs in these hotspots and observed that they also had fewer hotspots with MAPs ( $P < 0.05$ , Mann–Whitney–Wilcoxon test). Finally, recombination was 20% more frequent in core genes flanking hotspots in naturally transformable than in the remaining bacteria ( $P < 10^{-4}$ ,  $\chi^2$ -test). These results suggest that recombination at core genes flanking hotspots might be particularly important in driving genetic diversification of naturally transformable bacteria.

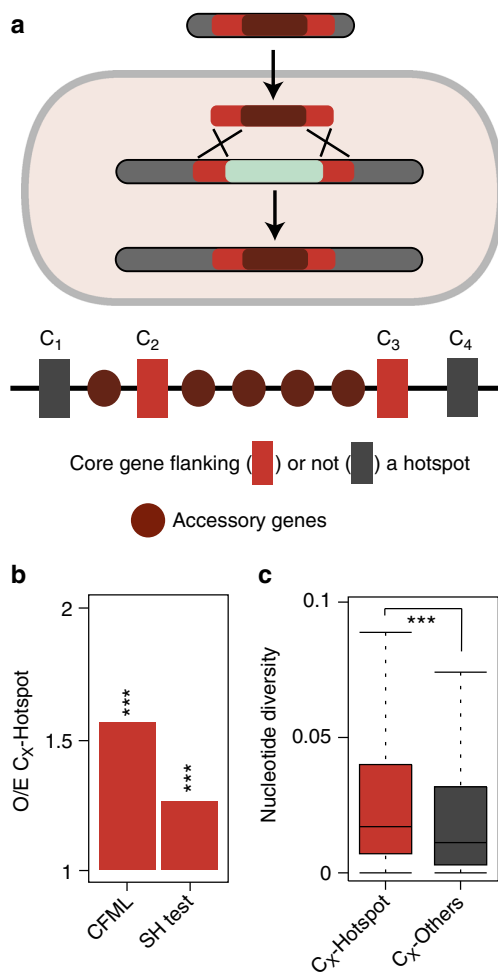
## Discussion

Our study showed high concentration of HTgenes in a small number of locations in the chromosomes of many bacterial species. These hotspots include most MGE-related genes, fitting

previous observations that the latter co-evolved with the host to use integrases targeting specific locations in the chromosome that minimize the fitness cost of chromosomal integration. For example, many temperate phages integrate tRNA genes without disrupting their function<sup>52</sup>. The concentration of most self-mobilizable MGEs at few loci might be thought sufficient to justify the existence of hotspots, but we found that few hotspots had identifiable prophages or conjugative elements and that most lacked integrases. These puzzling results could be caused by failure to identify MGEs, but our methods were shown to be highly accurate at identifying conjugative elements and prophages<sup>13, 33</sup>, or by the presence of many radically novel integrase-lacking MGEs in these model microbial species, which would be very surprising. Hotspots also contain degenerate MGEs that we have failed to identify. Yet, inactivated elements are not expected to drive the observed rapid genetic turnover of these regions.

Our results suggest that an MGE-independent mechanism, double homologous recombination at the flanking core genes, contributes to hotspot diversification. The mechanism only requires housekeeping recombination functions and exogenous DNA with homology to the flanking core genes. This last condition is easy to fulfill, because these genes are present in all genomes of the species (and usually in closely related species). In agreement with our hypothesis, we showed that naturally transformable species had more hotspots, and fewer MAPs in hotspots, than the others. There are other mechanisms of transfer that can bring homologous sequences without MAPs in non-transformable bacteria, including generalized transduction, gene transfer agents, or DNA-carrying vesicles<sup>34</sup>. Their role in hotspot diversification remains to be explored.





**Fig. 8** Evidence for more frequent homologous recombination in core genes flanking hotspots than in the other core genes. **a** Model for the creation and evolution of hotspots by homologous recombination at the flanking core genes. **b** We detected homologous recombination events in the core genes using ClonalFrameML, and searched for evidence of phylogenetic incongruence ( $P < 0.05$ ) between each core gene family and the whole core genome tree of the clade using the Shimodaira-Hasegawa (SH) test. The observed-expected ratios (O/E) for these two analyses are significantly higher than one.  $***P < 10^{-3}$ ;  $\chi^2$ -test. **c** Differences in nucleotide diversity between core genes flanking hotspots and the others. Nucleotide diversity was calculated using the R package "PopGenome" v2.1.6<sup>49</sup> by implementation of the diversity.stats() command.  $***P < 10^{-3}$ ; Mann-Whitney-Wilcoxon test

