Bioinformatic analysis of the genomes of epidemic pseudomonas aeruginosa

Panisa Treepong

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UFC
UNIVERSITÉ
DE FRANCHE-COMTÉ

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE FRANCHE-COMTÉ

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

PANISA TREEPONG

Bioinformatic analysis of the genomes of epidemic *Pseudomonas aeruginosa*

Analyse bioinformatique des génomes d'une souche épidémique de *Pseudomonas aeruginosa*

Thèse présentée et soutenue à Besançon, le 10 October 2017

Composition du Jury :

HOCQUET DIDIER	PHPU à l'Université de Franche-Comté	Président
PERRIÈRE GUY	DR à l'Université Claude Bernard, Lyon 1	Rapporteur
PUDLO PIERRE	PR à l'Université Aix-Marseille	Rapporteur
LE ROUZIC ARNAUD	CR à l'Université Paris-Sud	Examinateur
GUYEUX CHRISTOPHE	PR à l'Université de Franche-Comté	Directeur de thèse
VALOT BENOÎT	IR à l'Université de Franche-Comté	Encadrant de thèse

N° 2 1 4 1 0 8 2 5

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude and deep appreciation to my supervisor, Prof. Christophe Guyeux and to my co-supervisors Prof. Didier Hocquet HOCQUET and Dr. Benoît Valot, for their kindness, unfailing support, encouragement and invaluable guidance. They inspire me and keep an eye on every detail of my work. Without them, I would not be able to finish this thesis.

I would like to sincerely thank all thesis members, Prof. Pierre Pudlo, Dr. Guy Perrière and Dr. Arnaud Le Rouzic, for giving me the honor of accepting to be reviewers and examiner of this thesis. Moreover, they also give the valuable suggestion for improving my thesis manuscript.

I am grateful to Dr. Veronica N. Kos and Dr. Dominique S. Blanc for their precious recommendation and discussion to enhance the quality of this thesis.

I would like to thank Prince of Songkla University, Thailand, for the great opportunity and full scholarship to carry out my Doctorate degree.

My sincere thanks also goes to the Mésocentre de calcul de Franche-Comté team, who help and support in utilizing of supercomputer for launching the calculations this research.

I would like to thank DISC department, FEMTO-ST institute and Chrono-environnement laboratory for facilities supporting this work. Including the members of DISC and Chrono-environnement for their friendship, generousness and support. Especially, Bashar Al-Nauimi, Huda AL-NAYYEF, Bassam Alkindy, and Marie Petitjean.

I would also like express my thanks to all my close friends for being beside me, when I face the problems and give me the strength for overcoming those problems.

Last but not the least, I would like to thank my father, my mother, my sister and my love, for their support, encouragement and endless love throughout my life.

Contents

I

INTRODUCTION

1

General presentation

1.1/ Introduction

Pseudomonas aeruginosa is a Gram-negative bacterium that causes significant mortality and morbidity among compromised patients, like those suffering from cystic fibrosis. Its treatment is complicated by the noticeable ability of *P. aeruginosa* to develop resistance to almost all antibiotics. This is achieved through the selection of gene mutations, such as: (i) resistance to cephalosporins after mutations in the many regulators of AmpC cephalosporinase, (ii) to carbapenems after inactivation of the porin OprD, or (iii) to fluoroquinolones, after mutations in quinolone-resistance determining regions (QRDR) [Fournier et al., 2013, Schmidtke and Hanson, 2008, Poole, 2011], and the spread of horizontally acquired resistance [Livermore, 2002]. Environmental and clinical isolates owe their extraordinary ability to thrive in many ecological niches and to harm many hosts to the conservation of metabolic and virulence genes in the genome of the species [Valot et al., 2015, Grosso-Becerra et al., 2014, Wolfgang et al., 2003].

P. aeruginosa has a non-clonal structure, but some sequence types (STs) reported as 'high-risk clones' are frequent and widely distributed, such as ST111, ST175, ST235, ST244, and ST395. Up to now, ST235 is the most widespread clones associated with multi- and high-level antibiotic resistance [Curran et al., 2004, Maatallah et al., 2011, Pirnay et al., 2002, Woodford et al., 2011, Kos et al., 2015, Oliver et al., 2015].

Recently reported results found that the acquisition of resistance genes are related to approximately 100 different horizontally-acquired resistance elements in ST235 isolates [Livermore, 2002, Zeng and Jin, 2003, Potron et al., 2015, Oliver et al., 2015]. The expression of these acquired genes, together with intrinsic resistance mechanisms, considerably reduces the therapeutic options for the treatment of infections due to this ST235 *P. aeruginosa*. *P. aeruginosa* also has mechanisms for acute infection through its virulence factors, *e.g.*, this pathogen utilizes the type III secretion system to inject genetic elements into host cell and thus cope with host immune system.

1.2/ PRESENTATION OF THE PROBLEMS

While *P. aeruginosa* ST235 obviously has a wide influence on clinical problems, the molecular basis for the success of this clone was, up to now, hard to decipher. Even though *P. aeruginosa* ST235 isolates are highlighted in the literature due to their association with infection and extensive drug resistance, it remains unclear whether the dissemination of the ST235 lineage relies exclusively on resistance to antibiotics or if other features such as virulence factors are major contributors.

In addition, although the spread of ST235 has been documented in many locations, little is known about the relationship between isolates from different countries, and about the evolution and emergence of this clone. Indeed, available studies that have analyzed the genome content of a collection of ST235 isolates did not resolve the geographical or temporal origin of their ancestor [Kos et al., 2015, van Belkum et al., 2015].

1.3/ OBJECTIVE OF THE THESIS

Potentially all the *Pseudomonas aeruginosa* clones can be pathogen for the human being. However, epidemic outbreaks are most of the time due to the aforementioned high risk clones. And, among the diverse sequence types (STs) that have been reported according to their population structure, the so-called ST235 is the most frequent one. Such clones are found on every continents, and they frequently are multi-resistant to antibiotics thanks to the acquisition of resistance genes and to chromosomal mutations.

The objective of this thesis is to know the reasons why these clones have successfully spread around the world, through the example of ST235. Either these clones are ubiquitous and very frequent (more than the non pandemic clones) in the environment, which acquire in a second phase some resistance genes (hypothesis number one). Or the diffusion is more recent, promoted by the diffusion of subclones that have previously acquired some resistance genes (hypothesis number 2). This second hypothesis has been observed in the case of the pandemic clone of another pathogen *Escherichia coli* ST131. The first objective of this thesis is thus to determine, by the mean of bioinformatics studies on big genomic data, the phylogenetic structure of the ST235 clone, to understand the rhythm of its biological clock, to determine the antibiotics that could have favored its emergence and spread, and to date the worldwide diffusion of this high-risk clone.

To do so, we will have to collect a large number of isolates of this ST235 clone, coming from the five inhabited continents. We will then have to analyze the single nucleotide polymorphisms (SNPs) of the core genome of these isolates, for phylogenetic studies. These studies will necessitate the design of bioinformatics pipelines, encompassing existing tools based on Markov chains or maximum likelihood, and new algorithms to design. This pipeline will be applied on the set of

original data and on the obtained phylogeny, in order to provide evidences about the emergence (date and location) of the ST235 clone. We will have to pay special attention on the insertion sequences, which should necessitate to design *de novo* algorithms of detection.

1.4/ Organization of the thesis manuscript

This thesis is organized as follows. In the next chapter that follows this introductory part, various basic recalls and tools are presented. The first section deals with the bioinformatics analysis of bacterial genome evolution, through three main directions: genome sequencing, phylogenetic analysis, and comparative genomics. It is followed by various state of the art knowledges about *Pseudomonas aeruginosa*, encompassing its ecological niche, its pathogenicity, population structure, resistance, and genome evolution and adaptation.

The second part of this manuscript is devoted to our personal contributions. It is divided in two chapters, corresponding to the two main works we performed during this thesis. In the first chapter, we presented a complete pipeline applied on a large and original set of *P. aeruginosa* ST235. The objective of this first chapter is to understand the reasons of the emergence and the worldwide spread of this clone. In particular, materials and methods are introduced, a phylogenetic network is computed, and a most recent common ancestor analysis is performed. The spatiotemporal origin of the ST235 clone is discovered, and its emergence is associated with the extensive use of a particular class of antibiotics.

This first chapter of contribution is followed by the presentation of panISa, our new tool to find insertion sequences (ISs) from NGS data. We will present first all the elements of implementation of our tool, and then how we have constructed a dataset for validation. This validation has been achieved by artificially inserting ISs in reference genomes, and by checking whether panISa is able to recover well these insertion sequences. This tool has finally been applied to an epidemic strain of *Pseudomonas aeruginosa*.

This manuscript ends by a conclusion part, in which our contributions are summarized and intended future work are outlined. Finally, a complete bibliography brings this document to a close.

1.5/ Publications

1.5.1/ Peer-reviewed international Journals

1. P. Treepong, V. Kos, C. Guyeux, D. Blanc, X. Bertrand, B. Valot, and D. Hocquet. Global emergence of the high-risk *Pseudomonas aeruginosa* ST235

clone. Clinical Microbiology and Infection (I.F. 5.292 (2016), ranked Q1 (8/84) in Infectious diseases and Q1 (18/124) in Microbiology).

2. P. Treepong, C. Guyeux, B. Valot, and D. Hocquet. PanISa: a new tool to find insertion sequences from NGS data. Under submission.

1.5.2/ POSTER

1. Bassam Alkindy, Bashar Al-Nuaimi, Huda Al'Nayyef, Panisa Treepong, Christophe Guyeux, Jean-François Couchot, Michel Salomon, and Jacques Bahi. "Bioinformatics Approaches on Genomic Evolution in Femto-ST (Core Genome, Phylogenetic Analysis, Transposable Elements, and Ancestral Reconstruction)". Workshop of Femto-ST, June 2015, Besançon, France. Note: Poster.

1.5.3/ SEMINARS

1. Structure de population mondiale du clone épidémique *Pseudomonas aeruginosa* ST235. Panisa Treepong, Veronica N. Kos, Christophe Guyeux, Dominique Blanc, Xavier Bertrand, Benoît Valot, Didier Hocquet. 36e Réunion Interdisciplinaire de Chimiothérapie Anti-infectieuse. Décembre 2016 - PARIS, France.

2. P. Treepong. PanISa: a tool for insertion sequence detection in bacterial genomes. AND team seminar, 28th of June, 2017. Belfort.

3. The emergence history of the high-risk *Pseudomonas aeruginosa* ST235 clone. Chronoenvironnement seminar, 5th of September, 2017. Besançon.

2

STATE OF THE ART

2.1/ BIOINFORMATICS ANALYSIS OF BACTERIAL GENOME EVOLUTION

The next-generation sequencing technologies have continuously delivered high volume and variety of bacterial genomes since its emergence. Such huge data can be analyzed for the benefit of clinical studies such as outbreak investigation, or to understand the evolution and spread of drug resistance. But such investigations necessitate the development and use of bioinformatics tools on NGS data (see Figure 2.1), the latter being detailed in this chapter.

2.1.1/ SEQUENCING GENOMES WITH NEXT-GENERATION DNA SEQUENCING (NGS) TOOLS

The first complete genome of a bacteria, namely of *Haemophilus influenzae*, was sequenced by the so-called Sanger shotgun technique in 1995. This sequencing technique is based on the DNA chain terminator dideoxynucleotide (ddNTP). Although this technique is able to generate a complete genome with 99.99% of accuracy, but is expensive and time consuming. These limitations were fixed a decade later, where next-generation or high-throughput DNA sequencing became able to deal with a huge amount of data (of the order of a billion) in parallel. NGS technology is faster and cheaper than the Sanger chemistry [Loman et al., 2012, Loman and Pallen, 2015]. In addition, several NGS platforms are available for microbiological research; they are different in features, strengths, and weaknesses summarized in Table 2.1.

2.1.1.1/ NEXT-GENERATION SEQUENCE DATA

The sequence data, produced by NGS platforms for downstream analysis, are provided in various formats, depending on the sequencing platform:

Table 2.1: **The different features of next-generation sequencing platforms** [Loman et al., 2012].

Machine (manufacturer)	Chemistry	Read length (bases)	Run time	Gb per run	Cost ($)	Advantages	Disadvantages
454 GS FLX+ (Roche)	Pyrosequencing	700-800	23 hours	0.7	500,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
HiSeq 2000/2500 (Illumina)	Reversible terminator	2×100	2-11 days	120-600	750,000	• Cost-effectiveness • Steadily improving read lengths • Massive throughput • Minimal hands-on time	• Long run time • Short read lengths
5500xl SOLiD (Life Technologies)	Ligation	75+35	8 days	150	350,000	• Low error rate • Massive throughput	• Very short read lengths • Long run times
PacBio RS (Pacific Biosciences)	Real-time sequencing	3,000-15,000	20 minutes	3 per day	750,000	• Simple sample preparation • Low reagent costs • Very long read lengths	• High error rate • Expensive system • Difficult installation
GS Junior (Roche)	Pyrosequencing	500	8 hours	0.035	100,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
Ion Personal Genome Machine (Life Technologies)	Proton detection	100 or 200	3 hours	0.01-1	80,000	• Short run times • Appropriate throughput for microbial applications	• Appreciable hands-on time • High error rate in homopolymers
Ion Proton (Life Technologies)	Proton detection	Up to 200	2 hours	10-100	220,000	• Short run times • Flexible chip reagents	• Instrument not available at time of writing
MiSeq (Illumina)	Reversible terminator	2×150	27 hours	1.5	125,000	• Cost-effectiveness • Short run times • Appropriate throughput for microbial applications • Minimal hands-on time	• Read lengths too short for efficient assembly

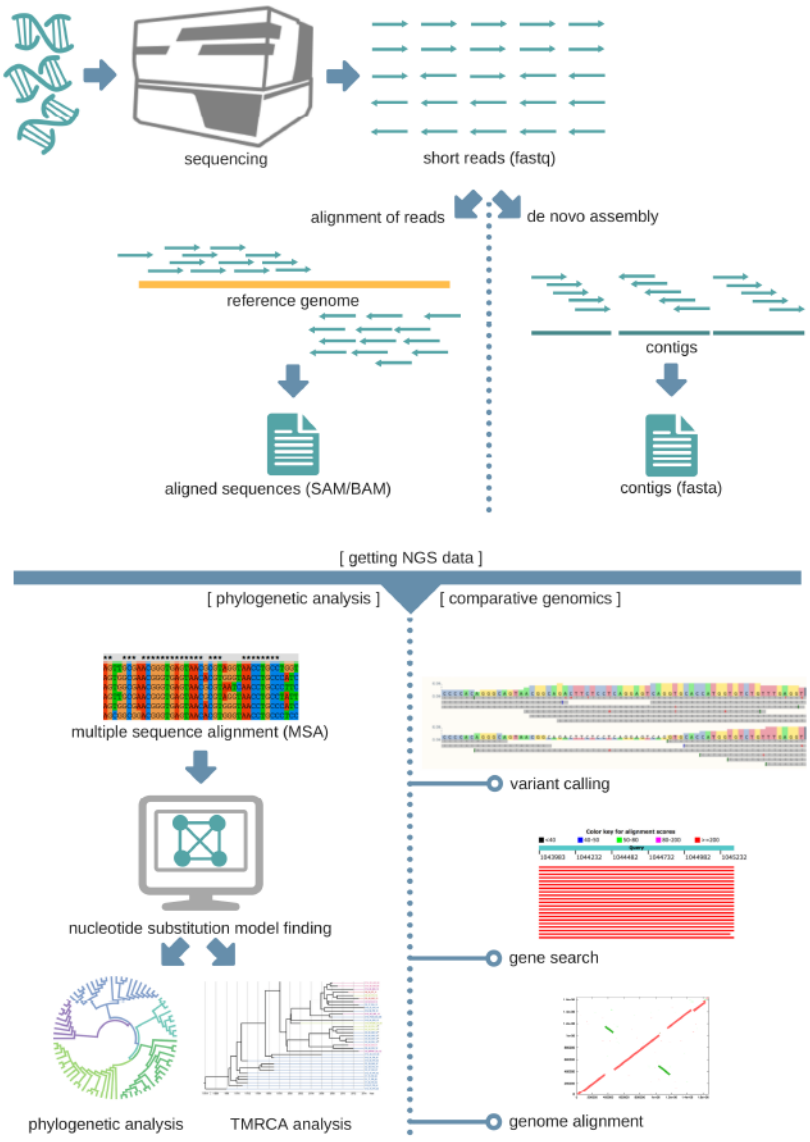


Figure 2.1: **An overview of bioinformatic analysis of bacterial genome evolution.**

- **SFF** (Standard Flowgram Format): a binary file format encoding pyrosequencing results from Roche 454 and IonTorrent which collects information of flowgram, called sequence, qualities, and the recommended quality and adaptor clipping;

- **HDF5**: a data model, library, and file format for storing and managing data of PacBio;

- **BCL** (basecall): a primary sequencing output for storing base call and its quality in per-cycle of illumina.

2.1.1.2/ ALIGNMENT OF READS

Even if short reads contained in FASTQ files are adequate in volume and of good quality, they are not ready to be analyzed if they are not addressed in the correct position. This is why alignment of reads is usually performed by mapping them against a reference genome.

This section introduces the algorithms and tools associated to read alignment. These alignment algorithms can be categorized into three groups: (1) hashing based methods, (2) Burrows-Wheeler Transform (BWT) based algorithms, and (3) merge sorting method. As the last one is only used in Slider [Li and Homer, 2010], we will thus only focus on the most commonly used algorithms, namely the hashing-based and BWT algorithms.

Hashing based algorithms use a hash table to transform subsequences in keys that are divided in two types: read hashes and reference genome ones. Handling with large sequences, a "seed and extend" method is applied to store the cut-sequence as k-mers for hash table mapping. For instance, in Figure 2.2 [Schbath et al., 2012], we start by cutting the genome "GCACAGCACA" into overlapping 3-mers. Then, we define their positions (Part A), and we also cut read "ACAGCA" into 3-mers (Part B). Then 3-mers of reads are mapped to the overlapping 3-mers of genome based on the hash table, after what the seed positions are sorted (Part C). Finally, the seed positions are verified and then saved (Part D).

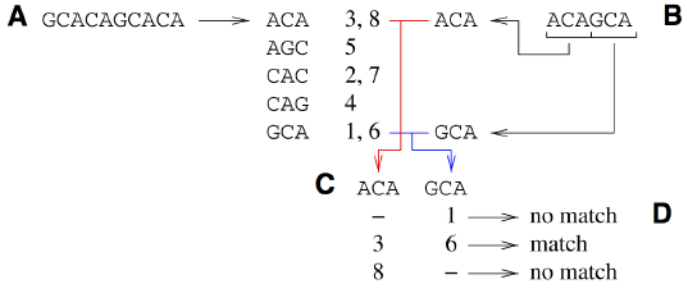


Figure 2.2: **Hash table based on seed and extend method.**

In the case of small mismatches, we can apply either a "pigeon hole" or a "q-gram filtering" method to perform seed and extend. Although the pigeon hole (Figure 2.3A) and q-gram filtering are similar in cutting the reads into smaller size and allowing small errors in mapping, the q-gram filtering cuts reads by overlapping, *e.g.*, read ACGT will be cut in AC, CG, and GT. If many mismatches are permitted, we will use the so-called spaced seeds method. Figure 2.3B shows how to map the GTCA read on the GATTACA genome using this spaced seeds method. In this example, the spaced seeds are "xT" and "xA", and the latter contain "x" or black box positions to allow errors/mismatches [Schbath et al., 2012].

In all cases described above, the "extend" phase is time consuming, while the Smith-Waterman algorithm is able to increase the alignment speed [Flicek and

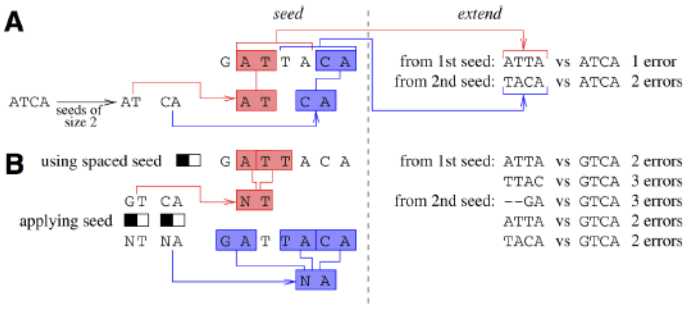


Figure 2.3: **Methods of hash table handling mismatches/indels.**

Birney, 2009, Li and Homer, 2010] (see a next section for details about this algorithm).

Burrows-Wheeler transform based algorithms are currently the most popular alignment methods, since BWT is able to rapidly search exact matches. The data structure based on suffix tree and suffix array with BWT are the keys for shrinking data.

A suffix tree is a tree storing suffixes of sequences from root to leaves in path format. Even though the suffix tree is proper for storing reads and reference genomes, it is not a good choice for large genomes like the human one. Figure 2.4 displays sample suffixes of "GATTACA", in which the dotted arrows represent the continued paths and double circles represent a suffix end. [Schbath et al., 2012].

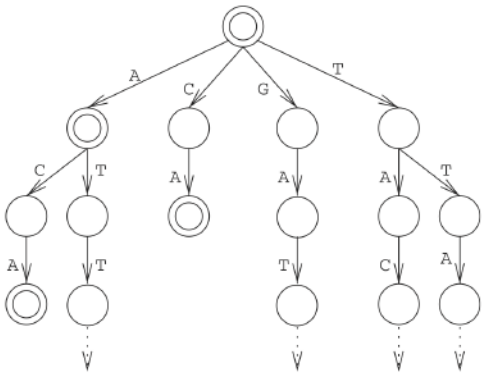


Figure 2.4: **The suffix tree of GATTACA.**

A suffix array is the set of all suffixes of a given genome stored in an array, leading to a better ability to deal with large genomes when compared with the aforementioned suffix tree. Figure 2.5 shows an algorithm of storage data in suffix array which consists in the 3 processes below:

1. mark "$" at the end of genome to generate the suffixes and repeatedly move "$" until it is in the first position;
2. sort all suffixes by lexicographic order;

3. take the last column into the BWT and start positions of the lexicographically sorted suffixes into suffix array.

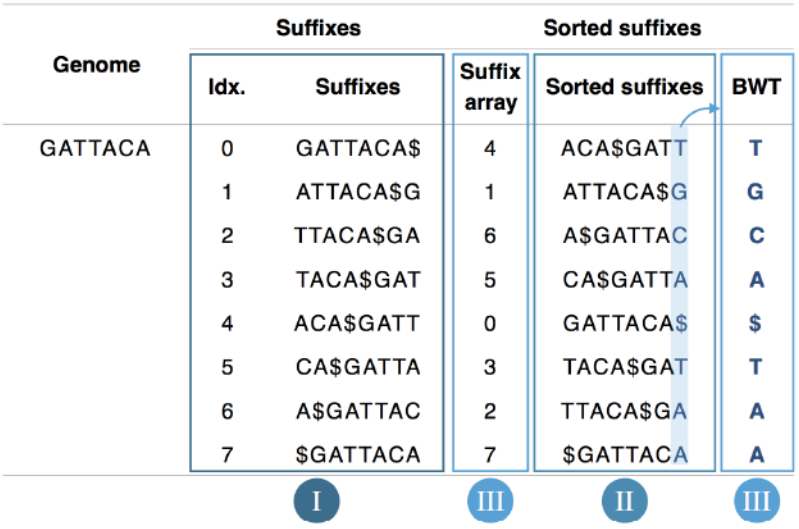


Figure 2.5: **The suffix array indexing data in three steps.**

With this data structure, the order of the original genome is updated in such a way that the character in BWT column comes before the one in the suffix array column. Let us consider the example provided in Figure 2.5. If the first position of genome "GATTACA" starts at index 0, the position of T is 3, while it is in the first BWT row. It is followed by A, which is the 4th BWT position. Let us illustrate now how to search "GAT". This search is performed one by one character, from the end of the sequence until its first position, that is, "T", "A", and then "G". First, considering "T" in the sorted suffix column, its possible positions according to the suffix array are 3 and 2. Then, due to BWT, the possible characters before "T" are "T" (in position 2) and "A" (in position 1). We then save "A" (position of 1), since it corresponds to the second finding. By continuing the process, we find the character before "A" or at the position 1 in suffix array, which leads to "G" according to BWT, and this is the 3rd character found. We can finally conclude that the "GAT" word is present in the genome GATTACA [Schbath et al., 2012].

Alignment tools More than 60 tools were available in 2012 [Fonseca et al., 2012]. Some of them are presented in Table 2.2 together with their features [Bao et al., 2011, Schbath et al., 2012, Fonseca et al., 2012]. Most of them return alignment output in SAM format and support multi-threading. Almost all the presented tools accept mismatched data in alignment which is beneficial for mutation studies. Performance of alignment tools have been measured for bacterial genomes, see Table 2.3, and also for human genome, as listed in Table 2.4. According to these studies, SOAP is the fastest tool for large scale alignments, while SHRiMP is the slowest one, see Table 2.4. The best tools for bacterial genomes are Bowtie, BWA, and PerM, as their numbers of unmapped reads are lower than what can be found using the other tools. Taking all factors under consideration, BWA is the

best tool for bacterial genomes: it has very good performances in terms of accuracy, sensitivity, and memory usage, while its processing time is quite normal.

Table 2.3: **Mapping results for each program run on the 3 mismatches bacterial genomes** [Schbath et al., 2012].