Many HTgenes are not adaptive (or even deleterious) and are rapidly lost by genetic drift (or purifying selection)<sup>6, 7, 35</sup>. Nevertheless, regions of high concentration of HTgenes must also include adaptive genes, as shown here for ARGs. In these circumstances, the high genetic turnover at hotspots might seem paradoxical, because it may lead to their loss. Actually, even adaptive genes can be lost with little fitness cost under certain circumstances. Genes under diversifying selection, such as defense systems, may be adaptive for short periods of time and subsequently lost (or replaced by analogous genes)<sup>36</sup>. Some costly genes may be adaptive in only very specific conditions, such as ARGs<sup>37</sup>, and become deleterious for the cell fitness upon environmental change. Finally, some genes under

frequency-dependent selection, such as toxins<sup>38</sup>, may stop being adaptive when their frequency changes in the population. Genetic drift, purifying, diversifying, and frequency-dependent selection can thus contribute to the rapid turnover of HTgenes. As a consequence of their high genetic turnover, hotspots are expected to be enriched in genes of specific adaptive value.

Hotspots may affect bacterial fitness not only by the genes they contain, but also by the way they drive genome diversification. According to the chromosome-curing model<sup>39</sup>, hotspots may facilitate the elimination of elements with deleterious fitness effects, such as certain MGEs, by double recombination at the flanking core genes. This fits our observation that core genes flanking hotspots endure higher rates of homologous recombination. As a response to chromosome curing, natural selection is expected to favor MGEs that inactivate genes encoding recombination and repair proteins<sup>39</sup>. Interestingly, we also found that hotspots tend to be flanked by recombination and repair core genes. Although these genes seem intact, at least they respect the constraints that we imposed for their classification as core genes, their expression may be affected by HTgenes in the neighboring hotspot. For example, excision of a MGE in *Vibrio splendidus* 12B01 from a *mutS* gene downregulates the expression of the latter leading to a hypermutator phenotype<sup>40</sup>.

Several selective effects can contribute to explain the very different number of hotspots per species, which were strongly correlated with the number of HTgenes and weakly with genome size (itself also correlated with the rate of HGT<sup>27</sup>). The first association may explain why species with little genetic diversity, such as *B. anthracis* and mycobacteria, have few hotspots in spite of their large genome size. It is also possible that our statistical tests lack power when species have few HTgenes. Some ecological determinants also affect the number of HTgenes, and their concentration in the genome. For example, sexually isolated species with few MGEs, such as obligatory endosymbionts, are expected to have few hotspots. Many of these species may also inefficiently select for hotspots because they have low effective population sizes. Conversely, the highest number of hotspots was found in facultative pathogens with very diverse gene repertoires, including *E. coli*, *Pseudomonas spp.*, and *Bacillus cereus*. A rigorous statistical assessment of the ecological traits affecting the organization of HTgenes will require the analysis of a larger panel of species representative of the different prokaryotic lifestyles.

Overall, our results suggest that hotspots are the result of the interplay of several recombination mechanisms and natural selection, presumably because they minimize disruption of genome organization by circumscribing gene flux to a small number of permissive chromosomal locations. For example, the increase in prophage-containing hotspots along the ori-> ter axis suggests co-evolution between these elements and the host to remove prophages from early replicating regions that are also rich in highly expressed genes in fast growing bacteria<sup>13</sup>. Interestingly, the spatial distribution of the remaining hotspots does not show similar patterns, which can be due to the lower fitness costs associated with their excision. Further work is needed to understand if there are other organizational traits that constrain the distribution of hotspots in the chromosome, and in particular in those devoid of recognizable MGEs. Knowing these traits might facilitate large-scale genetic engineering and should lead to a better understanding of the evolutionary interactions between horizontal gene transfer and genome organization.

Finally, our study focused on the dynamics of hotspots and how they contribute to genome diversification, but left unanswered the questions related to their origin and fate. Previous studies identified common prophage hotspots between *E. coli* and *Salmonella enterica*<sup>13</sup>. Hence, we will have to study

taxonomical units broader than the species level to unravel their origin. As for their fate, long-term adaptive HTgenes may become fixed in the population, explaining the patterns of nestedness of certain hotspots, and leading eventually to the split of the hotspot into two new (eventually hot) spots.

## Methods

**Data.** The sequences and annotations of 932 bacterial genomes from 80 bacterial species were retrieved from GenBank RefSeq (<ftp://ftp.ncbi.nih.gov/genomes>, last accessed in February 2014)<sup>41</sup>. We made no selection on the species that were to be analyzed, except that we required a minimum of four complete genomes per species. We have made no attempt to re-define species: we used the information presented in GenBank. Their list is available in Supplementary Data set 1. We excluded CDS annotated as partial genes, as well as those lacking a stop codon or having stop codons within the reading frame. Core genomes and phylogenetic reconstructions were obtained from our previous work<sup>27</sup> (Supplementary Data set 2). Our data set includes several species from the same genera. It also includes species with diverse numbers of genomes and HTgenes. To minimize the effects of these unavoidable biases most of our analyses are non-parametric and each species has the same weight. When they were done on the data cumulated from all species, we made a control where each species is analyzed separately. We also made complementary analyses where we aggregated the results per genus. The references for these supplementary controls are indicated in the main text, and the data are in the Supplementary Material.

**Identification of core genomes.** We used 80 core genomes previously published<sup>27</sup>. These core genomes were built for clades with at least four complete genomes available in GenBank RefSeq (Supplementary Data set 1, Supplementary Fig. 1). Briefly, a preliminary list of orthologs was identified as reciprocal best hits using end-gap-free global alignment, between the proteome of a reference genome (pivot, typically the first completely sequenced isolate) and each of the other strain's proteomes. Hits with <80% similarity in amino-acid sequence or >20% difference in protein length were discarded. This list of orthologs was then refined for every pairwise comparison using information on the conservation of gene neighborhood. Thus, positional orthologs were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighborhood of 10 genes (five upstream and five downstream). These parameters (four genes being less than half of the diameter of the neighborhood) allow retrieving orthologs at the edge of rearrangement breakpoints (positions where intervals were split by events of chromosome rearrangement) and therefore render the analysis robust to the presence of a few rearrangements. The core genome of each clade was defined as the intersection of pairwise lists of positional orthologs.

**Definitions of interval and spot.** The core genome is the collection of all gene families present in one and only one copy in each genome of a clade (Supplementary Fig. 2). Let  $C_X$  and  $C_Y$  be two families of core genes in a clade with  $N$  taxa where one of the taxa is a pivot (reference genome, see above). We call  $C_{AX}$  and  $C_{AY}$  contiguous core genes in a given chromosome  $A$  if they are adjacent in the list of core genes sorted in terms of the position in the chromosome. We defined an interval ( $I_{AX,AY}$ ) as the location between the pair of contiguous core genes  $C_{AX}$  and  $C_{AY}$  in chromosome  $A$ . The content of an interval is the set of accessory genes in the interval. The HTgenes content of an interval is the number of genes that were acquired by HGT in the interval. Multiple chromosomes, when present, were treated independently.

Intervals flanked by the same core gene families ( $C_X, C_Y$ ) as the pivot genome were defined as syntenic intervals (i.e., the members of the core gene families  $X$  and  $Y$  were also contiguous in the pivot). The intervals that do not satisfy this constraint were classed as breakpoint intervals and excluded from our analysis. They contain <2% of all genes. For every interval in the pivot genome, we defined spot as the set of syntenic intervals flanked by members of the same pair of core gene families (Supplementary Fig. 2).

**Identification of spot pan-genomes.** The pan-genome is the full complement of homologous gene families in a clade. We built a pan-genome for each species using the gene repertoire of each genome. Initially, we determined a preliminary list of putative homologous proteins between pairs of genomes (excluding plasmids) by searching for sequence similarity between each pair of proteins with BLASTP v.2.2.28+ (default parameters). We then used the  $e$ -values ( $<10^{-4}$ ) of the BLASTP output to cluster them using SILX (v1.2.8, <http://lbbbe.univ-lyon1.fr/SiLiX>)<sup>42</sup>. We set the parameters of SILX such that two proteins were clustered in the same family if the alignment had at least 80% identity and covered >80% of the smallest protein (options  $-I$  0.8 and  $-r$  0.8). We computed the diversity of gene families observed in each spot. The spot pan-genome is the set of gene families present in the intervals associated with the spot (Supplementary Fig. 2).