		Reads uniquely retrieved		Reads with multiple hits		
Software	Unmapped reads	Nb[b]	No retr.[c]	Nb[b]	Nb hits mean [sd]	No retr.[c]
BWA	212	7,301,261	66	2,698,527	8.99 [40.62]	2
Novoalign	89	7,333,124	31,764	2,666,787	8.41 [36.54]	20,154
Bowtie	214	7,301,200	3	2,698,586	8.99 [40.63]	14
BFAST	172,360	7,280,690	39,839	2,546,950	4.09 [2.31]	94,442
SSAHA2	17	6,994,428	2,377,585	3,005,555	6.98 [24.32]	1,367,027
GASSST	363,984	7,025,957	17,266	2,610,059	9.03 [40.90]	8,774
PerM	185	7,301,206	8	2,698,609	8.99 [40.63]	41

Table 2.4: **Performance of alignment tools in the experiment of simulated illumina reads** [Bao et al., 2011].

Task	Tools [version]	Reads mapped	Reads mapped correctly	Processed time (m)	RAM[d] (GB)
SE[e]	Bowtie [0.12.5]	79.19%	79.05%	271.37	5.09
	BWA [0.5.9]	96.11%	92.54%	324.31	3.17
	Mosaik [1.1.0021]	78.50%	77.61%	315.26	20.61
	PASS [1.2]	19.30%	9.23%	177.95	18.69
	RMAP [2.05]	75.28%	75.08%	397.845	6.1
	SeqMap [1.0.13]	79.19%	76.11%	5049.433	8.01
	SHRiMP [2.0.1]	99.93%	96.28%	9389.71	~32
	SOAP [2.2]	79.19%	78.67%	96.61	8.25
	SSAHA2 [2.5.3]	-	-	-	-
PE[f]	Bowtie [0.12.5]	0.02%	0.01%	227.25	5.09
	BWA [0.5.9]	99.46%	98.35%	616.8	3.2
	Mosaik [1.1.0021]	78.52%	77.59%	576.8	20.67
	PASS [1.2]	-	-	-	-
	RMAP [2.05]	0.39%	0.03%	1399.29	30.35
	SeqMap [1.0.13]	-	-	-	-
	SHRiMP [2.0.1]	95.13%	94.34%	15846.21	~32
	SOAP [2.2]	62.51%	62.43%	116.27	12.63
	SSAHA2 [2.5.3]	97.84%	96.01%	2884.5	13.38

[b]Number of mapped reads
[c]The original position has not been retrieved
[d]random-access memory
[e]single-end reads mapping
[f]paired-end reads mapping

Table 2.2: **Global characteristics of the alignment tools.**

Tool	Version	Format	Algorithm	Threads	Gaps	Mismatches
BFAST	0.6.5a	SAM	Hash the ref.	yes	yes	yes
Bowtie	0.12.7	SAM	BWT-index	yes	no	yes
BWA	0.5.8	SAM	BWT-index	yes	yes	yes
GASSST	1.28	SAM	Hash the ref.	yes	yes	yes
Mosaik	1.1.0021	BAM	Hash the ref.	yes	yes	yes
MPscan		personal	BWT-suffix tree	no	no	no
Novoalign	2.06.09	SAM	Hash the ref.	yes	yes	yes
PASS	1.2	SAM	Hash the ref.	yes	yes	yes
PerM	0.3.9	SAM	Hash the ref.	no	no	yes
RMAP	2.05	BED	Hash the reads	no	no	yes
SeqMap	1.0.13	ELAND[a]	Hash the reads	yes	no	at most 5
SHRiMP2	2.0.1	SAM	Hash the reads	yes	no	yes
SOAP2	2.20	personal	BWT-index	yes	no	at most 2
SSAHA2	2.5.2	SAM	Hash the ref.	no	no	yes

[a] Efficient local alignment of nucleotide data

2.1.1.3/ DE NOVO ASSEMBLY

De novo assembly is a self-alignment of reads that focuses on overlapping regions. It merges overlapping sequence reads into contiguous sequences, or *contigs*, without any reference genome. General methods of such de novo approaches are string-based and graph-based ones.

The string-based method is known as Greedy algorithm too. It joins couples of reads that have the best overlap, and it repeats the process until the quality becomes good enough according to a predefined quality threshold. Even though Greedy method is simple, it is unable to deal with highest repetitive genomic sequences [Simpson and Pop, 2015].

The graph-based methods use the graph theory to represent read sequences as nodes, and their relationships as edges. Contigs are created thanks to some particular travellings on the obtained graph. Since they define the erroneous paths caused by repeats, they are able to reduce the number of incorrectly assembled contigs. The main graph-based methods are described hereafter.

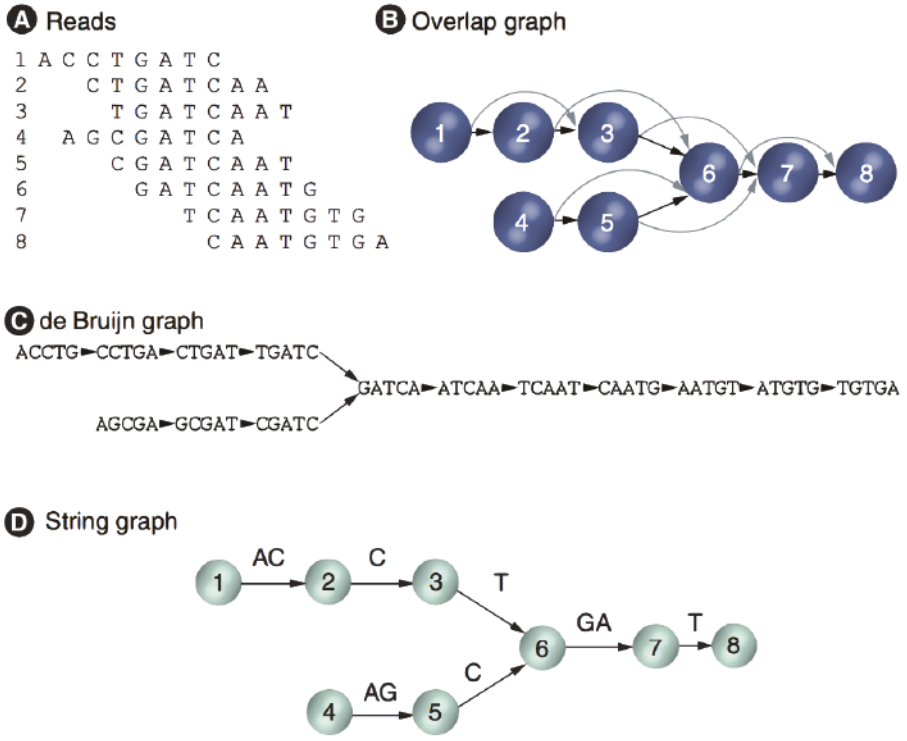


Figure 2.6: **Graph-based methods for genome assembly** [Henson et al., 2012].

OLC graph stands for "overlap, layout, and consensus". The detection of overlapping pairs of reads is firstly performed. Then, a proper ordering and orientation (layout) of the reads is achieved from the graph. Finally, layouts of the

reads are used to compute the consensus sequence.

Figure 2.6B depicts the OLC graph structure of the 8 sequence reads provided in Figure 2.6A. Each sequence read is a node. A pair of nodes that overlaps in at least 5 bases (in our illustration example) will be linked by an edge. For instance, the first read is linked to the second one due to 6 overlapping bases (CTGATC), and it is also transitively linked to the third read thanks to 5 overlapping bases (TGATC). This method is able to speed up the overlap detection via a k-mers indexation. However, if high-throughput short read data contain numerous overlaps, this will increase the graph size. This problem can be solved using de Bruijn graph described below [Simpson and Pop, 2015].

De Bruijn graph: nodes are k-mers while edges are couple of reads containing the continued k-mers.

Let us illustrate the De Bruijn graph on the same example as previously. The 3 first sequence reads from Figure 2.6A are transformed to the following 5-mers: ACCTG → CCTGA → CTGAT→ TGATC, which are connected in a path, see Figure 2.6C. The path depicted below, namely AGCCA → GCGAT → CGATC, corresponds for its part to the reads number 4 and 5. These two paths, corresponding to the reads from 1st to 5th, include too the 6th read. Thus they are linked to GATCA node, and the latter is finally followed by the paths of k-mers ranging from the 6th read to the 8th one, as depicted in Figure 2.6C.

This method is able to reduce the memory consumption, as its nodes contain not the reads but subsequences of length k called k-mers. The de Bruijn graph is thus an efficient method for short reads assembly of bacterial genomes [Simpson and Pop, 2015]. It needs however additional algorithms to store k-mers of large genome assembly, like hash table, bit array, and FM-index [Simpson and Pop, 2015].

String graph is an hybrid of string-based and graph-based methods. It takes the property of repeats reduction from the de Bruijn graph, and then it applies it with the overlap-based method. Therefore, in the string graph, it is not necessary to break the reads into k-mers, while some interesting properties of de Bruijn are still preserved.

In Figure 2.6D, each node contains a read from Figure 2.6A, while each edge represents the extended bases between the corresponding nodes. For example, “AC”, the label on the edge between the 1st node and the 2nd one, is the extension of 1st read when compared with the 2nd one.

Let us finally remark that, in the case of large genomes, the efficiency of this method can be improved by applying the FM index [Simpson and Pop, 2015].

De novo tools are presented in Figure 2.7. They have been developed during years 2005-2010, by following 4 major strategies: Greedy, OLC graph, de Bruijn graph, and String graph.

Magoc et al. [Magoc et al., 2013] have evaluated the assembly software in the bacterial genome case by measuring the "corrected N50 size", see Table 2.5. The N50 size is the size of the smallest contig, such that that 50% of the genome is contained in contigs of size at least equal to N50. And the corrected N50 size is the N50 size obtained after splitting contigs at each error. The results shown that MaSuRCA [Zimin et al., 2013], which is a string graph-based method, strongly provides the best results on HiSeq reads (3 of 3 species), while it provides good results on MiSeq reads (similarly to SPAdes [Bankevich et al., 2012], which is a de Bruijn graph-based method).

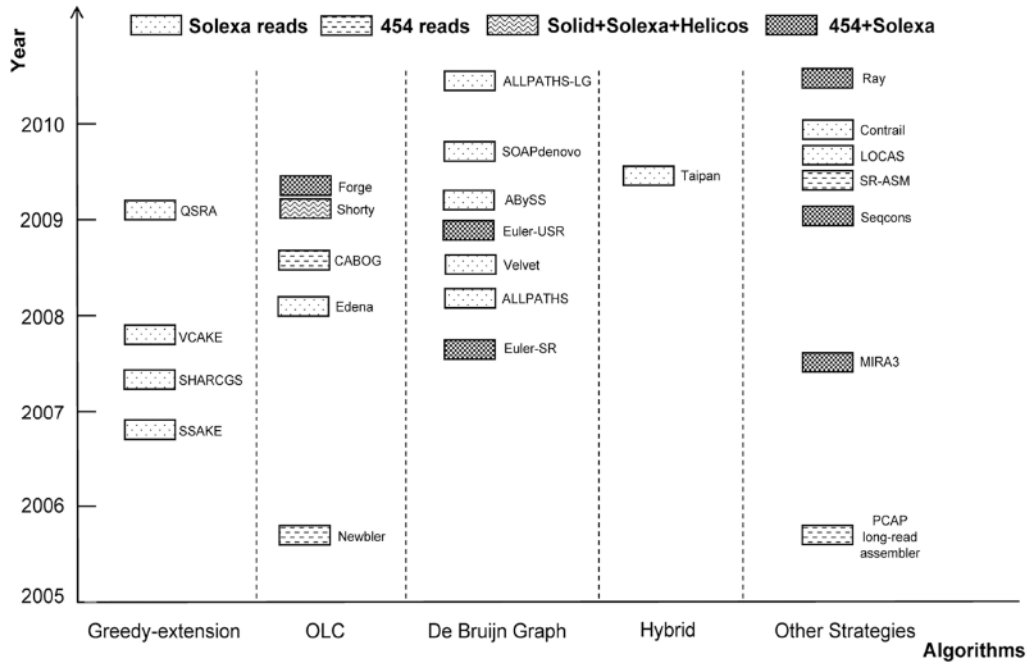


Figure 2.7: **Tools for de novo assembly which are implemented in 2005-2010** [Zhang et al., 2011].

Table 2.5: **Comparison of corrected N50 contig sizes, shown in kilobases** [Magoc et al., 2013].

Assembler	Species assembled						
	HiSeq (100 bp) reads				MiSeq (250 bp) reads		
	R.sphaeroides	*M.abscessus*	*V.cholerae*	*B.cereus*	*R.sphaeroides*	*M.abscessus*	*V.cholerae*
ABySS	13.0	115.7	93.0	130.6	21.4	68.5	60.3
CABOG	11.2	78.2	48.8	150.5	30.5	8.3	32.5
MIRA	17.7	129.2	87.1	100.0	15.4	75.0	108.7
MaSuRCA	**176.8**	**194.0**	**236.4**	**246.7**	**130.7**	36.2	71.6
SGA	12.1	27.9	23.4	25.5	9.1	12.8	27.3
SOAPdenovo	10.5	147.2	106.5	**246.3**	33.5	113.3	65.5
SPAdes	83.5	147.9	77.1	103.7	118.1	**215.4**	**246.6**
Velvet	13.1	60.3	39.5	24.5	24.2	41.5	67.1

2.1.2/ PHYLOGENETIC ANALYSIS

Any phylogenetic analysis mainly contains 3 steps, which are the multiple alignment of considered sequences, the selection of a mutation model, and the discovery of the best evolutionary tree that is able to recover the alignment profile, considering the mutation model. These three steps are detailed hereafter. Note that the state-of-the-art part related to multiple sequence alignment and their use for phylogenetic analysis has been studied in common with my colleague Bashar Talib, and written "four hands" as our investigations in this field have been performed together, in team.

2.1.2.1/ MULTIPLE SEQUENCE ALIGNMENT (MSA)

Multiple sequence alignment is an expansion of pairwise alignment to combine more than two sequences at a time. They are implemented to identify conserved regions among a set of sequences, evaluating by doing so if they are evolutionarily related. Alignments are also used to help in building evolutionary relationships on phylogenetic trees construction, as described in the next section.

The goal of MSA is to align all of the sequences in a given set if possible. A MSA is thus a collection of three or more nucleotide or amino acid sequences that are aligned partially or entirely. Identical residues are aligned in columns across the length of the sequences. These aligned residues are homologous in a fundamental sense or even in an Evolutionary one: they have probably derived from a common ancestor.

Figure 2.8 is an example of the result of MSA applied on Mycobacterium tuberculosis. However, as soon as the sequences exhibit some divergence, the problem of multiple alignments becomes extraordinarily difficult to solve. And if exact approaches produce optimal alignments, they are not feasible in time or space for more than a few sequences. Let us finally notice that the alignment accuracy can be hard to estimate and their actual biological significance can be ambiguous.

In practice, a very popular progressive sequence alignment tool is the Clustal family [Higgins and Sharp, 1988], in particular the weighted variant ClustalW that is incorporated in many web tools like GenomeNet[7] or EBI[8]. Another important progressive alignment approach is called T-Coffee [Notredame et al., 2000], which operates as a post processing on various MSAs of the same set of sequences that are provided by other existing methods like Clustal. Due to its principle of conception, T-Coffee is slower than Clustal and its derivatives but, in general, it yields more accurate alignments for distantly related sequence sets.

[7] http://align.genome.jp/
[8] http://www.ebi.ac.uk/clustalw

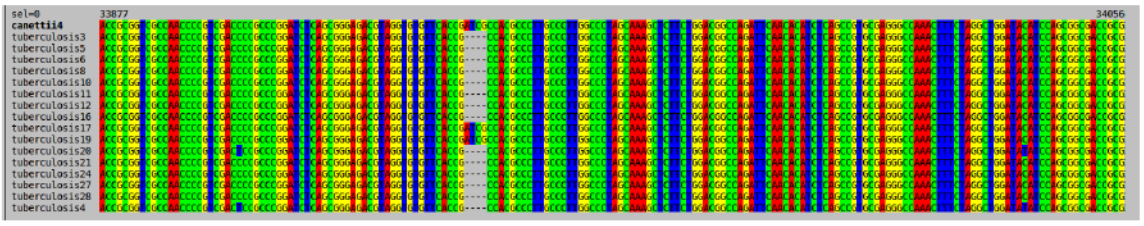


Figure 2.8: **Multiple sequence alignment of various sequences of *Mycobacterium tuberculosis*.**

2.1.2.2/ NUCLEOTIDE SUBSTITUTION MODEL AND TOOL

Substitution models Over time, nucleotide sequences can ”evolve” through substitution. This process can cause a nucleotide (A, C, T or G) to change into another nucleotide, and this is one of the most central driving force behind evolution. This modification in a DNA sequence may lead to an inactivation of a gene or to a mutation in the protein that the sequence codes. As proteins are the building blocks of organic life, this may cause significant variations in an organism's characteristics. Alternatively, this modification may have no effect at all, being silent.

Estimating the evolution of organisms over hundreds of millions of years, models of nucleotide evolution are helpful in speculating how one sequence of nucleotides may have evolved from another. These models can be inferred by either assuming that two given sequences had shared a common DNA ancestor or by assuming that one sequence evolved into the other.

At the simplest level, the proportion can be used to define such a matrix P of nucleotide substitution:

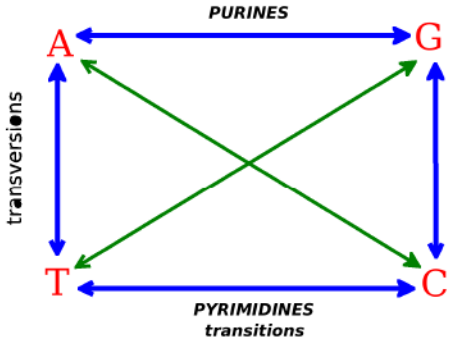
$$P_d = \frac{n_d}{n} \tag{2.1}$$

where n indicates the total number of nucleotides in the sequence, and n_d is the number of base d, where $d \in \{A, C, G, T\}$. Other richer probabilistic models have been proposed in the literature, to provide a more accurate estimation of the mutation matrix P, like Jukes and Cantor [Wilson and Sarich, 1969], Kimura [Kimura, 1980], and Tamura and Nei [Tamura and Nei, 1993]. Some of them are detailed hereafter.

A first model for genome evolution was proposed in 1969 by Thomas Jukes and Charles Cantor (JC model [Wilson and Sarich, 1969]). This first model is very simple, as it supposes that each nucleotide A, C, G, T has the same probability a to mutate to any other nucleotide, as shown in Figure 2.9.

$$\textit{Jukes Cantor}: a = \frac{-3}{4}\ln(1 - (\frac{4p}{3})) \tag{2.2}$$

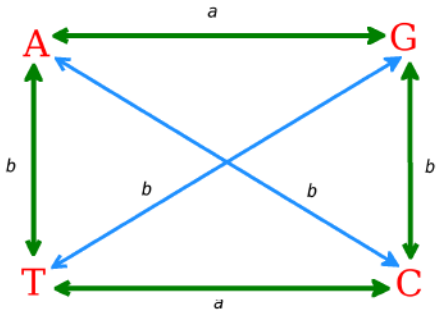
In 1980, Motoo Kimura [Kimura, 1980] proposed to consider different parameters for transversions and transitions. Let us recall that a transition is a nucleotide



	A	T	C	G
A		a	a	a
T	a		a	a
C	a	a		a
G	a	a	a	

Figure 2.9: **The Jukes & Cantor (JC) model assumes that all the substitution probabilities are equal.**

change between purines A and G, or between pyrimidines T and C, while a transversion is a purine ↔ pyrimidine change. Kimura thus proposes to consider a transition rate α (per unit time) and a different transversion rate β, as shown in Figure 2.10.



	A	T	C	G
A		b	b	a
T	b		a	b
C	b	a		b
G	a	b	b	

Figure 2.10: **The Motoo Kimura model assumes different rates of transitions and transversions.**

Since then, nucleotide substitution models have been developed into many different types that are biologically plausible. Among them, the most general nucleotide substitution model is the general time-reversible one (GTR model [Tavaré, 1984]), which has unequal base frequencies and also involves 9 parameters, including 3 different stationary probabilities and 6 different substitution rates among the four A, C, G, T states. Finally, as some researchers considered that some nucleotide sites may be invariant over time, it has been proposed to model the substitution rates of the variant sites by a Gamma distribution [Tamura and Nei, 1993, Nei and Gojobori, 1986, Yang, 1994]. This leads to an extension of the GTR model, which is called GTR+I+Γ. I stands for the proportion of invariant sites that fits the stable dataset, while Γ stands for gamma distributed rates among the variant sites that are able to model heterogeneity. Thus I and Γ are negatively correlated, as Γ reflects the absence of I in simulations [Lemey et al., 2009].

Note that, instead of considering the substitution rate among nucleotide symbols (A, T, C, G), we can focus on amino-acids for proteins as shown in Figure 2.11. Let us recall that each combination of three nucleotides, called a codon,

encodes one amino acid. Since there are 64 codons and only 20 amino acids, the genetic code is thus redundant, and a mutation in a nucleotide does not necessarily lead to a mutation in the amino-acid. This explains why larger (20×20) mutation matrices of amino acids have been investigated too, to refine the information provided by the nucleotide mutation ones[9].

As substitution models provide a mutation rate per time unit, they can be used to measure the "distance" between two sequences, estimating the date when they started to diverge from a common ancestor. And given a set of sequences, grouping them 2 by 2 according to their proximity, and putting at link node the sequence at mid-distance of both, leads to first ideas on how to infer a phylogenetic tree. This is why, in order to estimate the species tree, a model for the evolution of SNPs over time is needed. It is applied on aligned sequences, that are described below.

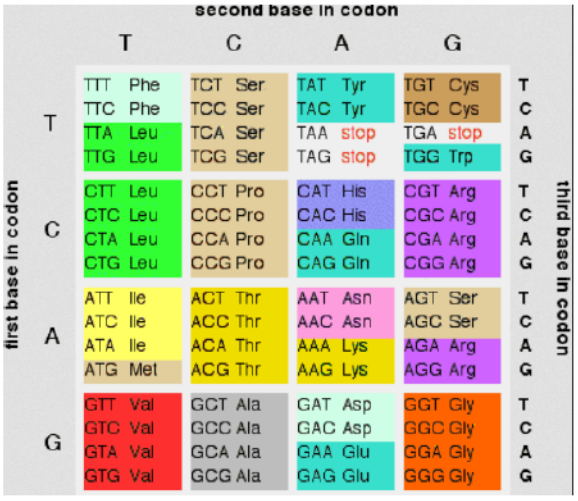


Figure 2.11: **The genetic code is the universal system that assigns amino acids according to codons** [Godfrey-Smith and Sterelny, 2008].

JModeltest According to their own website, jModelTest [Darriba et al., 2012] is a tool to carry out statistical selection of best-fit models of nucleotide substitution. Five different model selection strategies are embedded, namely: hierarchical and dynamical likelihood ratio tests, Akaike and Bayesian information criteria (as known as AIC and BIC), and a decision theory method. It also provides estimates of model selection uncertainty, parameter importance and model-averaged parameter estimates, including model-averaged tree topologies.

jModelTest 2, for its part, includes high performance computing capabilities. Additional features are embedded too, like heuristic filtering, new strategies for tree optimization, model-averaged phylogenetic trees (both topology and branch length), and automatic logging of user activity. For further information, see, *e.g.*, [Darriba et al., 2012]

[9] jModeltest 2.0 [Posada, 2008] can be used to estimated the best-fit substitution model for a given data set, see a further section.

2.1.2.3/ PHYLOGENETIC TREE AND PHYLOGENETIC NETWORKS

About phylogenetic trees An evolutionary or phylogenetic tree is an acyclic graph (or branching diagram, see Figure 2.12) that is used to emphasize evolutionary relationships among groups of biological species, that is, their phylogeny based upon similarities and variations in their genetic or physical characteristics. The nodes are connected in the tree by branches. The latter highlight the relationship between taxonomic units (TU) at the leaves of the tree and their ancestors, corresponding to the internal nodes of the graph. Each branch has a length that represents, for example, the number of expected mutations per site (in amino-acids or nucleotide sequences) that have probably happened in these branches, in a sequence based phylogeny. Thus, branch lengths provide the time of variation between two organisms and their common ancestor. There are basically two kinds of phylogenetic trees:

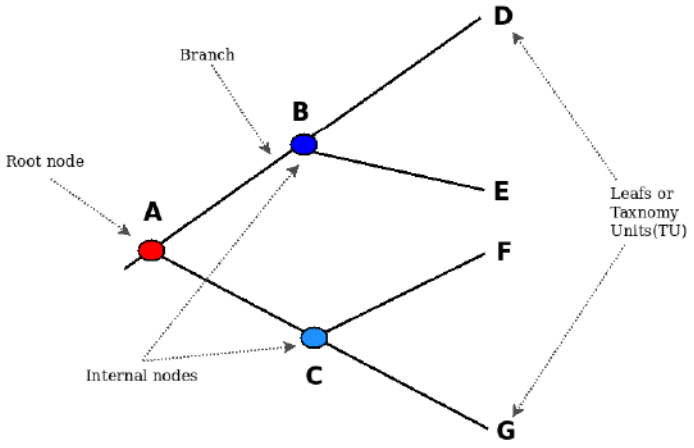


Figure 2.12: **Example of a phylogenetic tree structure.**

- **The unrooted Phylogenetic Trees:** evaluate the relationships between all the given TUs. However, they usually do not provide sufficient information to deduce the evolution from the last common ancestors.

- **The rooted Phylogenetic Trees:** embed a root node that represents the last common ancestor of all TUs in the tree. The main way to root a tree is to specify an *outgroup*, which is a TU known to be outside the group of TUs under consideration. This latter can be a species known to have diverged before the divergence of the considered TUs.

  It can be noticed that the time of evolution of a rooted species represented by a rooted phylogenetic tree can be computed from each sub-ancestor to the last common one when either the date of divergence or the divergence rate are known. Until now, however, this question is still an intensive subject of research. For further information, see, *e.g.*, [Soltis et al., 1999, Pevsner, 2005, Eaton and Ree, 2013].

The most known and commonly used methods of tree construction can be classified into two central divisions: *distance-based* and *character-based* methods.

Distance-based methods begin by transforming the original data into a matrix of pairwise distance values. The next stage is to infer a tree either by sequential joining approaches, or by estimating a set of candidate trees and applying a type of optimality criterion technique to select the best one. Under the minimum evolution criterion, the tree that has a minimum sum of branch lengths is selected as the best estimate. Distance-based algorithms encompass *UPGMA* and *Neighbor-Joining*, this latter being explained in the next subsection.

Character-based methods can depend on a divergence of phylogenetic characters such as genetic and molecular attributes to construct phylogenetic trees. As long as that there is divergence among taxa in the characteristic and that the characteristic is heritable, it could probably be accepted as a phylogenetic character. Protein, DNA, and RNA sequences can be considered as phenotypic attributes. Molecular phylogenetics attempts to estimate the rates and patterns of change occurring in the sequences and to reconstruct the evolutionary history of organisms.

Algorithms used to create phylogenetic trees using characters are more complicated than distance-based methods [Felsenstein, 1988]. The algorithms are based on an optimization criterion such as *Maximum Likelihood*, *Maximum Parsimony*, or *Bayesian* methods in order to find the best tree according to the considered characters. For the sake of illustrations, we will detail such methods at the end of this chapter.

Neighbor-Joining Algorithm As previously said, the neighbor-joining algorithm constructs unrooted phylogenetic trees using distance methods. Both topology and branch lengths are computed by iteratively specifying (based on a distance matrix) a neighbor as a pair of TUs that are joined in a single internal node X in an unrooted tree, depending on the previously computed distance matrix. An iteration of the neighbor-joining algorithm consists of the following steps:

1. Construct an unresolved tree with all TUs in a starlike structure with no hierarchy.
2. Construct a distance matrix by pairwise comparison and calculate the value of branch lengths, to identify the two most related sequences (TUs): NJ seeks to build a tree that minimizes the sum of all branch lengths.
3. Determine which TUs are connected to an internal node X. They are treated now as one single new TU.
4. Join the closest neighbors (TUs with similar characters), that is, the base pair that has the smallest sum-of-branch-lengths.
5. The algorithm is repeated until the topology of the tree is obtained.

The neighbor-joining method produces an unrooted tree. The sum of the branch lengths of N TUs in the tree is calculated as follows. Let us define D_{ij} and L_{AB} as the distance between TUs i and j and the branch length between nodes A and B respectively. The sum of branch lengths of the tree is defined based on the following formula:

$$S = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij/}$$

The distance between nodes X and Y is calculated as follows:

$$L_{XY} = \frac{1}{2(N-2)} \left[\sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^{N} L_{iY} \right].$$

The term inside the brackets is the sum of all distances including L_{XY}, and the outer term $\frac{1}{2(N-2)}$ is to eliminate unrelated branch lengths. For more information, see [Saitou and Nei, 1987, Bruno et al., 2000].

Neighbor-joining [Saitou and Nei, 1987] is a method which is especially suited for datasets comprising lineages with broadly varying rates of evolution. It can be used in combination with techniques that allow correction for superimposed substitutions.

Maximum Parsimony The maximum parsimony method [Tateno et al., 1994] aims at minimizing branch lengths by reducing the number of mutations. This approach predicts the evolutionary tree that minimizes the number of actions needed to generate the marked variation in the sequences from common ancestral sequences. In a maximum parsimony phylogenetic study, for given sequences, a MSA algorithm is used to align the sequences, and to identify the informative positions, that is, columns in the multiple sequence alignment with no gap and at least two characters.

The next step is to count the number of changes and assign this cost to each generated phylogenetic tree. The method then computes the total length L for each tree, which is calculated according to the following formula:

$$L = \sum_{j=1}^{C} w_j l_j,$$

where l_j is the cost for character j, C is the total number of characters, and w_i is the assigned weight for each character, which is set to 1 in most cases. The tree that maximizes this L value is finally selected.

Bayesian methods For the sake of completeness, we evoke here the well-known and frequently used Bayesian methods [Huelsenbeck et al., 2001], which estimate the phylogeny by calculating the conditional probability given the model,

based on the following formula:

$$Pr[Tree|Data] = \frac{\Pr[Data|Tree] \times \Pr[Tree]}{\Pr[Data]}$$

where $Pr[Tree|Data]$ is called a posterior probability distribution[10].

Bayesian methods can thus apply a model of sequence evolution and are ideal for building a phylogeny using sequence data.

Maximum Likelihood The Maximum Likelihood (ML) criteria requires a probabilistic model for the evolutionary process and finds the most likely tree, given the probabilistic model and the known sequences at the leaves. In other words, ML techniques are used to determine the topology and branch lengths that have the largest likelihood to produce the aligned data, providing the substitution model and the tree. The likelihood value is computed after the alignment stage and by considering some DNA or amino acids substitution models.

In ML method, the searching space is fulfill using a quartet program. This latter finds all possible sequence combinations for tree reconstruction, while the Maximum Parsimony criteria prefers solutions that minimize the number of mutations along the tree edges [Addario-Berry et al., 2004]. Bayesian methods and Maximum likelihood can apply a model of sequence evolution and are ideal to construct a phylogeny using data sequences. The main drawback of these methods is that they are computationally expensive. However, with today's computers, this is not too much a problem.

One of the most common tests used to evaluate the reliability of a deduced tree is the so-called Felsenstein's bootstrap test [Felsenstein, 1985], which is usually estimated using Efron's bootstrap resampling technique [Efron et al., 1996]. It is accomplished in practice by sampling the input data [Soltis et al., 2003] and measuring the proportion of deduced trees that support each branch of the best tree previously obtained. As a global rule, if the bootstrap value for a given internal branch is 95% or higher, the topology will be considered "valid" at that branch[11].

Phylogenetic networks and NeighborNet If phylogenetic trees are the usual mathematical models to represent the evolution of species or genes, they are not able to represent events like recombination, hybridization, or even horizontal gene transfer. When such events occur, some branches of the tree must combine into a reticulation node, and by doing so, the tree becomes a network. The ways to construct such networks depends on the available data. Such data can be of different kinds: sequences, distances between these sequences, rooted or unrooted trees, triplets (rooted trees on three leaves), quartets (unrooted trees

[10] A posterior probability is the probability that the tree is considered to be correct, if it has the maximum probability.

[11] In Bayesian approaches, this is the posterior probability itself that gives an evaluation of robustness: the support of a branch increases with its probability.

on four leaves), splits (bipartitions of the set of leaves), clusters (subset of leaves which should appear together in the network), and so on.

Various problems are related to phylogenetic networks, like how to graphically represent them or how to compare two networks. Some of them are polynomial while others are NP-complete ones. Fixed-parameter tractable, approximation algorithms, or heuristics have thus been developed to solve them. Depending on the problem addressed, and its complexity, the network used in the analysis may be an explicit phylogenetic network if it describes biological events, or an abstract phylogenetic network if its edges may not be interpreted biologically. A popular example of such abstract networks is the split network, which can be used as a visualization of a set of incompatible phylogenetic trees, therefore giving some understanding on the conflicts present in the data.

A well-known algorithm for constructing phylogenetic networks is the so-called NeighborNet [Bryant and Moulton, 2002], which is loosely based on the neighbor joining algorithm presented previously. Like neighbor joining, the method takes a distance matrix as input, and works by agglomerating clusters. However, the NeighborNet algorithm can lead to collections of clusters that overlap and do not form a hierarchy, and are thus represented using a split network. This method is implemented in the SplitsTree package [Huson and Bryant, 2006]. For further information, see, *e.g.*, [Huson et al., 2011].

2.1.2.4/ Time to the Most Recent Common Ancestor

The Time to the Most Recent Common Ancestor (TMRCA) is the amount of time, or number of generations, since individuals have shared their last common ancestor. Since mutations occur at random, the estimate of the TMRCA is not an exact number (*i.e.*, ten generations) but rather a probability distribution. Indeed, the TMRCA estimate becomes more refined when more information is compared.

In general, the divergence date analysis requires 2 elements of calibrations [Ho and Phillips, 2009].

- Calibrations of nodes in the evolutionary tree, which can be divided in: (i) terminal nodes for heterochronous sequence data analysis or including dated fossils in morphological character analysis, (ii) internal nodes (divergence of coalescent events) analysis that uses fossil record or dated biogeographic events together with an *a priori* knowledge of taxa relationships.
- Calibration of substitution rate which can be estimated from molecular clock model.

Interestingly, the structured coalescent method, which is in MultiTypeTree template of BEAST2, is valuable for epidemiological study since spatial data is taken into account to estimate TMRCA. Further additional assigns can be required for the general calibrations, such as the location of terminal nodes and a migration model for the internal ones [Müller and Vaughan, 2017].

To sum up, the estimation of the divergence time can be computed using BEAUti and BEAST [Bouckaert et al., 2014]. After having imported the sequences and computed their alignment, we have to set the dates of the taxa (by default, they have a date of zero, which will be correct if all the sequences were sampled at approximately the same point in time). Then, the evolutionary model must be chosen with its substitution rates, for instance provided by JModelTest2. Next, clock model and priors should be specified. Finally, the Markov chain Monte Carlo (MCMC) method can be run, to find the searched TMRCA.

2.1.3/ Comparative Genomic

In this section, some well-known tools in comparative genomics are recalled with their functioning. The objective here is not to be exhaustive, but to focus on tools we finally selected after various intensive tests.

2.1.3.1/ Genome alignment using MUMMER

MUMmer [Delcher et al., 1999b] is a system for rapidly aligning entire genomes. Using an efficient data structure called a suffix tree (*cf.* a previous section), the system is able to rapidly align sequences containing millions of nucleotides. MUMmer can also align incomplete genomes: as explained in its website, MUMmer can easily handle the 100s or 1000s of contigs from a shotgun sequencing project, and will align them to another set of contigs or a genome using the NUCmer program included with the system. If the species are too divergent for a DNA sequence alignment to detect similarity, then the PROmer program can generate alignments based upon the six-frame translations of both input sequences. MUMmer4 is planned to be released in 2017, in which the NUCmer program can handle genomes of unlimited size and runs multi-threaded.

2.1.3.2/ Gene search: BLAST, BLAT, and GMAP

BLAST Basic Local Alignment Search Tool (BLAST) is a database sequence search engine proposed by the National Center for Biotechnology Information (NCBI). The first version of BLAST was published in 1990 and it supported only ungapped searches. The second version, released in 1997 [Altschul and Gish, 1996], has been designed to determine high-scoring local alignments between sequences, without discrediting the speed of such searches. BLAST addresses thus an essential problem in bioinformatics research. It uses a heuristic process that attempts local as crossed to global alignments and, therefore, it is suitable to identify relationships between sequences (aminoacid sequences of proteins or the nucleotides of DNA sequences) which share only isolated regions of similarity [Altschul et al., 1990].

Table 2.6 displays the different BLAST programs available on the NCBI web server.

Table 2.6: **BLAST programs.** http://www.ncbi.nlm.nih.gov/BLAST/

Program	Comparison	Application
BLASTN	DNA vs. DNA. Compares a nucleotide query sequence against a nucleotide sequence database.	Find DNA sequences that match the query
BLASTP	Protein vs. Protein. Compares an amino acid query sequence against a protein sequence database.	Find identical (homologous) proteins
BLASTX	DNA vs. Protein. Compares a nucleotide query sequence translated in all reading frames against a protein sequence database.	Find protein databases using a translated nucleotide query
TBLASTN	Protein vs. DNA. Compares a protein query sequence against a nucleotide sequence database dynamically translated inall reading frames.	Find genes in unknown DNA sequences
TBLASTX	DNA vs. DNA. Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.	Find degree of homology between the coding region of the query sequence and known genes in the database.

BLAT BLAT [Kent, 2002], which stands for "BLAST-like alignment tool", is similar in many ways to BLAST. The program rapidly scans for relatively short matches and extends these into high-scoring pairs. However, BLAT differs from BLAST in some significant ways:

- BLAST builds an index of the query sequence and then scans linearly through the database. Conversely, BLAT builds an index of the database and then scans linearly through the query sequence.
- Where BLAST triggers an extension when one or two hits occur in proximity to each other, BLAT can trigger extensions on any number of perfect or near-perfect hits.
- BLAST returns each area of homology between two sequences as separate alignments, while BLAT stitches them together into a larger alignment.
- Finally, BLAST delivers a list of exons sorted by size, with alignments extending slightly beyond the edge of each exon. BLAT, for its part, effectively "unsplices" mRNA onto the genome – giving a single alignment that uses each base of the mRNA only once, and which correctly positions splice sites.

GMAP GMAP [Wu and Watanabe, 2005] is a standalone integrated genomic mapping and alignment program for cDNA sequences – both messenger RNAs (mRNAs) and expressed sequence tags (ESTs) – mapped and aligned to a genome. Compared to the state of the art, this program shows improved performances (in terms of speed and accuracy) and enhanced functionality, making it possible to:

1. map and align a single cDNA interactively against a large genome in about a second, without the startup time of several minutes typically needed by existing mapping programs;
2. switch arbitrarily among different genomes, without the need for a preloaded server dedicated to each genome;
3. run the program on computers with as little as 128 MB of RAM (random access memory);
4. perform high-throughput batch processing of cDNAs by using memory mapping and multithreading when appropriate memory and hardware are available;
5. generate accurate gene models, even in the presence of substantial polymorphisms and sequence errors;
6. locate splice sites accurately without the use of probabilistic splice site models, allowing generalized use of the program across species;
7. detect statistically significant microexons and incorporate them into the alignment; and finally
8. handle mapping and alignment tasks on genomes having alternate assemblies, linkage groups or strains.

The reader is referred to [Wu and Watanabe, 2005] for further information.

2.1.3.3/ Insertion Sequences (IS) tools and database

Basic recalls Let us firstly recall that insertion sequences are relatively short and genetically compact DNA segments (between 0.7 and 3.5 kbp) encoding no function other than those involved in their mobility. Many, but not all, carry short (<40 bp) imperfect inverted repeats (IR) at their ends and generate a small (between 2 and 14 bp) duplication of the target DNA flanking the point of insertion (DR, see below). At present, an estimated 3500 different ISs have been detected from both bacteria and archaea [Hickman et al., 2010, Zhou et al., 2008]. They have been observed in most bacterial genomes and plasmids where they may be present in high numbers [Siguier et al., 2006b, Siguier et al., 2012]. IS elements have an important role in the diversification and evolution of bacterial genomes [Ooka et al., 2009].

ISs are classified in families (see Figure 2.13, Tables 2.7, 2.8 [Siguier et al., 2014]) using a large variety of characteristics [Mahillon and Chandler, 1998], including the following ones:

1. the length and sequence of the short imperfect terminal inverted repeats (IRs) carried by many ISs at their ends (TIRs or ITRs in eukaryotes);

2. the length and sequence of the short flanking direct target DNA repeats DRs (TSD, target site duplication, in eukaryotes) often generated on insertion;
3. the organization of their open reading frames;
4. the target sequences into which they insert.

However, the principal factor in IS classification is the similarity [Siguier et al., 2014], at the primary sequence level, of the enzymes which catalyze their movement, namely their transposases (Tpases) – see “Major IS groups are defined by transposase type” in Figure 2.13.

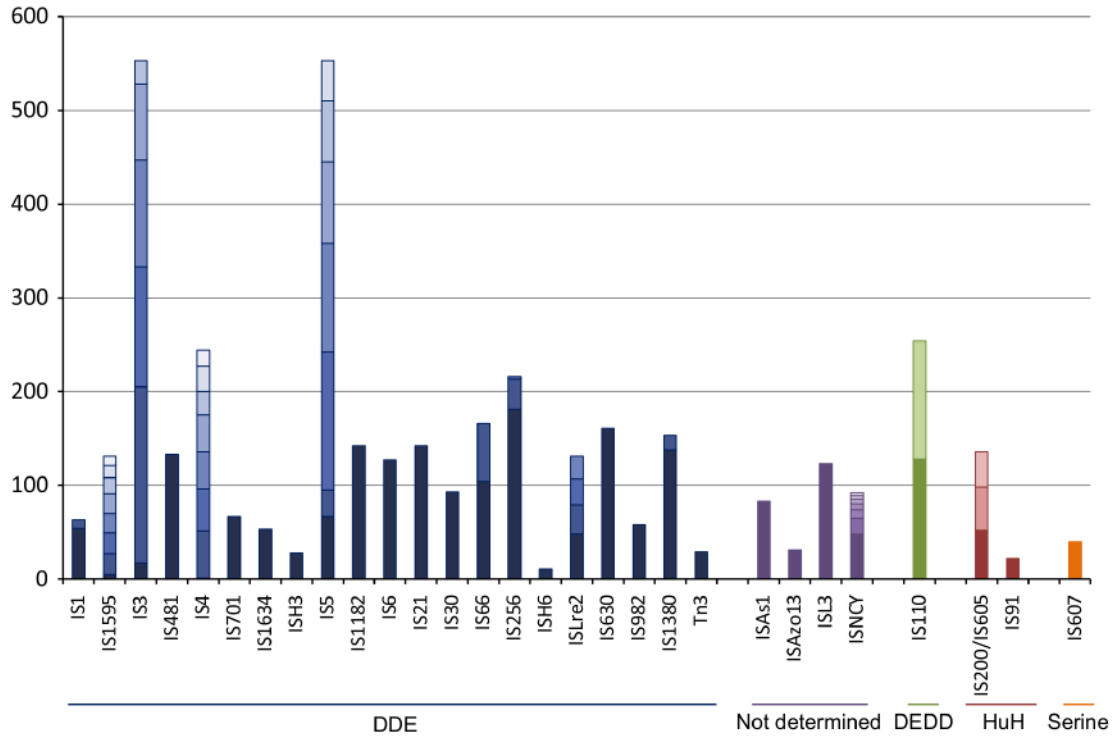


Figure 2.13: **Distribution of IS families in the ISfinder database.** The histogram shows the number of IS of a given family, as defined in the text, in the ISfinder database (June 2013). The horizontal boxes indicate the number and relative size of different subgroups (see Table 1 for the subgroup names) within the family. They are grouped by color to indicate the type of Tpase used: DDE, blue; undetermined, purple; DEDD, green; HUH, red; and Serine, orange.

To the best of our knowledge, there are only six tools that deal with detecting IS elements within prokaryotic genomes. ISA is not publicly available, while IScan, ISsaga, and OASIS are downloadable. The biggest IS database of bacteria and six state-of-the-art tools are presented thereafter.

ISFinder This is a database dedicated for bacterial insertion sequences developed in 2006 by P. Siguier *et al.* These sequences represent one of the

Table 2.7: **General characteristics of IS families.**

Family	Subgrps	Size-range	DR(bp)	Ends	IRs	Nb-ORF	ORF	Catalytic residus
IS1	-	740-1180	8-9	GGnnnTG	Y	2	ORFAB	DDE
IS1	Single ORF	800-1200	0-91					DDE
IS1	ISMhu11	900-4600	0-10		Y	2	ORFAB	DDE
IS1595	ISPna2	1000-1150	8	GGCnnTG	Y	1		DDNK
IS1595	ISPna2+pass	1500-2600	8					DDNK
IS1595	ISH4	1000	8	CGCTCTT		1		DDNK
IS1595	IS1016	700-745	7-9	GGGgctg		1		DDEK
IS1595	IS1595	900-1100	8	CcTGATT		1		DDNK+ER4R7
IS1595	ISSod11	1000-1100	8	nnnGcnTATC		1		DDHK+ER4R7
IS1595	ISNwi1	1080-1200	8	ggnnatTAT		1		DDEK+ER4
IS1595	ISNwi1+pass	1750-4750	8			1		DDEK+ER4
IS1595	ISNha5	3450-7900	8	CGGnnTT		1		DDER/K
IS3	IS150	1200-1600	3-4	TG	Y	2	ORFAB	DDE
IS3	IS407	1100-1400	4	TG				DDE
IS3	IS51	1000-1400	3-4	TG				DDE
IS3	IS3	1150-1750	3-4	TGa/g				DDE
IS3	IS2	1300-1400	5	TG				DDE
IS481	-	950-1300	4-15	TGT	Y	1		DDE
IS4	IS10	1200-1350	9	CT	Y	1		DDE
IS4	IS50	1350-1550	8-9	C				DDE
IS4	ISPepr1	1500-1600	7-8	-T-AA				DDE
IS4	IS4	1400-1600	10-13	-AAT				DDE
IS4	IS4Sa	1150-1750	8-10	CA				DDE
IS4	ISH8	1400-1800	10	CAT				DDE
IS4	IS231	1450-5400	10-12	CAT		1or+ (Passenger genes)		DDE
IS701	-	1400-1550	4		Y	1		DDE
ISH3	-	1225-1500	4-5	C-GT	Y	1		DDE
IS1634	-	1500-2000	5-6	C	Y	1		DDE

Table 2.8: **General characteristics of IS families (continued).**

Family	Subgrps	Size-range	DR(bp)	Ends	IRs	Nb-ORF	ORF	Catalytic residus
IS5	IS903	950-1150	9	GG	Y	1		DDE
IS5	ISL2	850-1200	2-31					DDE
IS5	ISH1	900-1150	8	-GC		1		DDE
IS5	IS5	1000-1500	4	Ga/g		1		DDE
IS5	IS1031	850-1050	3	GAa/g		1		DDE
IS5	IS427	800-1000	2-4	Ga/g		2	ORFAB	DDE
IS1182	-	1330-1950	0-60		Y	1		DDE
IS6	-	700-900	8	GG	Y	1		DDE
IS21	-	1750-2600	4-8	TG	Y	2 (istB : helper of transposition)		DDE
IS30	-	1000-1700	2-3		Y	1		DDE
IS66	-	2000-3000	8-9	GTAA	Y	3		DDE
IS66	ISBst12	1350-1900	8-9	GTAA	Y	1		DDE
IS91	-	1500-2000	0		N	1		HUH/Y2
IS110	-	1200-1550	0		N	1		DEDD
IS110	IS1111				Y			DEDD
IS200/IS605	IS200	600-750	0		N	1		HUH/Y1
IS200/IS605	-	1300-2000				2		HUH/Y1
IS200/IS605	IS1341	1200-1500				1		HUH/Y1
IS607	-	1700-2500	0		N	2		Serine
IS256	-	1200-1500	8-9	Ga/g	Y	1		DDE
IS630	-	1000-1400	2		Y	1 or 2	ORFAB	DDE
IS982	-	1000	3-9	AC	Y	1		DDE
IS1380	-	1550-2000	4-5	CC	Y	1		DDE
ISAs1	-	1200-1500	8-10	CAGGG	Y	1		
ISL3	-	1300-2300	8	GG	Y	1		
Tn3	-	¿3000	0	GGGG	Y	¿1		DDE
ISAzo13	-	1250-2200	0-4	Ga/g	Y	1		

largest groups of mobile genetic elements, which are segments of DNA that encode enzymes and other proteins that mediate the movement of DNA within genomes (intracellular mobility) or between bacterial cells (intercellular mobility), see, *e.g.*, [Chandler and Mahillon, 2002, Frost et al., 2005, Craig et al., 2002]. One of the ISfinder capability is to assign IS names and to provide a focal point for a coherent nomenclature. It is also a repository for insertion sequences. Each new IS is indexed together with information such as its DNA sequence and open reading frames or potential coding sequences, the sequence of the ends of the element and target sites, its origin and distribution together with a bibliography when available. Another objective is to continuously monitor ISs to provide updated comprehensive groupings or families.

IScan IScan identifies bacterial ISs and their sequence elements –inverted and target direct repeats– in multiple genomes using multiple flexible search parameters. This tool, developed by Wagner *et al.* [Wagner et al., 2007], has been proposed in 2007. Inverted repeats are found using Smith-Waterman[12] algorithm. IScan has been applied on 438 completely sequenced bacterial genomes by using BLAST with referenced transposases, to determine which transposases are related to insertion sequences. IScan identifies ISs in three major steps.

- **Identification of transposase ORFs:** IScan identifies the ORFs of an IS through a tblastn search (using WUBLAST [Altschul et al., 1997], http://blast.wustl.edu/) which matches the query amino acid sequence(s) to the translation products of the genomic sequence in all six possible reading frames.
- **Identification of (candidate) inverted repeats:** IScan applies a user-specifiable alignment algorithm, such as Smith-Waterman local alignment one [Smith and Waterman, 1981].
- **Identification of (candidate) direct repeats:** the final stage.

ISA *ISA* has been created by Zhou *et al.* in 2008 [Zhou et al., 2008]. This annotation program depends on both NCBI annotations and ISFinder. More precisely, authors manually collected 1,356 IS elements with both sequences and terminal signals from the ISFinder database, which have been used as templates for identification of all IS elements and map construction in the targeted genomes. ISA, which is not publicly available, has finally been used for an analysis of 19 cyanobacterial and 31 archaeal annotated genomes downloaded from NCBI.

[12]The Smith-Waterman algorithm performs local sequence alignment for determining similar regions between two strings (nucleotide or protein sequences). Instead of looking at the total sequence like in Needleman-Wunsch, Smith-Waterman compares segments of all possible lengths and optimizes the similarity measure. The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981.