**Reconstruction of the evolution of gene repertoires.** We assessed the evolutionary dynamics of gene repertoires of each clade using Count<sup>43</sup> (downloaded in

April 2015). This program uses birth-death models to identify the rates of gene deletion, duplication, and loss in each branch of a phylogenetic tree. We used the spots' pan-genomes matrices, and the phylogenetic birth-and-death model of Count, to evaluate the most likely scenario for the evolution of a given gene family on the clade's tree. Rates were computed with default parameters, assuming a Poisson distribution for the family size at the tree root, and uniform gain, loss, and duplication rates. One hundred rounds of rate optimization were computed with a convergence threshold of  $10^{-3}$ . After optimization of the branch-specific parameters of the model, we performed ancestral reconstructions by computing the branch-specific posterior probabilities of evolutionary events, and inferred the gains in the terminal branches of the tree. The posterior probability matrix was converted into a binary matrix of presence/absence of HTgenes using a threshold probability of gain higher than 0.95 at the terminal branches and excluding gains occurring in the last common ancestor with a probability higher than 0.5.

**Identification of hotspots.** We made simulations to obtain the expected distribution of HTgenes in the spots given the numbers of HTgenes and spots (Supplementary Fig. 4). We made the null hypothesis that the distribution of these genes was constrained by the frequency of genes in operons, and followed a uniform distribution in all other respects. Previous works have shown that two-third of the genes are in operons and one-third are in mono-cistronic units<sup>44</sup>, with little inter-species variation for the average length of poly-cistronic units ( $3.15 \pm 0.06$ )<sup>45</sup>. Hence, given  $N$  HTgenes per clade we created two groups of elements:  $N/3$  isolated genes and  $2N/3$  in operons with three genes. These elements were then randomly placed among the spots following a uniform distribution. For each of the 1000 simulations (per species), we recorded the maximal value of genes within a single spot ( $\text{Max}_{\text{HTg},i}$ ), which was used to identify the value of the 95th percentile ( $T_{95\%}$ ) of the distribution of  $\text{Max}_{\text{HTg},i}$ . Hence, 95% of the simulations have no spot with more than  $T_{95\%}$  genes (Supplementary Data set 1). Spots (in the real genomes) with more than  $T_{95\%}$  HTgenes were regarded as hotspots. Spots lacking accessory genes were called empty spots. The other spots were called coldspots.

As a control, we also made simulations considering that HTgenes were acquired independently of the structure in operons (i.e., considering  $N$  isolated genes). The values of  $T_{95\%}$  of the two analyses were highly correlated (Spearman's  $\rho = 0.89$ ,  $P < 10^{-4}$ , Supplementary Data set 1), but those of the latter were smaller (linear regression:  $T_{95\%}^{\text{isolated}} = -0.62 + 0.66 T_{95\%}^{\text{operons}}$ ,  $R^2 = 0.87$ ). This is expected because the operon structure should increase the variance of the genes per spot, and thus increase  $T_{95\%}$ .

**Measures of gene repertoire diversity.** Since most spots have few or no genes, and most gene families have few (or no gene) per genome, we computed the genetic diversity of spots using matrices of presence/absence of gene families (computed from the pan-genome).

We computed beta diversity per clade, using a multiple-site version (each interval is the equivalent of a site)<sup>46</sup> of the widely used Sorensen dissimilarity index ( $\beta_{\text{SOR}}$ ):

$$\beta_{\text{SOR}} = \frac{\sum_{i < j} \min(b_{ij}, b_{ji}) + \sum_{i < j} \max(b_{ij}, b_{ji})}{2 \cdot (\sum_i S_i - S_T) + \sum_{i < j} \min(b_{ij}, b_{ji}) + \sum_{i < j} \max(b_{ij}, b_{ji})}, \quad (1)$$

where  $S_i$  is the total number of accessory genes in genome  $i$ ,  $S_T$  is the total number of accessory genes in all genomes considered together, and  $b_{ij}, b_{ji}$  are the numbers of accessory genes present in genome  $i$  but not in  $j$  ( $b_{ij}$ ) and vice-versa ( $b_{ji}$ ).