As an illustrative application, the authors of ISA focused on the recently active IS elements (raIS), which are defined as IS elements with multiple copies of highly similar sequences in the same genome [Ray et al., 2007]. They found that:

1. the activities of IS elements heavily depend on the environments where the host organisms live;
2. the number of recently active IS elements in a genome tends to increase with the genome size;
3. the flanking regions of the recently active IS elements are significantly enriched with genes encoding DNA binding factors, transporters and enzymes; and
4. IS movements show no tendency to disrupt operonic structures.

ISsaga *ISsaga*[13], for its part, has been developed in 2011 by Varani *et al.* [Varani et al., 2011]. ISsaga is an ensemble of web-based methods for high throughput identification and semiautomatic annotation of insertion sequences in prokaryotic genomes. They used eight different bacterial genomes downloaded from NCBI, and produced a web application pipeline that allows semi-automated annotations based on BLAST against the ISFinder database. A modular construction allows the annotation process to be broken down into three interconnected steps: protein (IS-associated ORF identification); nucleotide; and validation steps. The validation step processes the result generated by the previous steps, and exports each predicted IS identified in the nucleotide step to the annotation table. This is an entirely manual procedure, where the annotator must verify each IS prediction result. Obviously, this requires some IS annotation expertise and, as a consequence, ISsaga cannot automatically identify new insertion sequences which are not already present in ISFinder database.

ISsaga can fundamentally contribute to insertion sequence studies in two ways. Firstly, by enriching the ISfinder database by high throughput annotation of completely assembled and scaffold-based genomes. And secondly, by a direct analysis of the metagenomes themselves. Although typical sequence runs in metagenomic analyses are short, enough information can be available to identify a particular IS from fragments at the DNA or protein level. Another advantage provided by a complete genome IS annotation is that it permits a detailed basis on which to compare strains and species [Parkhill et al., 2003].

OASIS A new computational tool for automated annotation of ISs has been releazed in 2012 by Robinson *et al.* [Robinson et al., 2012]. This tool has been called *OASIS*, which stands for “Optimized Annotation System for Insertion Sequences”. They worked with 1,737 bacterial and archaeal genomes downloaded from NCBI.

[13] http://issaga.biotoul.fr/ISsaga/issaga_index.php

OASIS identifies insertion sequences in each genome by finding conserved regions surrounding already annotated transposase genes. It uses a maximum likelihood algorithm to determine the edges of multicopy ISs based on conservation between their surrounding regions. For defining inverted repeats, the same strategy as IScan was used (Smith-Waterman alignment). Authors also used hierarchical agglomeration clustering to identify groups of ISs based on their lengths.

The ISs set is then classified according to the family and group after a BLASTP best hit in ISFinder database with an e-value lower than 10^{-12}. When a cluster cannot match with any entry of the database, the IS set is considered as new, see in Figure 2.14 the full work-flow of OASIS. Thus OASIS has the ability to discover new insertion sequences, that is, which cannot be found in ISFinder.

The outputs of OASIS is constituted by two files for each genome: a fasta file and a gene feature *gff* one. The gff file contains one line for each insertion sequence. An example line looks like:

NC_002516.2 OASIS IS 499832 501193 . + . set_id “1”; family “IS3”; group “IS3”; IRL “ATGGACTCCTCCC”; IRR “ATGGACTCCTCCC”;

Along with the usual gff data (chromosome, source, feature, start, end, score, direction, and strand), OASIS provides attributes that describe the insertion sequence:

- **set_id:** ISs within a genome are divided into sets, each containing nearly identical copies.
- **family:** The identified family of the insertion sequence, based on the ISfinder database of IS elements.
- **group:** The group of the insertion sequence (subclass of ISfinder family).
- **IRL:** The left (defined as upstream of the coding sequence) inverted repeat sequence, if one is found.
- **IRR:** The right (defined as downstream of the coding sequence) inverted repeat sequence, if one is found.

The fasta file contains the nucleotide sequence of each IS and one amino acid sequence for each open reading frame (ORF), see Figure 2.15. Recall that an ORF is the part of a reading frame that has the potential to code for a protein or peptide. An ORF is a continuous stretch of codons beginning with a start codon (usually ATG) and ending with a stop codon (usually TAA, TAG, or TGA), see [Brown, 2010]).

Table 2.9 shows the comparison between OASIS and the three other tools described previously.

In fact, IS detection can be performed on either finished genomes or on draft assembled ones. But, obviously, the number of identical ISs that are in multiple copies cannot be reported correctly in the first kind of genomes. This is why we

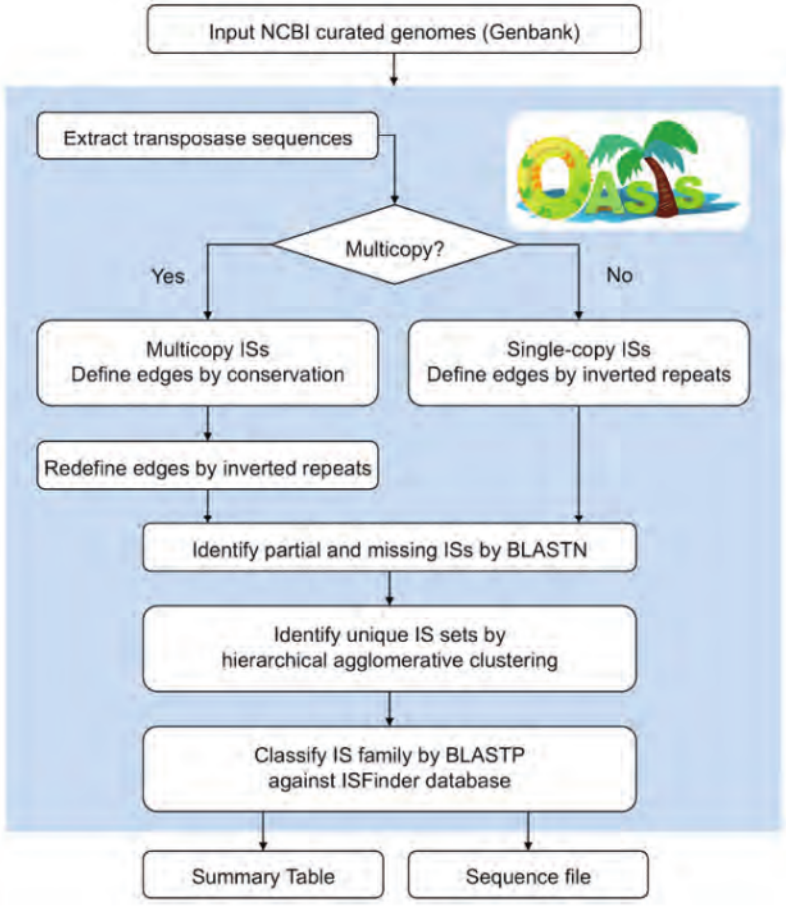


Figure 2.14: **Flowchart portraying the full workflow of OASIS** [Robinson et al., 2012].

have considered that, to detect the exact number of ISs, we have to work with read data.

breseq is a workflow to identify and annotate mutations of bacteria, by comparing read data (fastq) with a closely related reference genome. Its IS detection procedure is described below [Barrick et al., 2014].

- Map read data with reference genome using Bowtie2;
- Identify the junction evidence (JC) from a split-read matching at a location to find in the reference, and match it again in a distant site.
- Predict the mobile element insertion (MOB). This step is performed if we have been able to recover 2 JCs in the previous step.

Breseq has three main limitations, which are listed hereafter: (i) it is unable to find new sequences that do not exist in the reference, (ii) mutations in repeat regions are problematic, and (iii) chromosomal inversions and rearrangements through

110645304.fasta

```
>NC_002516.2_499832_501193|-|1
ATGGACTCCTCCCTCCTACGGCGTCAAGCGCCAGACTAATGGTGCAGCCACAAGTCCCGG
AGGGGGAGTCAGCATGAACCTTAGTCGCATTGGTCTGGATCTGGCAAAACAAGTATTCCA
GGTGCACGGCGTTGATCGTCACGAGCATGTGGTATGTCGTCGCCAACTCAAGCGGGCACA
GGTGCGGGATTTCTTTCGCCAACTGCCGCCGTGCCTGGTGGCGATGGAGGCCTGCGGCAG
TGCGCACTACTGGGCGCGCGAGTTGCGGGAGCTGGGCCACACGGTACGCCTGATCGCACC
GCAGTTCGTAAAACCCTACGTCAAGGGTGACAAGCACGATGCACACGACGCCGAGGCGAT
CTGCGAAGCCGCCAGTCGTCCGAGCATGCGCTATGTCCCGGTGAAGAGCGCAGAGCAGCA
GGCCGTGCAGTCGATGCATCGGGTGCGCAGTCGTCTAGTGCGGGCCCGCACGGCGTTGTG
CAATGAGGTGCGTGGCTTGCTGGGTGAGTTCGGGCTGATCGCGACTCGACGTGGGCGAGC
GGCGACGATGGCCTTGCTCGAAACGGTCATGGCTACCGAGCCGGCGCCCTTGCCGGCCCC
GATGGGCGAGCTGTTGCGTGAATTGAAAGATGAATTGCAGACGCTGGAGGCGCGCATCGC
TCGACTGGAGCGACAGATTCAGGCTCACGTACGTGGCGATGCCCGCATCCAGCGCCTGCT
GGCGGTGGAAGGCATCGGCCCGATCAGCGCCAGCGCGGTGGCGGCATCCGCCGGTGATGC
ACGGCAATTCCGCACGGGCCGCCAGTTTGCGGCCTGGCTGGGCCTGGTGCCACGGCAGCA
CTCCACGGGCGGGCAGCAGCGCTTGGGCAACATCAGCAAGCGCGGCGATACCTACTTGCG
GACCTTGCTCATCCATGGCGCCCGTGCAGTCGTGCGCTGTTGCGCCAACAAGACCGATGC
CCGCAGCCGCTGGCTGCAAGGTCTGCTGCAACGGCGTCCTGCCAATGTCGTCGCCGTTGC
CCTGGCCAACAAGAATGCCCGGATCCTTTGGGCTTTACTCAGCCGGGAGACATGCTACCG
GCCCGGTTGAGCGTTCCTGCCACACCGTAGTTGAAACATCCACCACGATTGCTCAGTGAA
TGACAATGATGACGAACCGGTCGAACCGGCCTGCATGAAACCTGGTTTATACGTGGGCTC
CCTGCTGCAGTGAAGCAAAGCCGTTAGGGCGATCAGGTATGCAGGCGCGCATTTCATCAG
GGCTCGGGAGTTGCAACACCACTCCATGAAGCCGGATATACGGATGCAGTCGTACACAGG
TTTGAAATCAAGACAAACACTGGCAAACCGGGAGGAGTCCAT
>NC_002516.2_499832_501193|-|1|ORF
MNLSRIGLDLAKQVFQVHGVDRHEHVVCRRQLKRAQVRDFFRQLPPCLVAMEACGSAHYW
ARELRELGHTVRLIAPQFVKPYVKGDKHDAHDAEAICEAASRPSMRYVPVKSAEQQAVQS
MHRVRSRLVRARTALCNEVRGLLGEFGLIATRRGRAATMALLETVMATEPAPLPAPMGEL
LRELKDELQTLEARIARLERQIQAHVRGDARIQRLLAVEGIGPISASAVAASAGDARQFR
TGRQFAAWLGLVPRQHSTGGQQRLGNISKRGDTYLRTLLIHGARAVVRCCANKTDARSRW
LQGLLQRRPANVVAVALANKNARILWALLSRETCYRPG
```

Figure 2.15: **Example of a fasta file outputted by OASIS.** It contains the IS nucleotides with start and end positions, and the amino acid sequence.

repeat sequences cannot be considered. Let us finally remark that breseq supports Unix/Linux, Mac OS, and Windows, as its programming languages are C++ and R.

ISMapper is an interesting tool for IS detection using short read data in fastq format and a reference genome (fasta or genbank). It also requires IS sequences in fasta format as a library for IS identification.

Firstly, it performs a mapping of reads to its IS library by BsingWA u. A SAM file that includes left-right flanking reads and soft clipped reads is then outputted at the end of this step. After that, it maps flanking reads and soft clipped reads to the reference genome, while removing low coverage regions by considering depth of mapped output. Next, sequences adjacent to the IS are identified by comparing

regions of the left and right end blocks to the reference genome. In the case of known IS positions, the compared regions will be intersected. Conversely, in the case of a novel IS position, the compared regions must be close and contain the overlapping sequences from left and right ends. Finally, the program returns 2 output files: (i) a genbank file of the reference sequence, and (ii) a tabular file indicating the position, orientation of each insertion site, and some information about the genes flanking the insertion site [Hawkey et al., 2015].

Let us finally remark that ISMapper is compatible with Unix/Linux, Mac OS, and Windows. Its programming language is Python. Before being able to use it, the user has to install the following software/libraries: BWA, SAMtools, Bedtools, BLAST, Samblaster, and Biopython.

Other related works Insertion sequences have been regarded too in the following research works:

- The study on the plant-pathogenic prokaryote *Xanthomonas oryzae pv. oryzae (Xoo)*, which causes bacterial blight (one of the most important diseases of rice) was published in 2005 by Ochiai *et al.* [Ochiai et al., 2005]. They used GeneHacker [Yada and Hirosawa, 1996], GenomeGambler version 1.51, and Glimmer program [Delcher et al., 1999a] for coding sequence prediction. Insertion sequences were finally classified by a BLAST analysis using ISFinder database evoked previously.

- Touchon *et al.*, for their parts, have analyzed 262 different bacterial and archaeal genomes downloaded from GenBank NCBI[14] in 2007 [Touchon and Rocha, 2007]. A coding sequence has then been considered as an IS element if its BLASTP best hit in ISFinder database has an e-value lower than 10^{-10}.

- In 2010, Plague *et al.* analyzed the neighboring gene orientations (NGOs) of all ISs in 326 fully sequenced bacterial chromosomes. They obtained primary annotations from the Comprehensive Microbial Resource database (release 1.0-20.0) at the Institute for Genomic Research[15]. Their approach for extracting IS elements from these genomes was to consider that a coding sequence with a best BLASTX hit e-value lower than 10^{-10} is an insertion sequence [Plague, 2010].

- Finally, in 2014, the analysis of the NGOs for all IS elements within 155 fully sequenced Archaea genomes was presented by Florek *et al.* [Florek et al., 2014]. To do so, they have launched a BLASTP in the ISFinder, with an e-value lesser than or equal to 10^{-10}, for all protein coding sequences downloaded from NCBI that are related to ISs.

[14] ftp://ftp.ncbi.nih.gov/genomes/
[15] http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi

Two major concerns with the tools detailed in this chapter can be emphasized. Firstly, most of them cannot detect new insertion sequences. Secondly, most of these tools are based on NCBI annotations, which are of very relative and variable qualities – except ISsaga, which could work with other annotation tools (but it depends only on transposase ORFs that have been already defined in ISFinder). Our objective in a next chapter will be to propose a pipeline that solves these two issues, being able to deal with unannotated genomes and to detect unknown ISs.

2.2/ PSEUDOMONAS AERUGINOSA

2.2.1/ GENERAL PRESENTATION AND ECOLOGICAL NICHE

Pseudomonas aeruginosa is one of the most frequent and severe causes of human opportunistic and acute infections like ventilator associated pneumonia and burn wound infections [Vincent, 2003, Gellatly and Hancock, 2013]. It is known to be a severe driver of chronic respiratory infections in cystic fibrosis patients [Oliver et al., 2008], and it is also related to various other infections, like otitis, bacteriemia, enterocolitis, meningitis, or folliculitis, to name a few.

Although not considered part of the resident human microbiota, gastrointestinal, upper respiratory tract or cutaneous colonization may occur, particularly among hospitalized patients [Oliver et al., 2015]. Indeed, due to its remarkable metabolic plasticity and versatility, this ubiquitous microorganism is able of colonizing a wide range of ecological niches, including aquatic and soil habitats, animals, and plants, as reported in [Silby et al., 2011].

2.2.2/ PATHOGENICITY

P. aeruginosa is considered as one of the major nosocomial pathogen due to its remarkable ability to face antimicrobial treatments, and its adaptability originated from its large number of regulatory genes and of virulence determinants [Gellatly and Hancock, 2013, Breidenstein et al., 2011]. Relevant treatments are dramatically reduced by the increasing prevalence of chronic and nosocomial infections caused by multidrug-resistant (MDR) or extensively drug-resistant (XDR) strains. As a consequence, *P. aeruginosa* is increasingly associated with significant morbidity and mortality [Mesaros et al., 2007, Livermore, 2009].

At the origin of this increasing threat, we find the noticeable ability of *P. aeruginosa* to develop resistance to almost all antibiotics. This is done thanks to the selection of mutations in chromosomal genes, to the increasing prevalence of transferable resistance determinants, like the ones encoding class β carbapenemases (metallo-β-lactamases) or extended-spectrum β-lactamases (ESBLs), see, *e.g.*, [Livermore, 2002, Lister et al., 2009] and to the enhanced spontaneous

Table 2.9: **Comparison of IS detection tools.**

Tool	New IS	Annotated	Transposase gene	Inverted Repeat	Mechanism
Iscan (2007)	Yes	yes	Local Database (Reference transposases)	smith waterman	automated
ISA (2008)	No	yes	collected 1,356 IS from ISFinder database	manually	not available
Issaga (2011)	No	no	ISFinder database	manually	semi-automated
OASIS (2012)	yes	yes	NCBI database	smith waterman	automated

mutation rates in mutator isolates, particularly found in chronically-infected patients [Maciá et al., 2005].

Relations between pathogenicity, epidemicity, and antibiotic resistance are usually considered when trying to understand the worldwide spread of any pathogen, which are related to [Oliver et al., 2015]:

- regulatory networks that establish interconnections between resistance and virulence [Balasubramanian et al., 2013, Gooderham and Hancock, 2009];
- the fitness cost of antibiotic resistance mechanisms [Andersson and Hughes, 2010, Beceiro et al., 2013];
- antibiotic resistance determinants and clonal success linked by so-called genetic capitalism, and achieved with natural genetic engineering [Baquero, 2004, Martínez and Baquero, 2002].

2.2.3/ POPULATION STRUCTURE

First studies in population structure of *P. aeruginosa* indicated a panmictic or fully sexual structure [Picard et al., 1994, Denamur et al., 1993], while further investigations emphasize a population structure in network, with a lot of recombinations between isolates [Kiewitz and Tümmler, 2000]. These early studies have finally led to a consensus, claiming the nonclonal epidemic population structure of *P. aeruginosa* [Maatallah et al., 2011, Curran et al., 2004, Kidd et al., 2012]. In other words, we have a reduced number of widespread clones that are extracted from a larger number of rare, unrelated genotypes that are highly recombining, as summarized in [Pirnay et al., 2002, Oliver et al., 2015]. Further structure analyzes, being possible due to an increasing access to whole genome sequence data [Dettman et al., 2013, Cramer et al., 2011, Jeukens et al., 2014], have emphasized a conserved core (see Section 2.2.5.1), while the accessory genome of *P. aeruginosa* is composed by extrachromosomal elements (*e.g.*, plasmids and blocks of DNA inserted in various loci), as described in [Klockgether et al., 2011]. This latter has probably been acquired during horizontal gene transfer from different sources (even other species, see [Oliver et al., 2015]).

The most popular approach to study *P. aeruginosa* populations is the multilocus sequence typing MLST scheme (see Figure 2.16) that has been introduced in [Curran et al., 2004]. It is based on the sequencing of 7 well-defined genes evenly distributed in the core genome of this bacteria, namely: *acs*, *aro*, *gua*, , *nuo*, *pps*, and *trp*. Being housekeeping, these genes are not subjected to positive selection. The MLST database[16] has referenced over 5,900 isolates, presenting more than 2,600 sequence types (STs). They are such that most of them are represented by single isolates, while about twenty STs are represented by more than 10 isolates from at least three different countries (successful clones). At this stage, we can remark that:

[16] Available at http://pubmlst.org/paeruginosa, Last updated: 2017-08-30

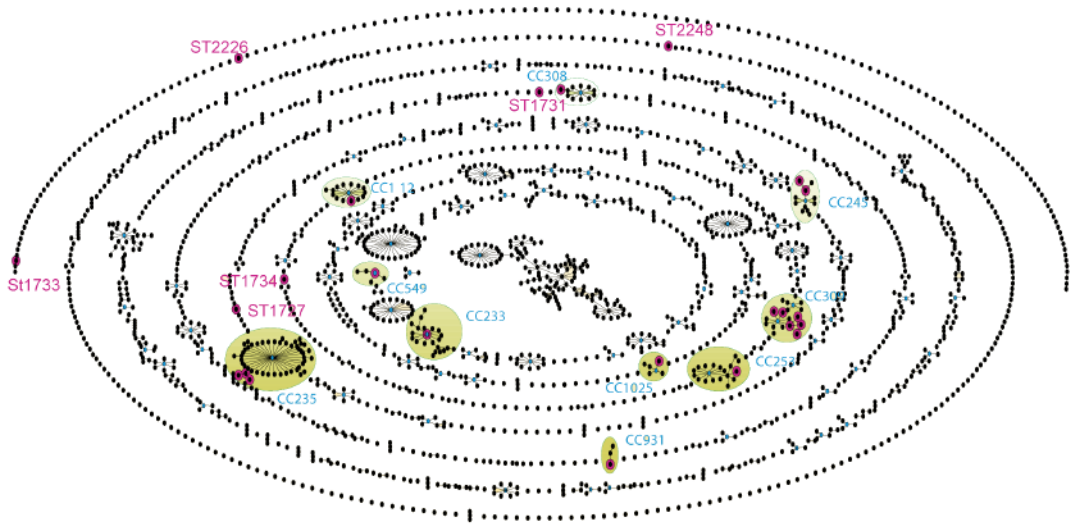


Figure 2.16: **Population snapshot in MLST scheme of *Pseudomonas aeruginosa* based on 2,266 sequence types (STs) from PubMLST database in April 2016** [Aguilar-Rodea et al., 2017]. Black dots represent STs while lines connect single-locus variants (SLVs). Blue points represent founder STs. The strict clonal complexes (CC) are highlighted in green which are the set of STs sharing 6 of 7 alleles.

- The MLST scheme only provides information on the core genome.
- The over-representation of particular types of isolates such as those showing antibiotic resistance, those recovered from cystic fibrosis patients or those obtained from certain geographic areas, is highly significant in terms of epidemiology [Oliver et al., 2015].
- Mutation of *mutL* is a frequent cause of the mutator phenotypes positively selected in chronic infections, as this gene encodes a component of the DNA mismatch repair system [Oliver et al., 2002,García-Castillo et al., 2012, Mena et al., 2008].

Among others frequent sequence types are found the widespread clone C (ST17) and PA14 (ST253) clones, high-risk clones associated with MDR/XDR nosocomial infections (like ST111, ST175, or ST235 that will be further studied in this manuscript) or CF epidemic clones such as ST146 (LES). Even if up to 838 STs are classified as singletons [Oliver et al., 2015], some of the aforementioned frequent clones are at the basis of clonal groups ranging from 2 to more than one hundred of STs:

- The largest clonal complex includes ST111, ST146 (LES), and ST17 (that is, clone C), which is probably the founder of this very large group. Note that, within this group, the international high-risk clone ST111 is the founder of a very large subgroup.

- International high-risk clone ST235 is the founder of the second largest clonal complex that has 43 STs.
- Finally, the international high-risk clone ST175 is the founder clone of a smaller clonal complex, which contains 12 STs.

For further details, please read [Oliver et al., 2015].

2.2.4/ RESISTANCE

The WHO defines *P. aeruginosa* resistant to carbapenems as major superbugs [Lawe-Davies and Bennet, 2017]. It can become readily resistant to multiple classes of available antimicrobial drugs, like aminoglycosides, fluoroquinolones, and β-lactamases. This is due to its capability to naturally resist to a wide range of antibiotics (called 'intrinsic resistance') and to acquire resistance determinants to antipseudomonal agents (called 'acquired resistance').

Intrinsic resistance mechanisms come from existing genes encoding inherent properties of cell structures and composition that prevent from antibiotics and toxic molecules. Figure 2.17A displays the extracellular polymeric substances (EPS) in which some non-specific porin proteins are replaced by specific ones, in order to limit the antibiotic uptaking. This mechanism includes AmpC production and overexpression of multidrug efflux pumps: MexAB-OprM and MexXY/OprM(OprA) to pump antibiotics out from cell [Breidenstein et al., 2011, Moradali et al., 2017].

Acquired resistance mechanisms are related to mutations on intrinsic genes or horizontal acquisition of genetic elements from plasmids, transposons, integrons, prophages, and resistance islands.

P. aeruginosa can readily acquire foreign genes, like those encoding the extended-spectrum β-lactamases (ESBLs) and carbapenemases as in Figure 2.17B: carbapenem-resistant genes have been found in all continents.

The mutations of chromosomal genes is the most frequent mechanism of resistance to antibiotic. Mutations can occur in (i) target genes as quinolone resistance determining region (QRDR) of DNA gyrase (*gyr A* and *gyr B*) and topoisomerase IV (*par C* and *par E*), (ii) *oprD* gene, (iii) transcriptional repressors leading to upregulation of resistance genes and efflux pumps, and (iv) genes related LPS (lipopolysaccharide) components [Lister et al., 2009, Moradali et al., 2017] (see Figure 2.17B). For instance, fluoroquinolone resistance can occur from mutation within the targets of fluoroquinolone (*gyr A*, *gyr B*, *par C*, and *par E*) or overexpression of efflux pumps. In addition, a mutation in *OprD* gene (as inactive) can lead to a loss or reduction of carbapenem transfer that is specific to OprD.