We then used a partitioned version of the ecological concept of beta diversity to characterize the gene diversity of spots<sup>46</sup>.  $\beta_{\text{SOR}}$  can be partitioned into two additive terms: turnover ( $\beta_{\text{SIM}}$ ) and nestedness ( $\beta_{\text{NES}}$ ) ( $\beta_{\text{SOR}} = \beta_{\text{NES}} + \beta_{\text{SIM}}$ , Fig. 7a).

To compute the turnover we used the multiple-site version<sup>46</sup> of the Simpson dissimilarity index ( $\beta_{\text{SIM}}$ ):

$$\beta_{\text{SIM}} = \frac{\sum_{i < j} \min(b_{ij}, b_{ji})}{2 \cdot (\sum_i S_i - S_T) + \sum_{i < j} \min(b_{ij}, b_{ji})}, \quad (2)$$

This index is a measure of the evenness with which families of genes are distributed across intervals of a spot (it is a measure of segregation). Turnover implies the replacement of some gene families by others.

By definition, the multiple-site dissimilarity term accounting only for nestedness ( $\beta_{\text{NES}}$ ) results from the subtraction<sup>46</sup>:

$$\beta_{\text{NES}} = \beta_{\text{SOR}} - \beta_{\text{SIM}}. \quad (3)$$

Nestedness occurs when intervals with fewer genes are subsets of intervals with larger gene repertoires. It reflects a non-random process of gene loss.

The above formulae were computed as follows: first, we plotted the distribution of the number of accessory genes from the hotspots of all clades analyzed. We took the minimum of this distribution ( $\text{min}_d$ ) and used it to select coldspots with a number of accessory genes equal or higher than  $\text{min}_d$ . By doing this, we eliminated coldspots with very few accessory genes, and likely to introduce a bias while computing diversity (leading to extreme situations where  $\sum S_i \approx S_j$ ;  $\beta_{\text{SIM}} \approx 1$ , and as consequence  $\beta_{\text{NES}} \approx 0$ ). After this filtering step, we put together all hotspots and all coldspots of each genome in two separate concatenates to avoid statistical artifacts

associated with poorly populated spots. The diversity was computed per clade for each of the concatenates.

**Inference of homologous recombination.** We inferred homologous recombination on the multiple alignments of the core genes of each clade using ClonalFrameML (CFML) v10.7.5<sup>47</sup> with a predefined tree (i.e., the clade's tree), default priors  $R/\theta = 10^{-1}$ ,  $1/\delta = 10^{-3}$ , and  $\nu = 10^{-1}$ , and 100 pseudo-bootstrap replicates, as previously suggested<sup>47</sup>. Mean patristic branch lengths were computed with the R package "ape" v3.3<sup>48</sup>, and transition/transversion ratios were computed with the R package "PopGenome" v2.1.6<sup>49</sup>. The priors estimated by this mode were used as initialization values to rerun CFML under the "per-branch model" mode with a branch dispersion parameter of 0.1.

**Functional assignment.** Gene functional assignment was performed by searching for protein similarity with HMMer (hmmsearch) on the bactNOG subset of the eggNOG v4.5 database<sup>50</sup> (downloaded in March 2016). We have considered the pivot (reference) genomes as good representatives of each clade, and limited our analysis to these. We have kept hits with an *e*-value lower than  $10^{-5}$ , a minimum alignment coverage of 50%, and when the majority (>50%) of non-supervised orthologous groups (NOGs) attributed to a given gene pertained to the same functional group. Hits corresponding to poorly characterized or unknown functional groups were discarded.