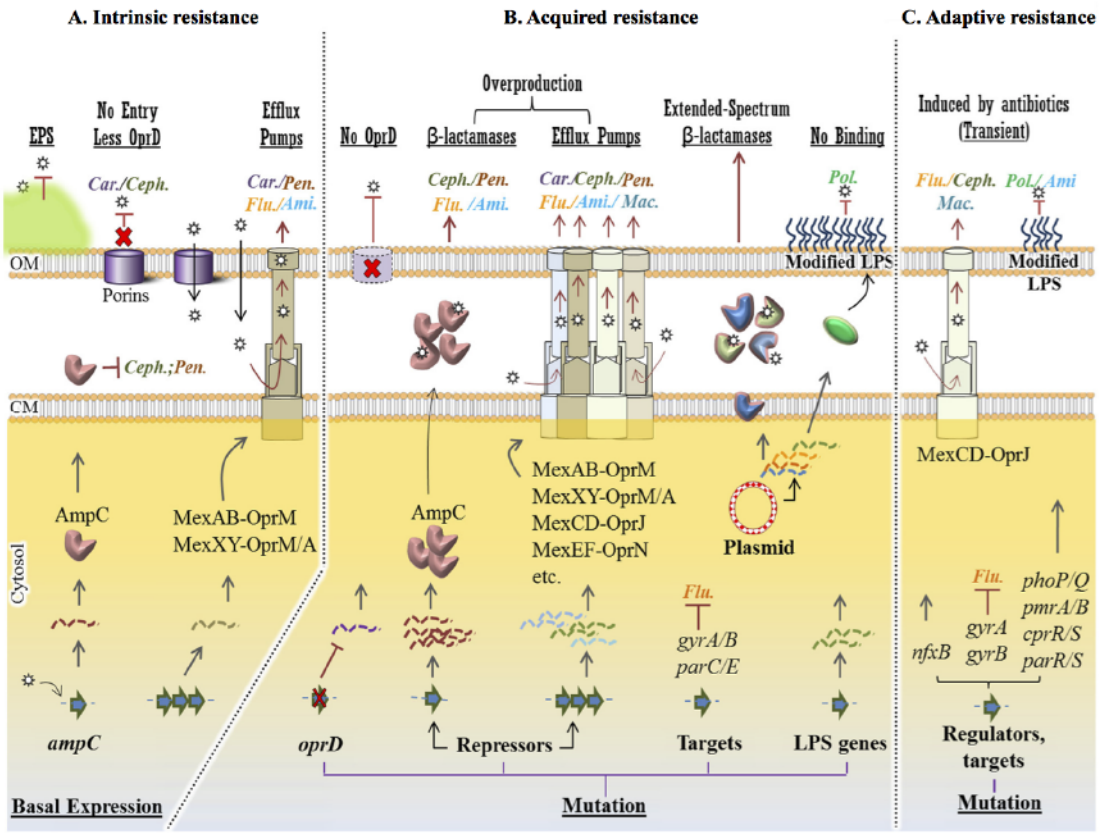


Figure 2.17: **Antibiotics resistance mechanisms of *P. aeruginosa*** [Moradali et al., 2017]. Stars represent antibiotics and dashed/wavy lines represent transcriptional levels including their colors depending on antibiotics. The abbreviation of antibiotics are Car. (Carbapenems), Ceph. (Cephalosporins), Pen. (Penicillins), Ami. (Aminoglycosides), Flu. (Fluoroquinolones), Mac. (Macrolides), and Pol. (Polymyxins). Others abbreviation are EPS (extracellular polymeric substances), LPS (lipopolysaccharide, CM (cytoplasmic membrane) and OM (outer membrane).

As a consequence, mechanisms of acquired genes can lead to multidrug and pandrug resistance which is a serious threat in clinical treatments [Potron et al., 2015].

Adaptive resistance mechanisms are induced by either antibiotics or other environmental events such as mutations in target genes or regulators, see Figure 2.17C. Since the two-component system PhoP-PhoQ, in which PhoQ acts as a sensor kinase that phosphorylates and activates the transcription factor PhoP [Breidenstein et al., 2011], is mutated, this leads to a mutation in the LPS structure and thus finally to a resistance against polymyxin.

2.2.5/ GENOME EVOLUTION/ADAPTATION

2.2.5.1/ CORE AND PAN GENOMES

The core and pan genomes have been investigated in [Valot et al., 2015] using a clustering approach. Using 17 distinct clones, authors of this article found that the average *P. aeruginosa* genome contained 5,972 genes, while the pan-genome is of 9,344 genes and the core genome has 5,233 genes. Using a smaller set of genomes, other authors found comparable results [Mathee et al., 2008, Klockgether et al., 2011, Roy et al., 2010]. The core genome is thus highly conserved and represents ca. 88% of the average genome, and the full sequencing of a *P. aeruginosa* strain allows the observation of ca. two-thirds of the pan-genome. In particular, most fundamental functions can be studied with a model strain and extrapolated to the species.

For the sake of comparison and following [Valot et al., 2015], let us emphasize that *P. aeruginosa* has a larger genome than those of the two other major nosocomial pathogens, namely *Escherichia coli* and *Staphylococcus aureus*: the former has an average size of 4,721 and the latter of 3,118 genes, see [Touchon et al., 2009, Boissy et al., 2011]. The ubiquity of *P. aeruginosa* relies on its metabolic versatility and on the large range of its hosts. According to [Valot et al., 2015], these properties are a consequence of the large genome of the species, while the small size of the accessory genome of *P. aeruginosa* reflects the absence of strain clustering during evolution and the minimal adaptation to environmental niches.

2.2.5.2/ VIRULENCE FACTORS

Multiple virulence factors have been reported concerning the pathogenesis of *P. aeruginosa* [Gellatly and Hancock, 2013]. It can be remarked that many of these factors are located in the accessory genome as part of pathogenicity (PAPI) or genomic (PAGI) islands, see, *e.g.*, [Battle et al., 2008]. As signaled in [Oliver et al., 2015], a habitat-independent interclonal gradient of virulence has been revealed through the pangenome analysis of isolates coming from various infections [Hilker et al., 2015]. Furthermore, it has been reported in the literature [Sawa et al., 2014, Linares et al., 2005, Martínez-Ramos et al., 2014, Skurnik et al., 2013] that antibiotic resistance mechanisms can:

1. have a direct effect (positive or negative) on virulence,
2. either determine or not a biological cost compromising bacterial virulence, and
3. be statistically related to some virulence traits.

The type III secretion system (TTSS) is one of the most relevant *P. aeruginosa* virulence factors. As recalled in [Oliver et al., 2015], this secretion system injects

potent cytotoxins, including ExoS, ExoT, ExoU (determining the greatest impact in bacterial virulence), or ExoY, into eukaryotic cells [Hauser, 2009]. The distribution of genes encoding such cytotoxins is not uniform among the strains, some of these genes being mutually exclusive. Note that each of the aforementioned enzymes determines a distinct host tissue injury. The TTSS genotype is thus linked to clonal lineages, and for this reason, it may play a major role in their intrinsic virulence levels.

As a consequence, TTSS genotype is considered as a major factor in the virulence and clinical outcomes associated to high-risk clones [Peña et al., 2012, Edelstein et al., 2013]. However, not all MDR isolates become widespread high-risk clones. So, if resistance plays a major role in the success of high-risk clones, additional factors of this success should exist, and recent studies emphasize that virulence factors should probably be part of the explanation [Kos et al., 2015, Turton et al., 2015].

2.2.5.3/ Horizontal gene acquisition

Horizontal gene transfer (HGT) is an important mechanism, already evoked in this manuscript, that acts on evolution of bacterial genomes by DNA transferring between the non parent-child organisms. It brings about broadly distribution of antibiotic resistance genes via 4 mechanisms, shown in Figure 2.18, which are listed below [von Wintersdorff et al., 2016]:

- **Conjugation** is the DNA transferring from cell to cell through a link of pilus or adhesin by plasmid, or integrative conjugative element (ICE) working as transporter.
- **Transformation** is the DNA uptake from environment into cell.
- **Transduction** is the DNA transferring from infected cell to another cell by bacteriophage (bacterial viruses).
- **Gene transfer agents (GTAs)** are DNA transferred by bacteriophage-like particles that are encoded from host cell.

The main mechanisms of HGT are frequently mentioned as conjugation, transformation, and transduction [Summers et al., 2005, Bennett, 2008, Kung et al., 2010, Daubin and Szöllősi, 2016]. Even if all described genetic elements have unique characteristics, they are alike in cell transfer mechanisms, being considered as mobile genetic elements (MGEs). Generally speaking, MGEs can be classified in 2 groups according to their transfer ability: the intercellular MGEs on the one hand, and the intracellular MGEs on the other hand. Intercellular MGEs like plasmids and bacteriophages are able to transfer themselves between bacterial cells, while intracellular MGEs (*e.g.*, transposons, gene cassettes, and insertion sequences) require an intercellular integration to achieve such a transfer. We will describe in what follows the main MGEs of *P. aeruginosa*.

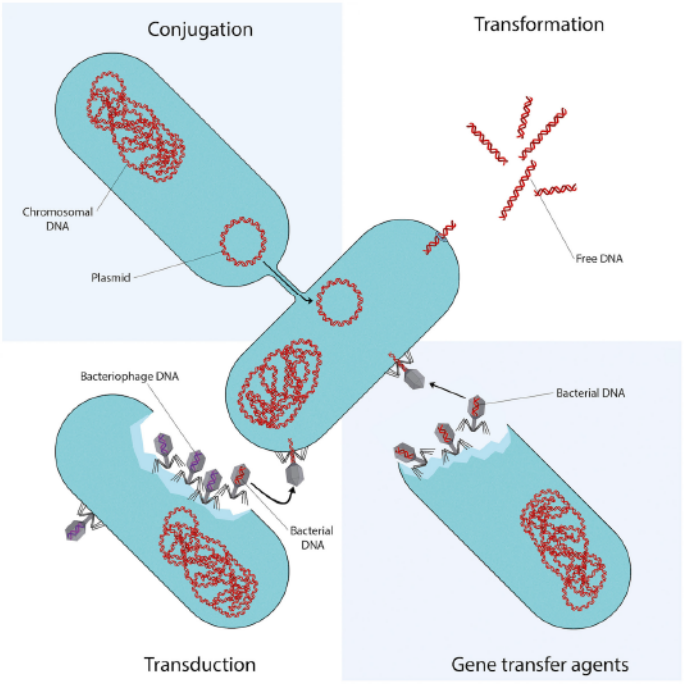


Figure 2.18: **Mechanisms of horizontal gene transfer in bacteria** [von Wintersdorff et al., 2016].

Plasmids Plasmids are small and circular, with a double-stranded DNA. They have some capability in stabilization, copy number controlling, and self-replication. Some of them are reported as resistant plasmids, since they carried antibiotic resistance genes such as plasmid RP1 [Bennett, 2008], plasmid pCOL-1, and plasmid pPA-2. Even though the two last aforementioned plasmids are different in size (pCOL-1: 31,529 bp and pPA-2: 7,995 bp), they are similar in carbapenem resistance, as they both contain *bla*KPC-2 genes [Naas et al., 2013].

Integrative and conjugative elements (ICEs) ICEs are self-transmissible genetic elements that encode the machinery for conjugation and integration into replicons of the host chromosome (see Figure 2.19). ICEs are similar to plasmids: they are mobile circular elements before their self or conjugative transfer. They share similarities too with bacteriophages, transposons, and insertion sequences, as they integrate themselves into host chromosome. The common size of *P. aeruginosa* ICEs ranges from 81 kb to 108 kb, and their 72 ORFs share more than 75% of identity [Kung et al., 2010]. They confer various phenotypes, *e.g.*, antibiotic resistance from PAGI-3, virulence factors and regulation of biofilm formation from PAPI-1 [Wozniak and Waldor, 2010], and so on (see Table 2.10).

Bacteriophages Bacteriophages are bacterial viruses that are composed by protein-coated DNA or RNA. They carry only the genetic elements needed to infect the host cell and the replication machinery. Since phages dock on host

Table 2.10: **Characteristics of ICEs in *P. aeruginosa*** [Kung et al., 2010].

ICE	Size (kb)	tRNA integration site	Features	NCBI accession no.
PAGI-2	105	$tRNA^{Gly}$	Cargo genes thought to function in complexing with and transport of heavy metals	AF440523
PAGI-3	103	$tRNA^{Gly}$	Cargo genes thought to confer metabolic, transport, and resistance capacities	AF440524
PAGI-4	23	$tRNA^{Lys}$	Contains genes with putative metabolic functions	AY258138
PAGI-5	99	$tRNA^{Lys}$	Contributes to virulence in mouse model of acute pneumonia	EF611301
PAGI-8	18	$tRNA^{Phe}$	Predicted to encode an ATPase, a Zn-dependent transcriptional regulator, and a DotA/TraY-like protein	EF611304
LESGI-1	46	$tRNA^{Pro}$	Contains several ORFs similar to those for predicted proteins in nonpseudomonads	FM209186
LESGI-3	111	$tRNA^{Gly}$	Cargo genes thought to confer transport capacities	FM209186
LESGI-5	29		Contributes to virulence in a rat lung chronic infection model; high prevalence of carriage in CF patients	FM209186
PAPI-1	108	$tRNA^{Lys}$	Cargo includes numerous genes shown to affect virulence in infection models; self-transmissible	AY273869
PAPI-2	11	$tRNA^{Lys}$	Contains gene encoding the type III secreted virulence factor ExoU	AY273870
Dit island	112	$tRNA^{Gly}$	Putative diterpenoid metabolism island	NZ-AAKW00000000
pKLC102	104	$tRNA^{Lys}$	Can exist episomally; ICE progenitor	AY257538
pKLK106	106	$tRNA^{Lys}$	Can exist episomally; ICE progenitor	AF285417-AF285424
clc element	105	$tRNA^{Gly}$	Found in *P. knackmussii* strain B13; contains genes for chlorocatechol degradation; self-transmissible; ICE progenitor	AJ617740

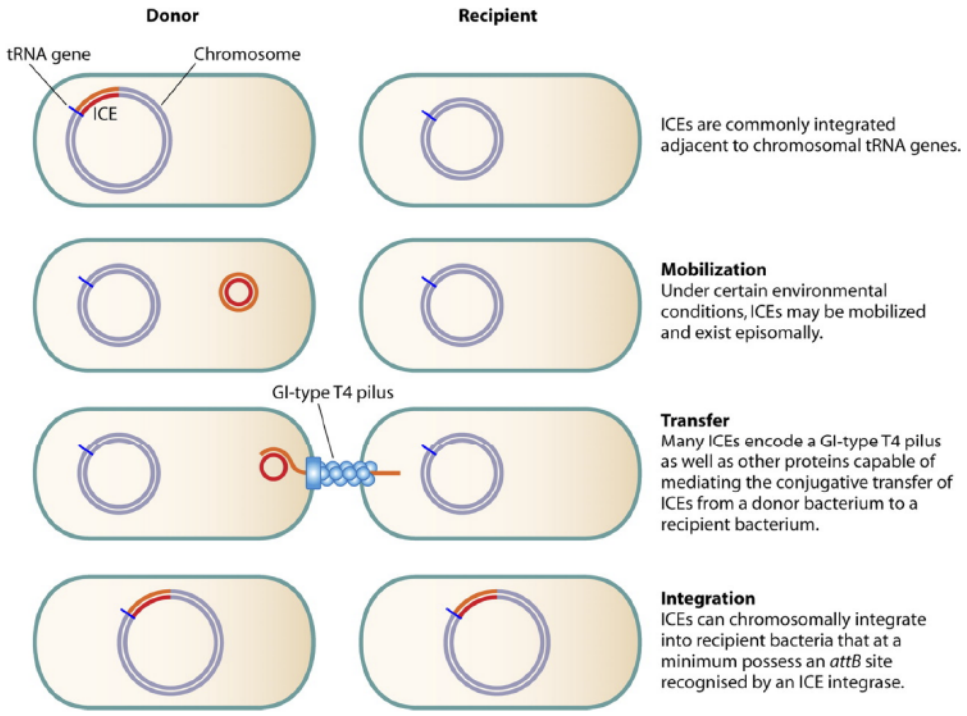


Figure 2.19: **Transfer mechanisms of ICE** [Kung et al., 2010].

cells, they transfer genetic elements to the host to complete infection, by using one of the three modes below:

- lytic phase: phages lyse their bacterial host cells;
- replication: phages, called prophages, are integrated into host chromosome, which leads to mutations;
- phages carrier: phages are not integrated with host chromosome, or phages are stable as episome. Thus, episome phages can asymmetrically segregate cells in place of replication [Blasdel et al., 2017].

P. aeruginosa phages act on pathogenicity, like for instance: (i) phage D3112 decreases biofilm development while lipopolysaccharide (LPS)-specific phage E79 increases it, (ii) ϕCTX contains a gene encoding a pore-forming toxin which increase virulence, (iii) phage OMKO1 causes a decrease of antibiotic resistance, due to the fact that this phage uses membrane porin M for binding site in lytic process – but, normally, porin M function as pump antibiotics outs of host cells [Blasdel et al., 2017]. Other examples are provided in Table 2.11.

Insertion sequences (ISs) As already mentioned in this chapter, ISs are small DNA fragments that encode enzymes for their movements. They must be integrated into plasmids or bacteriophages for cell-cell transferring. They consist

Table 2.11: **Phages and prophage-like elements in *P. aeruginosa*** [Kung et al., 2010].

Element	Size (kb)	Features
LES prophage 2	42	Contains regions of homology to *Siphoviridae* family phage F10; contributes to virulence in a rat lung chronic infection model
LES prophage 3	43	Contains regions of homology to phage F10, *P. aeruginosa* strain 2192, and LES prophage 5; contributes to virulence in a rat lung chronic infection model
LES prophage 4	37	Similar in sequence to transposable phage D3122
LES prophage 5	50	Shares regions of similarity with phage D3; contributes to virulence in a rat lung chronic infection model
LES prophage 6	8	Similar to filamentous phage Pf1
ϕCTX	36	Double-stranded DNA *Myoviridae* family phage encoding a pore-forming cytotoxin that contributes to the virulence of *P. aeruginosa* strains harboring it
PAGI-6	44	Exhibits a high level of sequence identity to ϕCTX; however, lacks genes encoding ϕCTX integrase and cytotoxin
D3	56	Double-stranded DNA *Siphoviridae* family phage containing a "seroconverting operon" that changes lysogenized *P. aeruginosa* strains from serotype O5 to O16
Pf1	7	Single-stranded DNA filamentous phage highly upregulated during *P. aeruginosa* biofilm development
Pf4	12	Single-stranded DNA filamentous phage affecting *P. aeruginosa* biofilm phenotypic variation and differentiation as well as virulence
PT-6	Unknown	Double-stranded DNA *Podoviridae* family bacteriophage that produces an alginase
R-type and F-type pyocins	12-15	Defective prophages related to phage P2 and phage lambda; encode phage tail particles with antibacterial activity; affect susceptibility to fluoroquinolone antibiotics
F116	65	Double-stranded DNA *Podoviridae* family generalized transducing phage; digests alginate and encodes several putative proteins with amino acid sequence similarity to proteins from fluorescent *Pseudomonas* species
D3112	38	Double-stranded DNA *Siphoviridae* family generalized transducing phage with transposase-mediated integration; representative of one of two groups of *P. aeruginosa* transposable phages (D3112-like and B3-like); like other tailed *P. aeruginosa* phages, it exemplifies a mosaic genetic structure; predicted to encode several proteins with sequence similarity to other phage proteins as well as proteins from the bacterial plant pathogen *Xanthomonas fastidiosa*

of terminal inverted repeats (IRs) for transposases (Tpases) binding sites and also splitting points leading to transposition [Mahillon and Chandler, 1998]. Most of ISs are flanked by short sequences, called direct target repeats (DRs) [Siguier et al., 2015]. Since ISs are various in mechanisms and features, they have been classified as families by (i) orf organization; (ii) transposition chemistry; (iii) features of their ending terminal inverted repeats (IRs); and (iv) their target sequences [Mahillon and Chandler, 1998]. ISs affect not only the gene expression but also have impact on neighbouring gene expression. The following ISs have an influence on antibiotic resistance:

- IS*1999* inserts at upstream of $bla_{\text{VEB-1}}$ gene that increase expression of the extended-spectrum-lactamase (ESBL) or resistance to ceftazidime [Kung et al., 2010].
- IS*Pa12* inserts at upstream of $bla_{\text{PER-1}}$ gene which increase expression of ESBL or resistance to penicillins, cefotaxime, ceftibuten, ceftazidime and aztreonam [Vandecraen et al., 2017].
- IS*6100* is associated with ESBL expression [Kung et al., 2010].
- other ISs insert as gene inactivation, they are in Table 2.12.

Table 2.12: **IS-mediated gene inactivation affecting resistance of *P. aeruginosa*** [Vandecraen et al., 2017].

Antibiotic resistance	Gene	IS element	IS family
β-lactam^{R}	*ampD*	IS*1669*	IS*5*
	mexR	IS*21*	IS*2*
Carbapenem^{R}	*oprD*	IS*Pa133*	IS*13*
		IS*Pa1328*	IS*256*
		IS*Pa46*	IS*256*
		IS*Pa1635*	IS*4*
		IS*Pa45*	IS*4*
		IS*Pa26*	IS*5*
		IS*Pa8*	IS*5*
		IS*Pre2-like*	IS*5*

2.2.5.4/ CRISPR-CAS SYSTEMS

Clustered regularly interspaced short palindromic repeats (**CRISPR**)–CRISPR-associated genes (**Cas**) **systems** are adaptive immune systems of bacteria for prevention from phage/plasmid infection.The systems are composed of CRISPR locus and Cas genes. The Figure 2.20 displays CRISPR–Cas systems work as defence in 3 phases:

1. **Adaptation**: CRISPR locus, which are spacers or non-repetitive sequences from acquired MGEs (colored regions in Figure 2.20) partition on repetitive sequences or repeats (white regions in Figure 2.20), are integrated with the new spacer.

2. **Biogenesis of crRNAs**: CRISPR loci are transcribed to pre-CRISPR RNA (pre-crRNA), and cleavage repeated into short elements called crRNAs. Next, crRNAs and Cas proteins are formed as CRISPR-Cas complex for phages/plasmids inspector.

3. **Interference**: CRISPR-Cas complex recognizes phages/plasmids DNA thanks to crRNAs.

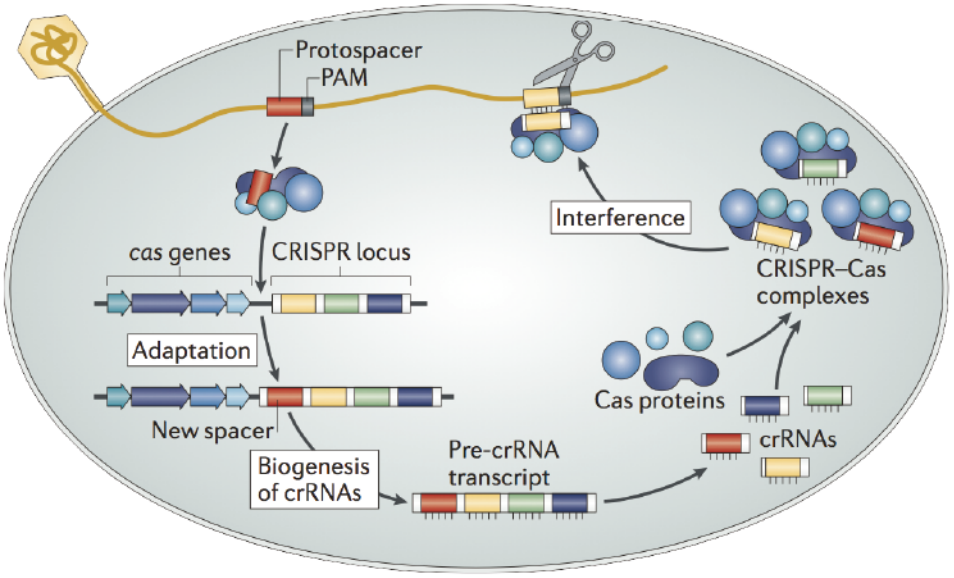


Figure 2.20: **Mechanism of CRISPR–Cas systems** [Samson et al., 2013].

Generally speaking, CRISPR–Cas systems are classified into 3 types, usually denoted by I, II, and III. In the particular case of *P. aeruginosa*, only type I has been reported but with 3 subtypes, namely: I-C, I-E, and I-F [van Belkum et al., 2015]. Interestingly, type I-F CRISPR–Cas systems in *P. aeruginosa* have been signaled to be involved in biofilm development. More precisely, systems of this type interact with a specific protospacer sequence in phage/plasmid DMS3 to inhibit biofilm formation [Cady et al., 2012]. This system participates to the bacteriophage infection defense, since it places a chromosomally integrated element without chromosomal degradation [Samson et al., 2013]. Finally, CRISPR-Cas systems play an important role in controlling horizontal gene transfer and potentially impact on the acquisition of antibiotic resistance genes.

II

CONTRIBUTION

3

Analyzes of global clone ST235

Despite the non-clonal epidemic population structure of *Pseudomonas aeruginosa*, several multi-locus sequence types are distributed worldwide and are frequently associated with epidemics where multidrug resistance confounds treatment. ST235 is the most prevalent of these so-called 'international' or 'high-risk' clones. The evolution of ST235 and the molecular basis for its success are poorly understood. Here we examine the genomes of 79 *P. aeruginosa* ST235 isolates collected worldwide over a 27-year period. Bayesian phylogenetic reconstruction suggests that the ST235 sublineage emerged in Europe in approximately 1984, coinciding with the use of fluoroquinolones as antipseudomonal treatments. The ST235 sublineage seemingly spread from Europe via two independent clones. ST235 isolates then appeared to acquire resistance determinants to aminoglycosides, β-lactams, and carbapenems locally. Additionally, we found that all the ST235 genomes contained the ExoU-encoded exotoxin and identified 22 ST235-specific genes clustering in blocks and implicated in transmembrane efflux, DNA processing and bacterial transformation. These unique genes may have contributed to the poor outcome associated with *P. aeruginosa* ST235 infections and increased the ability of this high-risk clone to acquire foreign resistance elements.

3.1/ Introduction

Pseudomonas aeruginosa is a major opportunistic pathogen responsible for nosocomial infections in humans and for morbidity in individuals afflicted with cystic fibrosis [Gellatly and Hancock, 2013, Lyczak et al., 2002]. This ubiquitous Gram-negative bacilli has a non-clonal epidemic population structure with however several sequence types (ST111, ST175, ST235, ST244, and ST395) distributed worldwide and frequently associated with outbreaks. ST235 is the most prevalent of these so-called 'international', 'high-risk', or 'widespread' clones associated with poor clinical outcomes in part due to multi- and high-level antibiotic resistance [Curran et al., 2004, Maatallah et al., 2011, Pirnay et al., 2002, Woodford et al., 2011, Kos et al., 2015, Oliver et al., 2015]. Treatment of *P. aeruginosa* infections relies on three major antibiotic families: the β-lactams, aminoglycosides, and fluoroquinolones. High levels of resistance to these compounds can

be readily rendered by chromosomal changes. Furthermore, acquisition of resistance genes borne by specific genomic islands and associated transposons is particularly common among *P. aeruginosa* clinical isolates, with nearly 100 different horizontally-acquired resistance elements currently reported in ST235 isolates [Livermore, 2002,Zeng and Jin, 2003,Potron et al., 2015,Oliver et al., 2015]. The expression of these acquired genes, together with the intrinsic resistance mechanisms, considerably reduces the therapeutic options for the treatment of infections caused by ST235 *P. aeruginosa*.

Additionally, *P. aeruginosa* has the ability to cause severe infections due to its many virulence factors. During acute disease, this pathogen utilizes the toxins of the type III secretion system to circumvent the host immune system and establish infection. Of the four exotoxins (ExoS, ExoT, ExoU, ExoY), ExoU a potent phospholipase that disrupts the plasma membrane and leads to rapid cell death is the most virulent [Sato et al., 2003].

Despite its clinical importance, the molecular basis for the success of the ST235 clone is poorly understood. In addition, although the spread of ST235 has been documented in many geographic locations, little is known about the evolution and emergence of this clone on a global scale.

To better understand the history of the *P. aeruginosa* ST235 lineage as a high-risk international clone, we carried out the process depicted in Figure 3.1. A Bayesian phylogenetic reconstruction using 79 *P. aeruginosa* ST235 isolates collected from five continents over a 27-year period. Genome comparison of these isolates identified antibiotic resistance determinants, virulence genes and ST235-specific genes that may have contributed to the success of this clone.

3.2/ Materials and Methods

3.2.1/ ST235 extraction from NCBI collection

All nucleotide sequences of *Pseudomonas aeruginosa* have been retrieved in November 2014 from the NCBI database. To identify the sequence type, we applied the *blastn* program from the Basic Local Alignment Search Tool (BLAST) 2.2.30+ standalone version, recalled in Section 2.1.3.2 of the state-of-the-art chapter.

It has been launched with 100% identical matches (*pident* parameter) and 100% query coverage per subject (*qcovs*). The query data in blastn was downloaded from the pubmlst database (http://pubmlst.org), according to the following profile: acs-38, aro-11, gua-3, mut-13, nuo-1, pps-2, and trp-4 (7 loci sequences of ST235). Finally, we validated the blastn results with MLST 1.7 tool [Larsen et al., 2012] and found that 14 strains from the NCBI database belong to ST235.

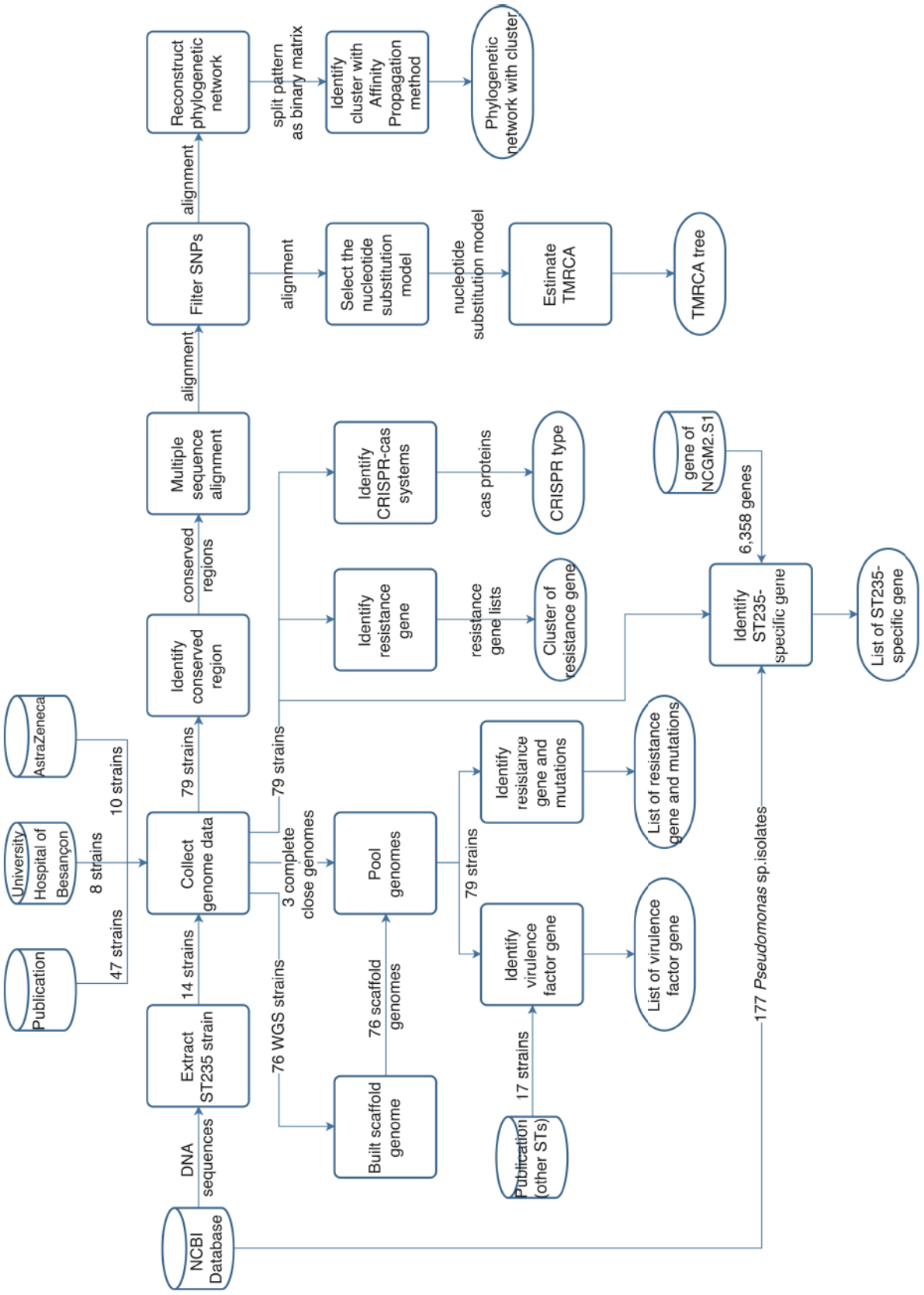


Figure 3.1: **The proposed pipeline.**

3.2.2/ SEQUENCE TYPE (ST) 235 GENOME COLLECTION

The genomes of 79 isolates representing ST235 *P. aeruginosa* clonal cluster were obtained from various sources collected over a 27-year period [van Belkum et al., 2015, Cholley et al., 2014, Hocquet et al., 2012, Kos et al., 2015]. The 17 newly sequenced isolates used in this study were deposited at DDBJ/EMBL/GenBank (see Supplementary material at the end of this chapter, Table 3.2). The isolates came from 23 countries in 5 inhabited continents as follows: Africa ($n = 6$), Asia ($n = 7$), Europe ($n = 34$), North America ($n = 21$), and South America ($n = 11$), see Figure 3.2.

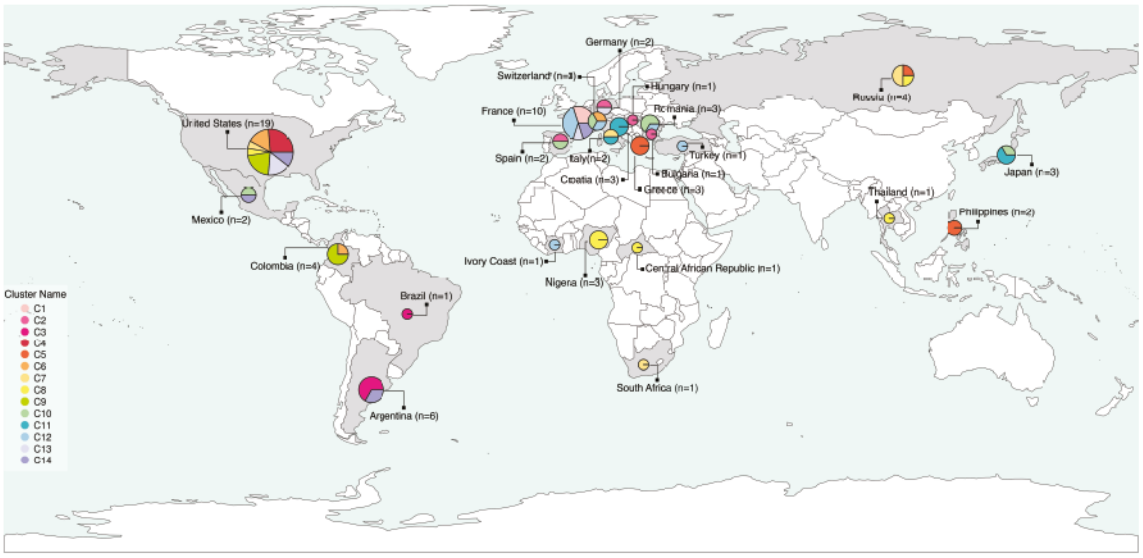


Figure 3.2: **Worldwide distribution of the 79 ST235 isolates of *P. aeruginosa* which genomes were used in this study.** The countries of origin of the isolates are shaded in gray. Pie chart diameters are proportional to the number of isolates collected from each country. The area of each slice is proportional to the quantity of isolates of each cluster.

3.2.3/ CORE GENOME DETERMINATION

To define the 'core' genome of ST235 sublineage, the 79 genomes were aligned to *P. aeruginosa* reference isolate NCGM2.S1 using MUMmer [Kurtz et al., 2004]. Since our data are close strains, we applied the delta-filter of MUMmer with 98% identity to find highly conserved regions and also apply show-coords to extract information such as position, identity percentage, and length of conserved regions.

Among these conserved regions, we only kept the ones that were found in all strains, which approximately represents 61.2% of the size of the reference genome. However, some isolates contain more than one region in the same strict conserved one, which raises difficulties during the alignment process. T-Coffee [Notredame et al., 2000] described previously was used to align multiple

sequences and measure the similarity of each sequence alignment by using the sim_mat_idmat parameter. Thus the chosen alignment of each strict conserved region is the one that has the largest sim_mat_idmat score. The MSA has been preprocessed by removing all gaps before concatenation in the complete alignment.

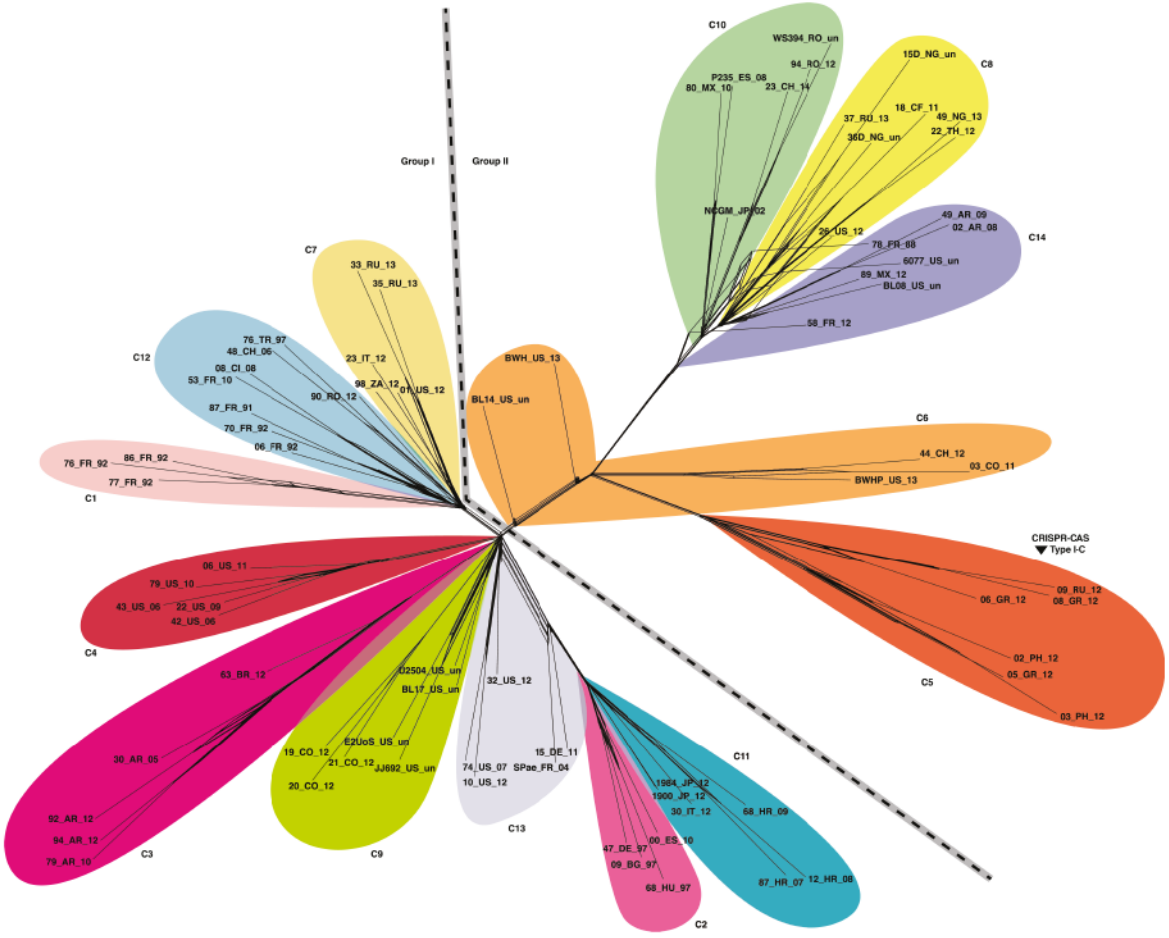


Figure 3.3: **Phylogenetic network of ST235 *P. aeruginosa*.** The alignment of strict conserved regions from 79 isolates was used to construct the phylogenetic network by SplitsTree4 with the NeighborNet method. Bootstrapping was computed for 1,000 replicates then clustered them with APCluster package in R. The clusters, labeled from C1 to C14 and shaded with different colors for clarity, formed two groups (Group I and Group II) separated by a dashed line. Moreover CRISPR-Cas systems have been found in C5 of which is I-C type

3.2.4/ PHYLOGENETIC NETWORK

The evolutionary relationship between the 79 ST235 isolates was investigated using an unrooted phylogenetic network constructed from the sequence alignment

generated by SplitTree4 [Huson and Bryant, 2006] The split network was obtained from a Hamming distance-based method, namely the NeighborNet [Bryant and Moulton, 2002] already presented in this manuscript. Bootstrapping technique was applied to generate 1,000 networks. The average score of the bootstrapped networks is approximately 88%. The consensus is drawn from the bootstraps, so reliable relationship among strain can be captured.

We identified the taxa clusters in consensus network using the split patterns obtained from SplitsTree4. Accordingly, the split patterns were transformed in matrix format with row representing split positions while column are taxa. Then, the transformed split patterns were used to generate clusters of strains by the affinity propagation clustering (APcluster) R-package [Bodenhofer et al., 2011, Frey and Dueck, 2007].

APcluster is an exemplar-based agglomerative clustering technique, which identifies cluster exemplars using a similarity matrix. This package provides many functions for generating the similarity matrix and measuring distances. In this study, we used the following negDisMat function:

$$s(x, y) = -d(x, y)r,$$

where $r = 2$ and d is the Euclidean distance. As a result, 14 clusters were obtained for 79 taxa, which have been labeled from C1 to C14.

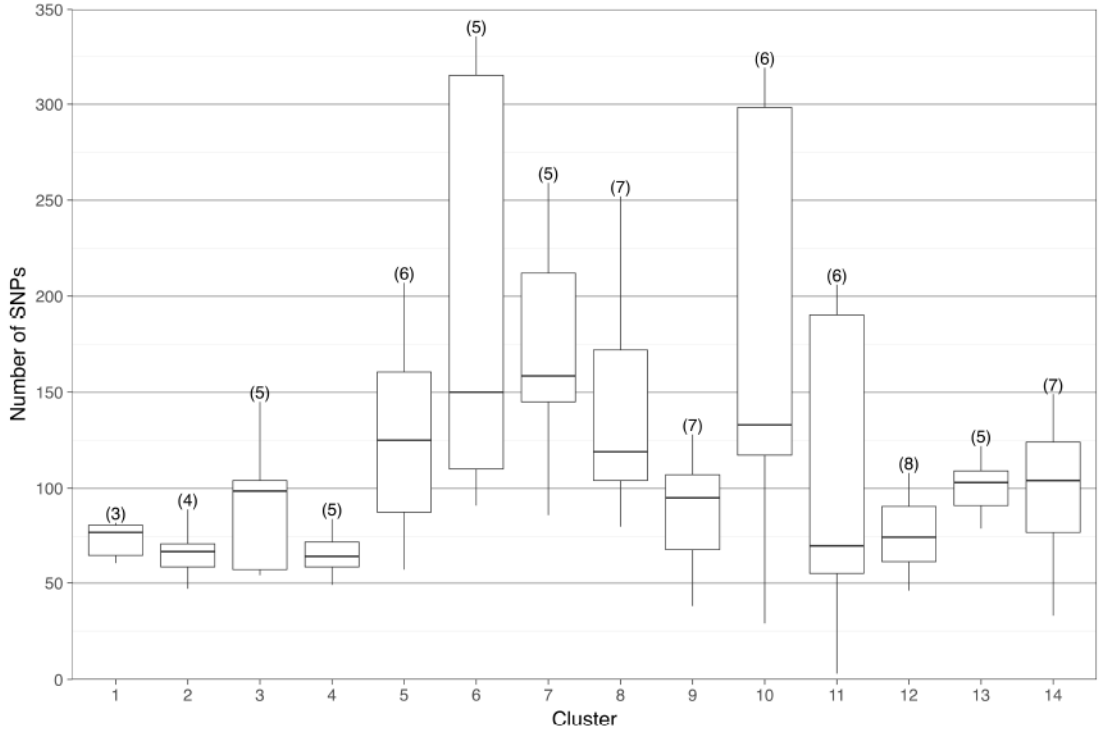


Figure 3.4: **Relations between clusters and SNPs.**

The network is depicted in Figure 3.3, while the relation between SNPs and obtained clusters is represented in Figure 3.4. As can be seen, single nucleotide polymorphism profiles cannot be incriminated in the appearance of our clusters.

3.2.5/ NUCLEOTIDE SUBSTITUTION MODEL SELECTION

As stated in the previous chapter, several substitution models are available for phylogenetic studies, and we need to find the one that fit the most our data. To achieve this goal, one of the most interesting tool is jModelTest 2.1.7 [Darriba et al., 2012, Guindon and Gascuel, 2003], which provides the best-fit models of nucleotide substitution together with other important parameters for phylogeny reconstruction. As recalled in Section 2.1.2.2, it requires an alignment file in phylip format to compute the likelihood scores. This latter has been computed, and we then have launched jModelTest with the following settings: 11 for number of substitution schemes, F for base frequencies, I and G as rate variation, a number of categories of 4, a ML optimized base tree for likelihood calculations, and finally the best of both NNI and SPR for the base tree search.

The likelihood score for the 88 models has been computed, and the best model has been selected according to the AIC, AICc, and BIC criteria. As this model selection has been applied in order to date the most recent common ancestor, our alignment file only contained 69 alignment taxa (the ones whose collection date were known). The obtained model is the so-called general time reversible (GTR), which has been increasingly used over the past decade. Obtained proportion of invariable sites were equal to I = 0.997, while gamma distribution where G = 0.17. Finally, estimated rates were respectively R(a) 0.7306, R(b) 1.9489, R(c) 1.1260, R(d) 0.8309, R(e) 1.3081 and R(f) 1.

3.2.6/ TIME OF THE MOST RECENT COMMON ANCESTOR (TMRCA) ANALYSIS

To determine the most likely evolutionary scenario for the ST235 clone, we calculated the time of the MRCA from the core genome of 69 isolates with known collection dates, using the BEAST2 package [Bouckaert et al., 2014] on a multiTypeTree template. The following parameters have been considered:

- tip dates: year since some time in the past;
- gamma site model: GTR+I+G;
- gamma category count: 4;
- clock model: strict clock;
- clock rate: 1;
- priors model: coalescent constant population;
- and MCMC chain length: 5 millions.

The generated XML file has been parsed with BEAST for tree inference, while the output are log and trees files that represent the evolution events with timing

information and confidence level. The log file is viewed by Tracer to provide the effective sample sizes (ESS) on each parameter. The trees file contains the inferred trees used to construct a consensus tree by summarizing the maximum clade credibility tree to estimate the posterior using TreeAnnotator. The last step is the tree visualization, the final tree from TreeAnnotator can be viewed using FigTree v1.4 [Rambaut, 2014]. It contains details such as posterior probability, age of each node and with credibility interval (height 95% HPD), etc.

3.2.7/ RESISTANCE GENE HIERARCHICAL CLUSTERING

We computed a hierarchical clustering of resistance genes (see Figure 3.2.7) in order to investigate their evolution. ResFinder v2.0 [Zankari et al., 2012] has been used to identify the acquired antimicrobial resistance genes on our data. Genome sequences of our 79 isolates have been uploaded to the CGE server with the following setting: all-antimicrobial, 98% threshold for %ID, and 60% minimum length.

We found 838 genes with some of them containing more than one replicate (812 without duplication). We then focused on the presence status, leading to a binary matrix of resistance genes that has been clustered by the *pvclust* R-package [Suzuki and Shimodaira, 2006]. This package provides a hierarchical clustering with both AU (approximately unbiased) p-value and BP (bootstrap probability) value. AU p-value is computed by multiscale bootstrap resampling, while BP value is computed by a normal bootstrap resampling. Their thresholds were set to 0.95.

3.2.8/ RESISTANCE GENE SEARCH AND MUTATIONS IDENTIFICATION

Genomes of all ST235 isolates were scaffolded using Ragout v2.0b [Kolmogorov et al., 2014] with the *P. aeruginosa* NCGM2.S1 genome as a reference [Kolmogorov et al., 2014], to identify the acquired antimicrobial resistance genes with ResFinder v2.0 [Zankari et al., 2012]. In order to detect mutations that confer high-level resistance to β-lactams and fluoroquinolones, we looked for non-synonymous mutations and insertion sequences in the genes encoding the cephalosporinase AmpC: *ampD*, *ampDh2*, *ampDh3*, *ampG*, *ampO*, *ampP*, *ampR*, *dacB* and its regulators, in oprD, and in the quinolone-determining regions (QRDR) of *gyrA*, *gyrB*, *parC*, and *parE*.

3.2.9/ VIRULENCE FACTOR GENE IDENTIFICATION

All possible virulence factor (VF) genes of *Pseudomonas* were downloaded from virulence factor database (VFDB [Chen et al., 2005]) for using as query. We

searched for the virulence factor genes on 79 scaffold and complete genomes with gmap version 2016-04-04 [Wu and Watanabe, 2005] and filtered the results with 95% identity and 60% coverage. Then we have identified the ST235-specific VFs by comparing VFs list from 79 ST235 isolates with VFs list from 17 other STs isolates of *P. aeruginosa* [Valot et al., 2015].

3.2.10/ ST235-SPECIFIC GENE IDENTIFICATION

ST235-specific genes were defined by firstly searching for all possible ST235-similar genes from alignment to the 6,358 genes present in NCGM2.S1 by GMAP with 95% identity and 60% coverage [Wu and Watanabe, 2005]. Then ST235-specific genes were defined as those present in $\geqslant$ 95% of the genomes ST235 isolates (*i.e.*, $\geqslant 75$ out of the 79 tested ones) and absent from all other isolates of the genus *Pseudomonas* with completely closed genomes available via the NCBI reference sequence database in Oct. 2016 (*i.e.*, 65 non-ST235 *P. aeruginosa* isolates and 177 *Pseudomonas sp.* ones).

3.2.11/ CRISPR-CAS SYSTEMS

Clustered regularly interspaced short palindromic repeat (CRISPR) and CRISPR-associated proteins (CRISPR-Cas) were detected with PILER-CR and BLAT, respectively [Edgar, 2007, Kent, 2002]. The CRISPR-associated proteins subtypes were obtained from NCBI database (accession no. WP_019726852, WP_019726853, WP_034009156, WP_019726855, WP_025981579, WP_025981580, and WP_031628761). Obtained results are depicted in Figure 3.3.