**Identification of MGEs and proteins associated to mobility.** Temperate phages integrated in the bacterial chromosome (prophages) were identified using Phage Finder v4.6<sup>51</sup> (stringent option). Prophages with > 25% of the predicted genes belonging to ISs, and partially degraded prophages (shorter than 30 kb) were removed<sup>52</sup>. Integrons were identified using IntegronFinder v1.4 with the -local\_max option<sup>53</sup>. Integrative conjugative elements (ICEs) and integrative mobilizable elements (IMEs) were identified using MacSyFinder v1.0.2<sup>54</sup> with TXSScan profiles<sup>55</sup>. Elements with a full conjugative apparatus were classed as ICE, the others as IME (see ref. <sup>33</sup> for criteria). Integrases were identified using the PFAM profiles PF00589 for tyrosine recombinases, and the pair of profiles PF00239 and PF07508 for serine recombinases (<http://pfam.xfam.org/>)<sup>56</sup>. All the protein profiles were searched using hmmsearch from the HMMer suite v.3.1b1 (default parameters). Hits were regarded as significant when their *e*-value was smaller than  $10^{-3}$  and their alignment covered at least 50% of the protein profile. Insertion sequences (ISs) were detected combining two approaches (i) using hmmsearch from the HMMer with IS HMM profiles (as previously proposed)<sup>57</sup> and (ii) by a BLAST-based method using the ISFinder database<sup>58</sup>. Integrases and transposases were defined as mobility-associated proteins (MAPs). tRNA genes were identified using tRNAscan SE v.1.21<sup>59</sup>, tmRNA genes were identified using Aragorn v.1.2.37<sup>60</sup>, and the location of the rRNA genes was taken from the Genbank annotation file. Antibiotic resistance genes were detected using HMMer against the curated database of antibiotic resistance protein families ResFams (Core v.1.2, <http://www.dantaslab.org/resfams>)<sup>61</sup> using the '-cut\_ga' option. A hotspot was considered to encode a peculiar MGE or MAP when at least one genome of the clade contained such element (Supplementary Data set 3).

**Identification of origin and terminus of replication.** The *ori* and *ter* of replication were predicted using Ori-Finder in the pivot genome of each clade<sup>62</sup>. When the ratio of the predicted replicore length was greater than 1.2, the clade was removed from the analysis (Supplementary Data set 4). Then, we divided each replicore in 10 equally sized regions from the *ori* to the *ter* of replication.

**Phylogenetic analyses.** We retrieved the 16S rRNA sequences of the sequenced type strains (also used as reference genomes, see above) of the 80 bacterial clades (Fig. 3a). We made a multiple alignment of them with MAFFT v7.305b<sup>63</sup> using default settings, and removed poorly aligned regions with BMGE v1.12<sup>64</sup> using default settings. The tree was computed by maximum likelihood with PHYML v3.0<sup>65</sup> under the general time reversible (GTR) +  $\Gamma(4)$  + I model (Supplementary Data set 2a). This tree is never used in the calculations; it is only used in Fig. 3a to display the relative position of each clade in the phylogeny of bacteria.

We built core genome trees for each clade using a concatenate of the multiple alignments of the core genes (see main text). Each clade's tree was computed with RAXML v8.00<sup>66</sup> under the GTR model and a gamma correction (GAMMA) for variable evolutionary rates. All trees are shown in Supplementary Data set 2b. We performed 100 bootstrap experiments on the concatenated alignments to assess the robustness of the topology of each clade's tree. The vast majority of nodes were supported with bootstrap values higher than 90% (Supplementary Data set 2b). We inferred the root of each phylogenetic clade's tree using the midpoint-rooting approach of the R package "phangorn" v1.99.14<sup>67</sup>. The alignment and the tree for each individual core gene were used for topology testing against the clade's tree (i.e., the concatenate tree of all the core genes of the clade) using the Shimodaira-Hasegawa (SH) congruence test<sup>68</sup> (1000 replicates) implemented in IQ-Tree v1.4.3<sup>69</sup>.

**Data availability.** The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files and from the corresponding author upon reasonable request.

Received: 6 March 2017 Accepted: 31 July 2017

Published online: 10 October 2017

## References

- Wilmes, P., Simmons, S. L., Deneff, V. J. & Banfield, J. F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.* **33**, 109–132 (2009).
- Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
- Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
- van Passel, M. W., Marri, P. R. & Ochman, H. The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLoS Comput. Biol.* **4**, e1000059 (2008).
- Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
- Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
- Koskineniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS Genet.* **8**, e1002787 (2012).
- Rocha, E. P. The organization of the bacterial genome. *Annu. Rev. Genet.* **42**, 211–233 (2008).
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
- Vieira-Silva, S. & Rocha, E. P. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808 (2010).
- Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W. H. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246**, 808–810 (1989).
- Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Bobay, L. M., Rocha, E. P. & Touchon, M. The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* **30**, 737–751 (2013).
- Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
- Burrus, V., Pavlovic, G., Decaris, B. & Guedon, G. Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* **46**, 601–610 (2002).
- Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276 (2003).
- Asadulghani, M. et al. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* **5**, e1000408 (2009).
- Sullivan, J. T. & Ronson, C. W. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a *phe*-tRNA gene. *Proc. Natl Acad. Sci. USA* **95**, 5145–5149 (1998).
- Murphy, K. C. Phage recombinases and their applications. *Adv. Virus Res.* **83**, 367–414 (2012).
- Balbontin, R., Figueroa-Bossi, N., Casades, J. & Bossi, L. Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica*. *J. Bacteriol.* **190**, 4075–4078 (2008).
- Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* **17**, 47–53 (2009).
- Williams, K. P. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–875 (2002).
- Touchon, M. et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
- Croucher, N. J. et al. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
- Chancey, S. T. et al. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. *Front. Microbiol.* **6**, 26 (2015).
- Everitt, R. G. et al. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nat. Commun.* **5**, 3956 (2014).
- Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663 (2016).
- Homma, K., Fukuchi, S., Nakamura, Y., Gojobori, T. & Nishikawa, K. Gene cluster analysis method identifies horizontally transferred genes with high