3.3/ RESULTS AND DISCUSSION

3.3.1/ ST235 POPULATION STRUCTURE BY PHYLOGENETIC NETWORK ANALYSIS

A phylogenetic network built from the core genome alignment depicted 14 distinct clusters (C1 to C14; Figure 3.3). The largest sampling of isolates within this collection came from the United States (US, 19 isolates) and France (10 isolates). These isolates were scattered throughout the phylogenetic network, with isolates collected in the US appearing in 7 clusters and those from France in 4 clusters. Focusing on the spatiotemporal distribution of isolates within each cluster, 4 clusters were found to be comprised of isolates from a single country or continent of origin:

- C1 and C2 included 3 and 4 isolates, respectively, from France (C1) and Europe (Spain, Germany, Hungary, and Bulgaria; C2);

- C3 included 5 isolates from South America;
- and C4 included 5 isolates from North America (Figure 3.3).

Although a majority of the ST235 clusters suggests a global dispersion, country-wide spreads are confirmed (Figure 3.2). Isolates from C1, C2, and C3 shared a geographical origin (France for C1, Spain/Germany/Bulgaria/Hungary for C2, Argentina/Brazil for C3; Figure 3.3, and also Table 3.2). In contrast, the remaining clusters included isolates obtained from 2 to 4 continents. Most exemplary of the worldwide spread of the clone ST235 was identified in C8 (7 isolates) which were collected within a 2-year time frame (2011-2013) from 4 distinct continents (Africa, North America, Asia and Europe).

3.3.2/ SPATIOTEMPORAL ORIGIN OF THE ST235 CLONE

The MRCA appeared approximately 32 years ago (≈ 1984) with a very high credibility interval (95% CI, 29.12 to 30.82 years from the date of isolation of the latest isolate in 2014) (Figure 3.5A). The oldest isolate within the current collection was isolated in France in 1988 (78_FR_88), 4 years after the MRCA. The clade of French isolates in the time scale analysis suggests that the clone emerged in this country (Figure 3.5A). A split of more recent isolates in two sub-lineages (Groups I and II, Figure 3.3) was retrieved by both phylogenetic analyses. Hence, 100% of the 1,000 phylogenetic trees that were aggregated to build the network (Figure 3.3) identified the split into the 2 groups I and II. This split was confirmed independently with a Bayesian approach (Figure 3.5) with a posterior probability score of 1.0. These data suggested that the international spread of ST235 implies at least two independent clades. The SNPs that discriminated these two clades were neither in resistance genes nor in virulence genes. Group I consisted of clades that shared a common ancestor that emerged in ≈2001 (C3, C4, C7, C9, C11, and C13) with older isolates that were identified in Europe (C1, C2, and C12; Figure 3.3 and 3.5A). Group II (C5, C6, C8, C10, and C14) shared a common ancestor which emerged in ≈1999 and then appeared to spread worldwide.

3.3.3/ CUMULATIVE RESISTANCE TO ANTIBIOTICS BY CHROMOSOMAL MUTATIONS

We specifically searched for non-synonymous mutations in the QRDRs of each isolate and correlated them with their date of isolation. The two earliest isolates (78_FR_88 and 87_FR_91, retrieved in France in 1988 and 1991, respectively) had wild-type QRDR sequences compatible with a full susceptibility to fluoroquinolones (Figure 3.5B). Among the five isolates retrieved in 1992, four displayed a T83I or a D87Y change in GyrA. Mutations in the QRDR of *parC* first appeared in 1992 in isolate 06_FR_92. Analysis of isolates collected post-1992 showed that the number of QRDR mutations per an isolate seemingly grew over time with the

late isolates (2011 to 2014) accumulating mostly T83I change in GyrA (33 of 38 isolates) and S80L change in ParC (24 of 38 isolates), with occasional additional substitution in position 87 of GyrA, in GyrB, and in ParE (Figure 3.5B).

Mutation-dependent overproduction of intrinsic β-lactamase AmpC is the main cause of resistance of clinical strains of *P. aeruginosa* to antipseudomonal penicillins and cephalosporins [Berrazeg et al., 2015]. Additionally, loss or alteration of the outer membrane porin protein OprD is by far the most common mechanism of resistance to the carbapenems (including imipenem) in *P. aeruginosa* [Fournier et al., 2013]. Hence, we searched for non-synonymous mutations and insertion sequences in genes whose inactivation upregulates AmpC cephalosporinase production and in oprD (Figure 3.5B). Out of the 79 isolates, 36 had acquired one or more AmpC regulator mutations (Figure 3.5B, and also Table 3.3). Of these 36 isolates, 27 produced a mutated transcriptional regulator AmpD. Sequence analysis of the OprD porin showed 39 isolates had acquired mutations which presumably affect the porin activity. The genes *ampD* and *oprD* had 15 and 29 different types of non-synonymous mutations, respectively. Mutations among the AmpC regulators and OprD were mostly unique in contrast to the relative uniformity in the QRDR mutations (T83I in GyrA, S80L in ParC) in the late isolates.

3.3.4/ High diversity of foreign antibiotic resistant determinants among the ST235 isolates

Figure 3.7 details the antibiotic resistance genes acquired by the isolates (see also Figure 3.8). Analysis of the aminoglycoside-modifying enzyme (AME) content among the collection revealed that a majority (72 out of 79) of ST235 isolates harboured at least one AME potentially conferring a decrease in aminoglycoside susceptibility (cf. Figure 3.8).

Approximately 40% (32 out of 79) of the isolates possessed at least one of 23 β-lactamases characterized as having extended spectrum activity(Figure 3.7). Seven isolates accumulated two β-lactamases with an extended spectrum, mostly combining the production of an extended-spectrum oxacillinase (ES-OXA) with that of an extended-spectrum β-lactamase (ESBL) or a metallo-β-lactamase (MBL). Extended-spectrum enzymes were generally accumulated within specific sub-lineages. For example, bla_{OXA-17} and $bla_{OXA-129}$ were uniquely identified in isolates belonging to C3, bla_{OXA-19} to C5, bla_{KPC-2} to C9, bla_{GES-19}, bla_{PER-1} and bla_{OXA-74} to C10, and bla_{IMP-34} to C11 (Figure 3.7). Within C5, the three Greek isolates harboured either bla_{VIM-2}, bla_{VIM-4}, bla_{OXA-19}, or bla_{OXA-35} while the two isolates from Philippines had bla_{IMP-26} or bla_{VIM-2}. Other rare extended-spectrum enzymes were shared by clusters with bla_{IMP-1} harboured by isolates from C8 (1 of 7 isolates) and C10 (1 of 6 isolates), and bla_{VIM-2} present in isolates of C5 (2 of 6 isolates) and C7 (2 of 5 isolates) which were most likely acquired independently. Acquired resistance determinants were not detected within six of the isolates of the collection (C6_03_CO_11, C6_44_CH_12, C6_BL14_US_un, C6_BWH_US_13, C6_BWHP_US_13, and C14_58_FR_12) with five of these isolates originating from

diverse geographic locations clustering in C6 (see Figure 3.8). The number of acquired resistance genes was significantly lower (p-value, 0.05) in the isolates collected at an earlier time point (1988-1997, 11 of isolates) than those collected between 2011-2014 (38 of isolates).

The variety of antibiotic resistance genes indicates independent and local acquisition of resistance genes by members of the clone ST235, consistent with observations for other 'high-risk' international clones and that the success of these lineages in causing infection is not exclusively dependent on antibiotic resistance [Turton et al., 2015, Oliver et al., 2015].

Altogether, the mosaic of either acquired or mutational resistance determinants to extended-spectrum cephalosporins and to carbapenems suggests that these antimicrobial agents played a minor role in the international spread of ST235 clone (Figure 3.7, Figure 3.5B).

3.3.5/ CRISPR-Cas type I-C detected within Cluster 5

CRISPR-Cas bacterial adaptive immune systems play an important role in shaping the accessory genomes of *P. aeruginosa* but are not thought to prevent the acquisition of antibiotic resistance elements [van Belkum et al., 2015]. A search of the 79 ST235 isolates identified only six ST235 isolates harbouring a CRISPR-Cas system; all being of type I-C and falling within C5 of the phylogenetic tree (Figure 3.3). The cluster C5 included isolates from the Philippines ($n = 2$) and Greece ($n = 3$) along with one isolate from Russia. Interestingly, all C5 isolates were from 2012 and shared a common ancestor dating from 2007 (Figure 3.5A). Such a low prevalence of CRISPR-Cas in the ST235 lineage is in line with previous data [van Belkum et al., 2015, Touchon et al., 2012].

The C5 type I-C CRISPR-Cas isolates also carried acquired resistance determinants previously identified to be associated with class I integrons [Kos et al., 2015]. This is in line with the absence of clear correlation between the acquisition of antibiotic resistance genes via plasmids and class I integrons in Gram-negative bacilli and the presence of CRISPR-Cas systems [Touchon et al., 2012, van Belkum et al., 2015]. The fact that the average size of the chromosome of CRISPR-Cas containing C5 isolates (6.88 Mbp) is similar to that of the rest of the ST235 clone (6.87 Mbp) gives indirect evidence of the poor role of CRISPR-Cas systems in the acquisition of foreign DNA. *P. aeruginosa* CRISPR-Cas systems may then be considered preferably as a protection against temperate bacteriophages or genomic islands rather than against the transfer of resistance genes directly.

3.3.6/ ST235-specific determinants

One could imagine that the global success of *P. aeruginosa* international clones (e.g. ST111, ST175, ST235, ST244, ST395) could rely on specific determi-

nants. Unfortunately, we did not retrieve any gene signature shared among these widespread clones (data not shown). We then identified genes present in the genome of all ST235 isolates but absent from a majority of the non-ST235 isolates. The type III secretion system exotoxin encoded by the *exoU* gene was present in all of the ST235 genomes, although not specific of this lineage. Hence, the analysis of 65 non-ST235 genomes of *P. aeruginosa* also detected the presence of the *exoU* gene in 18 strains representative of ST253 (e.g. strain PA14), ST313, ST316, ST357, ST823, ST1024, ST1047, ST1971, and an unassigned ST (strain MTB-1). We further identified 22 ST235-specific genes. A majority of these genes clustered into three blocks (Table 3.1).

Block 1 contained 9 genes and was a part of ExoU island A. Three genes (NCGM2_1830 to NCGM2_1832) encoded homologs of the components (TolC, EmrA, EmrB) of a tripartite efflux pump. Two others contiguous genes (NCGM2_1836, NCGM2_1837) encoded a putative transporter and a periplasmic adaptor, respectively, that could act together as another efflux pump. The production of these two efflux pumps could specifically enhance the resistance of ST235 isolates to antibiotics. Block 2 included 10 genes with 9 encoding proteins implicated in DNA processing (a P-loop NTPase, two type-1 restriction endonucleases - HsdR and HsdS, a UvrD/REP helicase, a SMC domain-containing protein, two DNA methylases, DprA and RecQ).

3.4/ Discussion

3.4.1/ ST235-specific determinants

We identified *dprA* as a specific determinant of ST235 sublineage. DprA is required for protection of incoming single-stranded DNA and interacts with the ubiquitous recombinase RecA to integrate the acquired DNA into the host chromosome of naturally transformable bacterial species such as *Streptococcus pneumoniae* [Johnston et al., 2014]. RecQ is a DNA helicase that affects DNA transformation in Gram-negative and Gram-positive bacteria [Bolotin et al., 2001]. Of note, the rest of the transformation machinery (*i.e.*, secretion channel PilQ, DNA receptor ComE, transmembrane channel ComA, translocase ComFA, and the competence activator Tfox) was conserved in the species *P. aeruginosa* [Johnston et al., 2014, Valot et al., 2015].

ST235 displays a very high diversity of acquired resistance genes (Figure 3.7 and 3.8). DprA universality among transformable species and its demonstrated role in homologous recombination provide evidence that its presence in ST235 possibly increases the ability of this widespread clone to acquire and maintain foreign resistance elements at a greater rate than other *P. aeruginosa* clones. Overall, 22 genes were found to be unique to the ST235 sublineage and encoded proteins implicated in DNA processing, transport through the membrane, and bacterial transformation. The role of these ST235-specific proteins in the suc-

Table 3.1: **Description of the 22 genes highly conserved in and specific to *P. aeruginosa* ST235 lineage.**

Block number	Gene symbol in strain NCGM2.S1	Domain	Additional description	Accession no. of the closest homolog (name, bacterial species, % identity)
1	*NCGM2_1826*	Transposase	-	-
	NCGM2_1828	Alpha/beta hydrolase family protein	-	-
	NCGM2_1829	Pirin-related protein	Putative transcriptional regulation	-
	NCGM2_1830	Putative RND outer membrane protein (TolC family)	-	NP_417507 (TolC, *E. coli*, 22%)
	NCGM2_1831	Putative RND membrane fusion protein (ErmA family)	-	NP_417170.1 (EmrA, *E. coli*, 50%)
	NCGM2_1832	Putative MFS multidrug efflux transporter (EmrB family)	-	NP_418166.1 (EmrB, *E. coli*, 24%)
	NCGM2_1836	Putative transporter membrane protein	-	WP_058142560.1 (*P. aeruginosa*, 99%)
	NCGM2_1837	Putative RND membrane fusion protein (HlyD/EmrA family)	-	-
	NCGM2_1838	PucR C-terminal helix-turn-helix	Probable transcriptional regulator	NP_391122.1 (*B. subtilis*, 45%)
2	*NCGM2_3761*	Hypothetical protein	-	-
	NCGM2_3762	P-loop NTPase	Involved in replication	WP_025991883.1 (*P. aeruginosa*, 99%)
	NCGM2_3765	Type-I restriction endonuclease HsdR	Restriction-modification system	-
	NCGM2_3766	Type-I restriction endonuclease HsdS	Restriction-modification system	-
	NCGM2_3767	UvrD/REP helicase	-	-
	NCGM2_3768	SMC domain-containing protein	Replication, recombination and DNA repair	-
	NCGM2_3769	N-6 DNA methylase	-	-
	NCGM2_3770	N-7 DNA methylase	-	-
	dprA	DNA protection protein	Dedicated to natural bacterial transformation	-
	recQ	ATP-dependent DNA helicase	Involved in genome maintenance	-
	leuS	Leucyl-tRNA synthetase	-	-
3	*NCGM2_6332*	Hypothetical protein	-	-
	NCGM2_6333	DEAD/DEAH box helicase	-	-

cess of the clone (via higher intrinsic resistance to antibiotics or easier acquisition of foreign resistance determinants) is speculative but deserves experimental verification.

Although not fully specific of the ST235 clone, exoU-encoded exotoxin was retrieved in all ST235 isolates. The presence and production of ExoU is a marker for early mortality associated with P. aeruginosa infections. The virulence in *P. aeruginosa* is multifactorial and combinatorial, however these data suggest that ExoU production could participate to the poor outcome of infections due to ST235.

3.4.2/ High diversity of resistance determinants to aminoglycosides and to β-lactams in ST235 isolates

The variety of both chromosomal and acquired determinants of resistance to aminoglycosides and β-lactams in the clone ST235 (Figure 3.5, Figure 3.7) indicates independent and local acquisition in line with previous observation. Hence, one can hypothesize that the global spread of these lineages does not fully rely on resistance to aminoglycosides and to β-lactams. This high diversity of foreign resistance determinants in ST235 contrasts with that observed in another widespread clone, ST175, in which the resistance to antibiotics mainly occurs via chromosomal mutations. Altogether, the mosaic of either acquired or mutational resistance determinants could argue for a limited role of the selective pressure of these antimicrobial agents in the international spread of ST235 clone (Figure 3.7, Figure 3.5B).

3.4.3/ Role of the fluoroquinolones in the spread and emergence of ST235

Interestingly, the emergence of ST235 sublineage in 1984 coincides with the beginning of the use of antipseudomonal fluoroquinolones (pefloxacin, ofloxacin, and ciprofloxacin) between 1984 and 1987 [Rubinstein, 2001]. The emergence of fluoroquinolone-resistant mutants of *P. aeruginosa* after treatment with these compounds has been well documented [Ball, 1990, Carmeli et al., 1999]. The expansion of ST235 may have been favoured by the extensive use of fluoroquinolones, selecting for QRDR mutations. Similarly, the worldwide expansion of the major pathogens *Escherichia coli* ST131 H30-Rx, methicillin-resistant *Staphylococcus aureus* EMRSA-15 to the mid-1980s, and that of *Clostridium difficile* 027 in the early 1990s also incriminated the clinical use of fluoroquinolones [Ben Zakour et al., 2016, Holden et al., 2013, He et al., 2013].

Although QRDR mutations have been observed in many other clones of *P. aeruginosa*, it has been experimentally demonstrated that the fitness cost of mutations in the QRDR of *P. aeruginosa* depends on the genetic background of the

strains [Kugelberg et al., 2005]. Hence, it has been recently shown that mutations in *parC* increase or have no effect on the fitness of *exoU*-strains, while they decrease that of *exoS*-strains [Agnello et al., 2016]. This difference was due, at least in part, to a better ability of ExoU-strains to maintain the DNA supercoiling levels of the parent strain. In addition, Agnello et al. also demonstrated that *parC*-mutants of ExoU-strains, after 7 days of culture have a higher mutation frequency than *parC*-mutants of *exoS*-strains [Agnello et al., 2016]. Overall, although other features of ST235 may have contributed to its spread, fluoroquinolone use could have highly favored the spread of *exoU*-strains (including ST235) in which the fitness burden of the resistance to these major antibiotics is lower.

3.4.4/ Weaknesses and limitations of the study

Although a thorough genotypic analysis was completed on these isolates, the corresponding phenotypic antibiotic resistance level of a majority of the isolates was unknown. However, previous studies have illustrated significant correlation between the genotype and phenotype for β-lactams, fluoroquinolones and cephalosporins. Despite the particular efforts we made to collect 'historic' ST235 genomes collected pre-1992 and the high degree of confidence of the MRCA age, the relative low number of early isolates could have biased the result. Additionally, the isolate collection explored here (n=79) is of limited size and suffers from a bias in geographical distribution, with a low representation of isolates coming from Asia and Africa when compared with isolates coming from North America and Europe.

3.5/ Conclusion

Analysis of the genome sequences of 79 isolates of *P. aeruginosa* ST235 obtained over a 27-year period from diverse regions of the globe were used to gain an understanding of the epidemiology of this international high-risk clone. Clustering analysis confirmed that clonal spreads occur on a country or regional scale but also revealed that ST235 subclones can spread across continents. Analysis of a time scale of a phylogenetic tree suggested that the ST235 disseminated very recently since the late 20st century ($\approx$ 1984). The ST235 seemingly spread from Europe worldwide in the 1998-2000 with at least two independent groups. The very high diversity of acquired resistance genes, which contrasts with the relatively recent age of the ST235 ancestor, is typical of high-risk clones and has also been reported within the ST111 clone [Turton et al., 2015]. It indicates independent and local acquisition of mobile resistance genes. In addition, the mosaic of mutational and acquired resistance determinants to extended-spectrum cephalosporins and carbapenems suggests that these antimicrobial agents played a minor role in the international spread of the ST235 clone. In contrast, the date of emergence of the clone and the similar QRDR change in nearly all isolates retrieved after 1992

strongly suggests that ST235 worldwide spread has been favored by the extensive use of fluoroquinolones since the mid-1980s. Although pinpointing the genetic basis of the success of an epidemic pathogen is complex, two recent studies demonstrated that the emergence of *Escherichia coli* ST131 H30-Rx and that of *Staphylococcus aureus* EMRSA-15 relied on the acquisition of a virulence factor or on genetic changes leading to an adaptation to hospital environment along with the acquisition of drug resistance [Ben Zakour et al., 2016, Holden et al., 2013] (31, 32). Similarly, once arriving in a region, ST235 *P. aeruginosa* subsequently acquire resistance determinants to aminoglycosides, β-lactams and carbapenems to create local outbreaks with poor outcome due in part to the production of ExoU. We found that most ST235 isolates did not harbor CRISPR-Cas systems with the exception of isolates of cluster C5. Twenty-two genes were found to be unique to ST235 sublineage and encoded proteins implicated in DNA processing, transport through the membrane, and bacterial transformation. These two latter features could have helped this high-risk clone to enhance its intrinsic resistance to toxic compounds and to acquire new determinants of resistance to major antipseudomonal antibiotics.

In summary, *P. aeruginosa* ST235 clone has become prevalent across the globe due to the selective pressure of antimicrobials and their ability to adapt through mutation and readily acquire resistance elements.

3.5.1/ DATA ACCESS

The sequence data from this study have been submitted to the DDBJ/EMBL/GenBank under NCBI project id. PRJNA311177. Raw sequencing data are available under accession no. SRP076542.

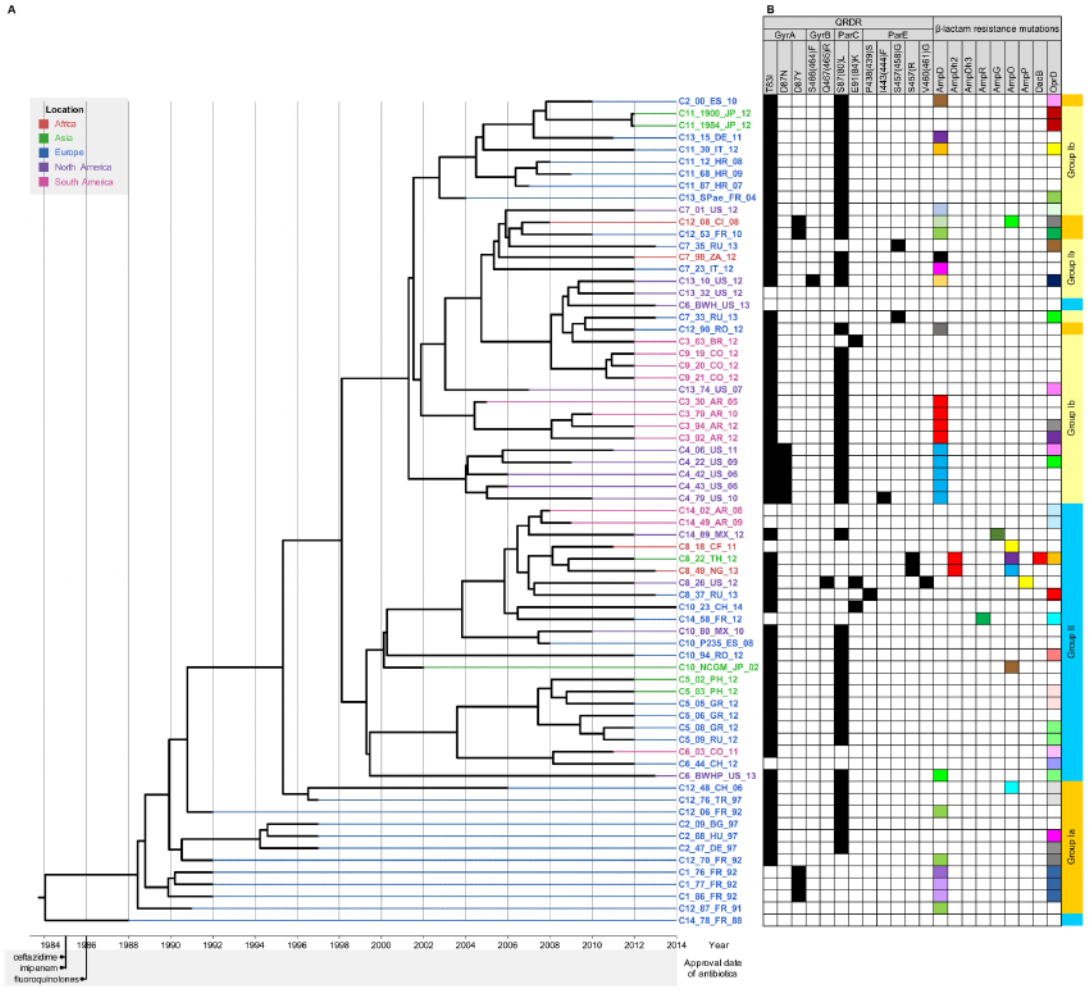


Figure 3.5: **Time of Most Recent Common Ancestor (MRCA) of *P. aeruginosa* ST235 and chromosomal mutations conferring high-level resistance to fluoroquinolones, extended-spectrum cephalosporinases, and carbapenems.** (A) Phylogenetic tree with time scale was calculated from the 69 genomes of isolates with known isolation date with BEAST2 using continent origin as co-variable. The estimated mutation rate was 4.85 10-6 (95% CI, 4.59 10-6 - 5.15 10-6) per site per year. The time of MRCA is ≈ 30 years ago from 2014. The tips are labeled with the isolate name and are colored by continent of origin (see insert). Names of the isolates are prefixed with the cluster to which they belong. Isolates with type I-C CRISPR-Cas system are marked with an asterisk. (B) Mutations in the QRDR of *gyrA*, *gyrB*, *parC*, and *parE*, in the regulators of the cephalosporinase AmpC and in *oprD*. For the QRDR, the numbers of the corresponding codons in *Escherichia coli* are in parentheses. Black cells and white cells indicate the presence or absence of a given mutation, respectively. For mutation in each regulator of the cephalosporinase AmpC and in *oprD*, every single mutation is represented by a single color. White cells indicate an intact protein. The detail of the mutations is given in the Table 3.3. Every protein was compared to its closest homolog born by a β-lactam susceptible isolate of *P. aeruginosa* (strain M18 for *ampD*, *ampP*, *dacB* ; strain PA14 for *ampDh2*, *ampG*, *ampO*, *oprD* ; strain MBT-1 for *ampDh3*, *ampR*). *ampO* polymorphism for all the isolates of the collection: S125T, D256E. *ampP* polymorphism for all the isolates of the collection: L74F and L98F.