- reliability and indicates that they provide the main mechanism of operon gain in 8 species of gamma-Proteobacteria. *Mol. Biol. Evol.* **24**, 805–813 (2007).
29. Leclercq, S. & Cordaux, R. Do phages efficiently shuttle transposable elements among prokaryotes? *Evolution* **65**, 3327–3331 (2011).
  30. Koleff, P. Measuring beta diversity for presence–absence data. *J. Anim. Ecol.* **72**, 367–382 (2003).
  31. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J. P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).
  32. Campbell, A. Prophage insertion sites. *Res. Microbiol.* **154**, 277–282 (2003).
  33. Guglielmini, J., Quintais, L., Garcillan-Barcia, M. P., de la Cruz, F. & Rocha, E. P. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* **7**, e1002222 (2011).
  34. Garcia-Aljaro, C., Balleste, E. & Muniesa, M. Beyond the canonical strategies of horizontal gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **38**, 95–105 (2017).
  35. Kuo, C. H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
  36. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
  37. Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–271 (2010).
  38. Levin, B. R. Frequency-dependent selection in bacterial populations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **319**, 459–472 (1988).
  39. Croucher, N. J. et al. Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol.* **14**, e1002394 (2016).
  40. Chu, N. D. et al. A Mobile element in *mutS* drives hypermutation in a marine *Vibrio*. *MBio.* **8**, e02045–16 (2017).
  41. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
  42. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
  43. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
  44. Touchon, M. & Rocha, E. P. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb. Perspect. Biol.* **8**, a018168 (2016).
  45. Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J. & Kasif, S. Computational identification of operons in microbial genomes. *Genome Res.* **12**, 1221–1230 (2002).
  46. Baselga, A. Partitioning the turnover and nestedness components of beta diversity. *Global Ecol. Biogeogr.* **19**, 134–143 (2010).
  47. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
  48. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
  49. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
  50. Powell, S. et al. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**, D231–D239 (2014).
  51. Fouts, D. E. Phage\_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
  52. Bobay, L. M., Touchon, M. & Rocha, E. P. Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS Genet.* **9**, e1003825 (2013).
  53. Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
  54. Abby, S. S., Neron, B., Menager, H., Touchon, M. & Rocha, E. P. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9**, e110726 (2014).
  55. Abby, S. S. et al. Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080 (2016).
  56. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
  57. Kamoun, C., Payen, T., Hua-Van, A. & Filee, J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* **14**, 700 (2013).
  58. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
  59. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
  60. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
  61. 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).
  62. Gao, F. & Zhang, C. T. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics* **9**, 79 (2008).
  63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
  64. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC. Evol. Biol.* **10**, 210 (2010).
  65. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
  66. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
  67. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
  68. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
  69. 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).
  70. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).

## Acknowledgements

This work was supported by an European Research Council grant (EVOMOBILOME no. 281605). J.C. is a member of the 'Ecole Doctorale Frontière du Vivant (FdV)—Programme Bettencourt'. We thank the members of the Microbial Evolutionary Genomics group for comments and suggestions on the manuscript.

## Author contributions

P.H.O., M.T. and E.P.C.R. designed the research; P.H.O., M.T. and E.P.C.R. analyzed the data; J.C. provided data and tools; P.H.O., M.T. and E.P.C.R. wrote the paper.

## Additional information

**Supplementary Information** accompanies this paper at doi:10.1038/s41467-017-00808-w.

**Competing interests:** The authors declare no competing financial interests.

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

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