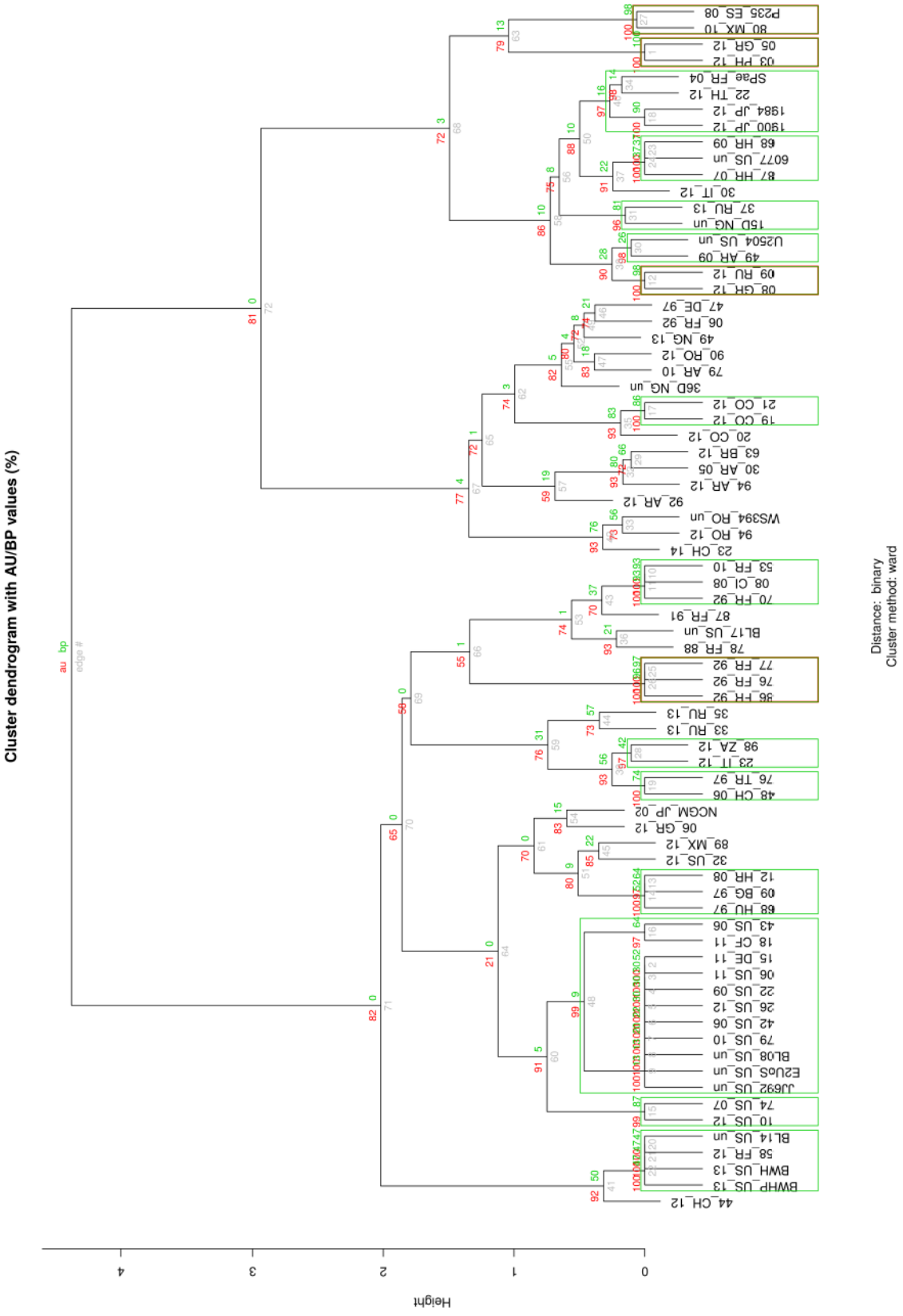


Figure 3.6: **The hierarchical clustering of resistance gene.**

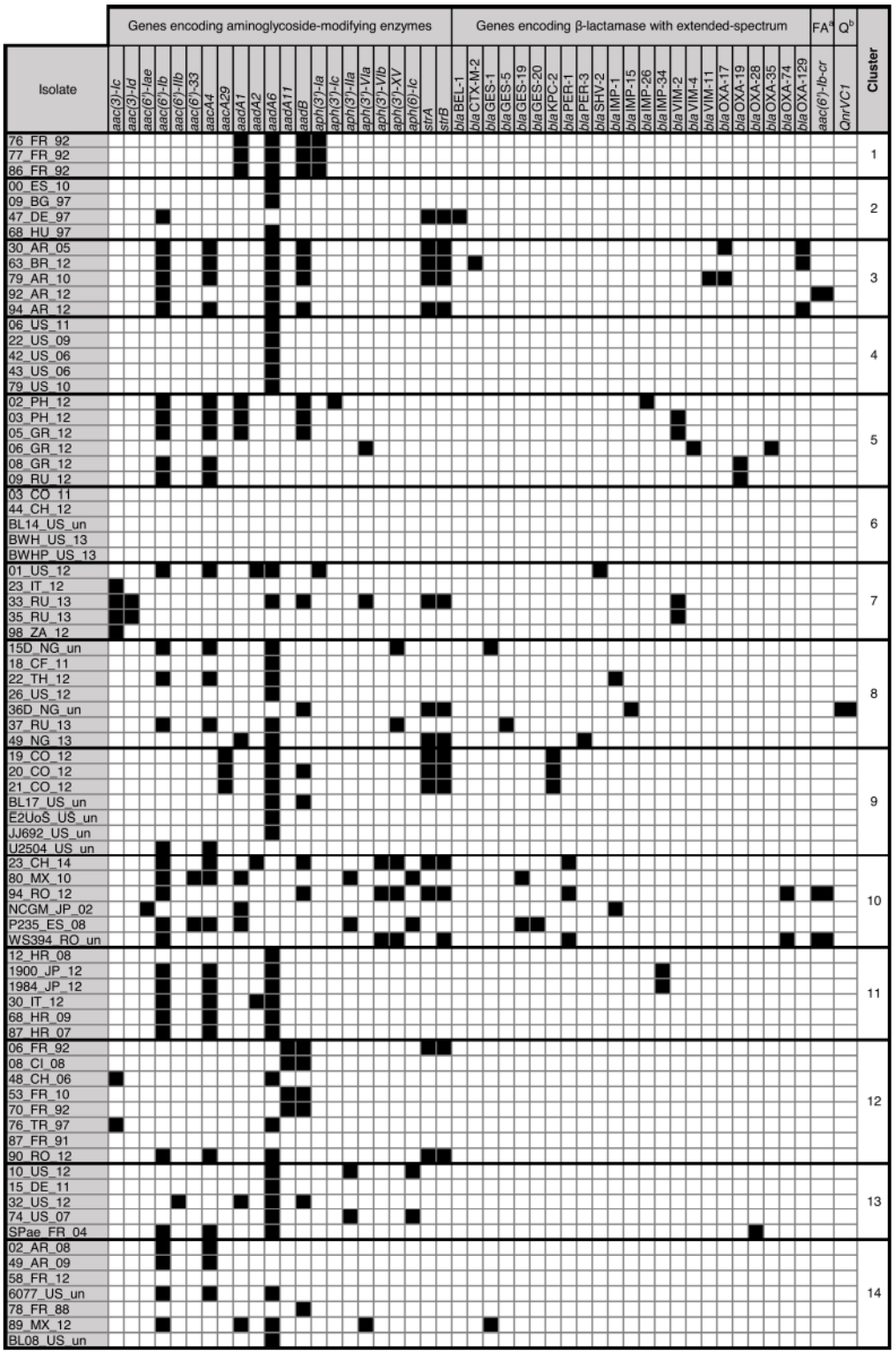


Figure 3.7: **Genes acquired by ST235 isolates that confer resistance to aminoglycosides, to extended-spectrum β-lactams, to carbapenems, and fluoroquinolones.** [a] Genes encoding resistance determinants to fluoroquinolones and to aminoglycosides. [b] Genes encoding resistance determinants to fluoroquinolones. The gene *aph(3')-IIb* was present in all the isolates (except 44_CH_12, 43_US_06 and 18_CF_11). The genes *bla* $_{OXA-50}$, *bla* $_{AmpC}$, and *fosA* were present in all the isolates tested.

Table 3.2: **Origin and metadata of the 79 *P. aeruginosa* ST235 isolates used to determine the worldwide population structure of this clone.**

Strain name[a]	Origin	Site of isolation	Country of isolation	Year of isolation	Total sequence length (bp.)	Number of contigs	GI number	NCBI Reference Seq.	Reference
1900_JP_12	Human	Urine	Japan	2012	6814936	1	AP014622	AP014622	Unpublished
1984_JP_12	Human	Urine	Japan	2012	6850954	1	AP014646	AP014646	Unpublished
6077_US_un	Human	Eye	United States	Unknown	6971023	45	553794551	NZ_AXRE00000000	Unpublished
BL08_US_un	Human	Eye	United States	Unknown	6946790	31	553772209	NZ_AXPS00000000	Unpublished
BL14_US_un	Human	Eye	United States	Unknown	7101427	24	553768504	NZ_AXPM00000000	Unpublished
BL17_US_un	Human	Eye	United States	Unknown	6865653	18	553766645	NZ_AXPJ00000000	Unpublished
BWH_US_13	Human	Urine	United States	2013	6451019	66	685985247	NZ_JIED00000000	Unpublished
BWHP_US_13	Human	Lung	United States	2013	6930020	20	553793043	NZ_AXQW00000000	Unpublished
E2UoS_US_un	Environment	Soil	United States	Unknown	6634585	341	514930350	NZ_ASQV00000000	[1]
JJ692_US_un	Human	Urine	United States	Unknown	6735382	17	553797072	NZ_AXRH00000000	Unpublished
NCGM_JP_02	Human	Urine	Japan	2002	6764661	1	AP012280	AP012280	[2], [3]
P235_ES_08	Human	Blood	Spain	2008	6929551	123	671279092	JNHD00000000	[4]
U2504_US_un	Human	Urine	United States	Unknown	7049946	28	553796923	NZ_AXRG00000000	Unpublished
WS394_RO_un	Human	Skin	Romania	Unknown	6754356	198	698162563	CBYA000000000	[5]
01_US_12	Human	Urine	United States	2012	6735127	137	1125616863	LYTT00000000	[6][b]
09_RU_12	Human	Urine	Russia	2012	6736869	154	1125616898	LYTU00000000	[6][b]
22_TH_12	Human	Lung	Thailand	2012	6840910	178	1125617038	LYTV00000000	[6][b]
33_RU_13	Human	Lung	Russia	2013	7068739	363	1125617870	LYTW00000000	[6][b]
35_RU_13	Human	Urine	Russia	2013	6881324	210	1125617615	LYTX00000000	[6][b]
37_RU_13	Human	Abdomen	Russia	2013	7109932	352	1125618029	LYTY00000000	[6][b]
49_NG_13	Human	Abdomen	Nigeria	2013	6888143	161	1125618040	LYTZ00000000	[6][b]
92_AR_12	Human	Lung	Argentina	2012	6829516	163	1125616452	LYTQ00000000	[6][b]
94_AR_12	Human	Lung	Argentina	2012	6819863	186	1125616528	LYTR00000000	[6][b]
98_ZA_12	Human	Abdomen	South Africa	2012	6803192	160	1125616460	LYTS00000000	[6][b]
00_ES_10	Human	Urine	Spain	2010	6858205	108	728982976	JTOU00000000	[7]
02_AR_08	Human	Lung	Argentina	2008	6886932	98	728993746	JTSN00000000	[7]
02_PH_12	Human	Lung	Philippines	2012	7131510	160	729010743	JTXD00000000	[7]
03_CO_11	Human	Lung	Colombia	2011	6882469	106	728982655	JTOR00000000	[7]
03_PH_12	Human	Abdomen	Philippines	2012	6892377	201	729010642	JTXC00000000	[7]
05_GR_12	Human	Urine	Greece	2012	6888268	240	729010543	JTXA00000000	[7]
06_GR_12	Human	Abdomen	Greece	2012	6921381	149	729009856	JTWZ00000000	[7]
06_US_11	Human	Lung	United States	2011	6832413	109	728982336	JTOO00000000	[7]
08_GR_12	Human	Abdomen	Greece	2012	6707536	187	729009836	JTWX00000000	[7]
10_US_12	Human	Lung	United States	2012	6748304	147	729009409	JTWW00000000	[7]
12_HR_08	Human	Lung	Croatia	2008	6835920	89	728992820	JTSD00000000	[7]
15_DE_11	Human	Urine	Germany	2011	6804229	113	728981344	JTOF00000000	[7]
19_CO_12	Human	Lung	Colombia	2012	6855346	190	729007718	JTWN00000000	[7]
20_CO_12	Human	Urine	Colombia	2012	6802126	235	729007823	JTWM00000000	[7]

To be continued

Strain name[a]	Origin	Site of isolation	Country of isolation	Year of isolation	Total sequence length (bp.)	Number of contigs	GI number	NCBI Reference Seq.	Reference
21_CO_12	Human	Abdomen	Colombia	2012	6867543	164	729007613	JTWL00000000	[7]
22_US_09	Human	Urine	United States	2009	6732223	90	729015496	JTYK00000000	[7]
23_IT_12	Human	Lung	Italy	2012	6639036	171	729007139	JTWJ00000000	[7]
26_US_12	Human	Urine	United States	2012	6885273	164	729006706	JTWG00000000	[7]
30_AR_05	Human	Abdomen	Argentina	2005	6816534	150	729003105	JTVE00000000	[7]
30_IT_12	Human	Lung	Italy	2012	6865760	198	729005870	JTWC00000000	[7]
32_US_12	Human	Lung	United States	2012	6793558	196	729005659	JTWA00000000	[7]
42_US_06	Human	Urine	United States	2006	6889407	80	729001609	JTUS00000000	[7]
43_US_06	Human	Urine	United States	2006	6737478	125	729001522	JTUR00000000	[7]
49_AR_09	Human	Lung	Argentina	2009	6899528	77	728988868	JTQS00000000	[7]
53_FR_10	Human	Abdomen	France	2010	6835254	123	729016788	JTZA00000000	[7]
58_FR_12	Human	Lung	France	2012	6567883	103	728977063	JTMO00000000	[7]
63_BR_12	Human	Lung	Brazil	2012	6768150	113	728976370	JTMJ00000000	[7]
68_HR_09	Human	Lung	Croatia	2009	6897223	123	728986708	JTQA00000000	[7]
74_US_07	Human	Lung	United States	2007	6850351	170	728997582	JTTO00000000	[7]
79_AR_10	Human	Unknown	Argentina	2010	6948193	132	729017191	JTZD00000000	[7]
79_US_10	Human	Urine	United States	2010	6751299	125	728985228	JTPP00000000	[7]
80_MX_10	Human	Unknown	Mexico	2010	6855952	112	729016970	JTZC00000000	[7]
87_HR_07	Human	Abdomen	Croatia	2007	6903570	114	728995829	JTTB00000000	[7]
89_MX_12	Human	Lung	Mexico	2012	7252889	216	729013444	JTXP00000000	[7]
90_RO_12	Human	Urine	Romania	2012	6730564	106	729013122	JTXO00000000	[7]
94_RO_12	Human	Urine	Romania	2012	6964286	268	729012972	JTXK00000000	[7]
06_FR_92	Human	Unknown	France	1992	6847965	74	956329076	LLOT00000000	[8]
09_BG_97	Human	Skin	Bulgaria	1997	6701776	136	957353945	LLUY00000000	[8]
47_DE_97	Human	Skin	Germany	1997	6868312	143	957106286	LLTH00000000	[8]
68_HU_97	Human	Skin	Hungary	1997	6763374	115	957253774	LLTZ00000000	[8]
70_FR_92	Human	Unknown	France	1992	6813449	119	955923662	LLLZ00000000	[8]
76_FR_92	Human	Unknown	France	1992	6730794	163	955940558	LLMF00000000	[8]
76_TR_97	Human	Skin	Turkey	1997	6723240	103	957283235	LLUC00000000	[8]
77_FR_92	Human	Unknown	France	1992	6761710	180	955995171	LLMG00000000	[8]
78_FR_88	Human	Unknown	France	1988	7489959	199	956281714	LLOC00000000	[8]
86_FR_92	Human	Unknown	France	1992	6754761	167	956029546	LLMQ00000000	[8]
87_FR_91	Human	Unknown	France	1991	6566958	88	956314600	LLOL00000000	[8]
08_CI_08	Human	Unknown	Ivory Coast	2008	6780875	1584	-	SRX1844435	[9][c]
15D_NG_un	Human	Unknown	Nigeria	Unknown	7368341	2204	-	SRX1844429	[9][c]
18_CF_11	Human	Wound	Central African Republic	2011	7024837	1937	-	SRX1844430	[9][c]
23_CH_14	Human	Urine	Switzerland	2014	7527497	1667	-	SRX1844434	Unpublished, D.S. Blanc[c]
36D_NG_un	Human	Unknown	Nigeria	Unknown	7160894	992	-	SRX1844431	[9][c]
44_CH_12	Human	Ear	Switzerland	2012	6658646	1979	-	SRX1844433	Unpublished, D.S. Blanc[c]

To be continued

Strain name[a]	Origin	Site of isolation	Country of isolation	Year of isolation	Total sequence length (bp.)	Number of contigs	GI number	NCBI Reference Seq.	Reference
48_CH_06	Human	Nose	Switzerland	2006	7316038	1727	-	SRX1844432	Unpublished, D.S. Blanc[c]
SPae_FR_04	Human	Lung	France	2004	6717652	130	1125603363	LYLN00000000	[10]
End of table									

[1] Stewart L, Ford A, Sangal V, Jeukens J, Boyle B, Kukavica-Ibrulj I, et al. Draft genomes of 12 host-adapted and environmental isolates of *Pseudomonas aeruginosa* and their positions in the core genome phylogeny. Pathog Dis 2014;71:20-5

[2] Shimizu W, Kayama S, Kouda S, Ogura Y, Kobayashi K, Shigemoto N, et al. Persistence and epidemic propagation of a *Pseudomonas aeruginosa* sequence type 235 clone harboring an IS26 composite transposon carrying the blaIMP-1 integron in Hiroshima, Japan, 2005 to 2012. Antimicrob Agents Chemother 2015;59:2678-87.

[3] Sekiguchi J-I, Asagi T, Miyoshi-Akiyama T, Fujino T, Kobayashi I, Morita K, et al. Multidrug-resistant *Pseudomonas aeruginosa* strain that caused an outbreak in a neurosurgery ward and its *aac(6')-Iae* gene cassette encoding a novel aminoglycoside acetyltransferase. Antimicrob Agents Chemother 2005;49:3734-42.

[4] Viedma E, Villa J, Juan C, Oliver A, Chaves F. Draft genome sequence of colistin-only-susceptible *Pseudomonas aeruginosa* strain ST235, a hypervirulent high-risk clone in Spain. Genome Announc 2014;2. doi:10.1128/genomeA.01097-14.

[5] Vorhölter F-J, Arnold M, Wibberg D, Blom J, Winkler A, Viehoever P, et al. Draft genome sequence of *Pseudomonas aeruginosa* strain WS394, a multidrug-resistant and highly cytotoxic wound isolate from chronic ulcus cruris. Genome Announc 2014;2. doi:10.1128/genomeA.01325-14.

[6] Unpublished, International Health Management Association, AstraZeneca

[7] Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, et al. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. Antimicrob Agents Chemother 2015;59:427-36.

[8] van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras J-B, Barbu EM, et al. Phylogenetic distribution of CRISPR-Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. MBio 2015;6:e01796-15.

[9] Cholley P, Ka R, Guyeux C, Thouverez M, Guessennd N, Ghebremedhin B, et al. Population structure of clinical *Pseudomonas aeruginosa* from west and central African countries. PLoS One 2014;9:e107008.

[10] Hocquet D, Llanes C, Thouverez M, Kulasekara HD, Bertrand X, Plésiat P, et al. Evidence for induction of integron-based antibiotic resistance by the SOS response in a clinical setting. PLoS Pathog 2012;8:e1002778.

[a] The name of isolates was defined as the concatenation of the isolate number or strain name, the country of collection (using the ISO 3166 2-letter country code), and the year of collection (2 figures, or “un” for unknown when the date of isolation was missing).

[b] Sequences generated by Illumina and raw read data was assembled by CLC Genomic Workbench 7.0.4.

[c] Sequences generated by Ion Torrent PGM and raw read data was assembled by Ray.

Table 3.3: **Details of the mutations in genes whose inactivation up-regulates *AmpC* cephalosporinase production and in *oprD* in a collection of 79 isolates of *Pseudomonas aeruginosa* ST235.**

Cluster	Isolate	*ampD*	*ampDh2*	*ampDh3*	*ampR*	*ampG*	*ampO*	*ampP*	*dacB*	*oprD*
C1	76_FR_92	IS*Pa7*[a]	-	-	-	-	-	-	-	IS*Ppu21*[b]
	77_FR_92	IS*Pa7*[c]	-	-	-	-	-	-	-	IS*Ppu21*[b]
	86_FR_92	IS*Pa7*[a]	-	-	-	-	-	-	-	IS*Ppu21*[b]
C2	00_ES_10	GCC→ACC (A96T)	-	-	-	-	-	-	-	Insert (GTCCG)1206
	09_BG_97	-	-	-	-	-	-	-	-	-
	47_DE_97	-	-	-	-	-	-	-	-	CAG→TAG (Q415stop)
	68_HU_97	-	-	-	-	-	-	-	-	GAA→TAA (E171stop)
C3	30_AR_05	AAC→AGC (N111S)	-	-	-	-	-	-	-	-
	63_BR_12	-	-	-	-	-	-	-	-	-
	79_AR_10	AAC→AGC (N111S)	-	-	-	-	-	-	-	-
	92_AR_12	AAC→AGC (N111S)	-	-	-	-	-	-	-	Insert C1203
	94_AR_12	AAC→AGC (N111S)	-	-	-	-	-	-	-	CAG→TAG (Q415stop)
C4	06_US_11	GGC→GAC (G84D)	-	-	-	-	-	-	-	AAG→TAG (K296stop)
	22_US_09	GGC→GAC (G84D)	-	-	-	-	-	-	-	ΔT907
	42_US_06	GGC→GAC (G84D)	-	-	-	-	-	-	-	-
	43_US_06	GGC→GAC (G84D)	-	-	-	-	-	-	-	-
	79_US_10	GGC→GAC (G84D)	-	-	-	-	-	-	-	-
C5	02_PH_12	-	-	-	-	-	-	-	-	-
	03_PH_12	-	-	-	-	-	-	-	-	Δ935-952
	05_GR_12	-	-	-	-	-	-	-	-	Δ935-952
	06_GR_12	-	-	-	-	-	-	-	-	-
	08_GR_12	-	-	-	-	-	-	-	-	Δ260-405
	09_RU_12	-	-	-	-	-	-	-	-	Δ260-405
C6	03_CO_11	-	-	-	-	-	-	-	-	Δ450-459
	44_CH_12	-	-	-	-	-	-	-	-	Insert C1016
	BL14_US_un	-	-	-	-	-	-	-	-	-

To be continued

Cluster	Isolate	*ampD*	*ampDh2*	*ampDh3*	*ampR*	*ampG*	*ampO*	*ampP*	*dacB*	*oprD*
	BWH_US_13	-	-	-	-	-	-	-	-	-
	BWHP_US_13	Δ267-283	-	-	-	-	-	-	-	ΔC825
C7	01_US_12	GGG→GAG (G121E)	-	-	-	-	-	-	-	Δ588-603
	23_IT_12	ATC→ACC (I69T)	-	-	-	-	-	-	-	-
	33_RU_13	-	-	-	-	-	-	-	-	ΔT907
	35_RU_13	-	-	-	-	-	-	-	-	ΔT323
	98_ZA_12	CGC→TGC (R176C)	-	-	-	-	-	-	-	-
C8	15D_NG_un	-	-	-	-	-	-	-	-	-
	18_CF_11	-	-	-	-	-	Δ699-1176	-	-	-
	22_TH_12	-	GCC→GTC (A239V)	-	-	-	GCG→GAG (A378E), CCA→GAT (P379D), CCC→CAA (P380Q)	-	Δ1-142, Δ895-1515	TCC→TAG (D100stop)
	26_US_12	-	-	-	-	-	-	GGG→TGG (G410W), GAA→GGA (E413G), TGA→CGA (Stop415R)	-	-
	36D_NG_un	-	-	-	-	-	Δ994-1176	-	-	-
	37_RU_13	-	-	-	-	-	-	-	-	TGG→TGA (W417stop)
	49_NG_13	-	GCC→GTC (A239V)	-	-	-	CCA→GCT (P379A)	-	-	-
C9	19_CO_12	-	-	-	-	-	-	-	-	-
	20_CO_12	-	-	-	-	-	-	-	-	-
	21_CO_12	-	-	-	-	-	-	-	-	-
	BL17_US_un	-	-	-	-	-	-	-	-	Insert A983
	E2UoS_US_un	-	-	-	-	-	-	-	-	-
	JJ692_US_un	-	-	-	-	-	-	-	-	-
	U2504_US_un	-	-	-	-	-	-	-	-	-
C10	23_CH_14	-	-	-	-	-	-	-	-	-
	80_MX_10	-	-	-	-	-	-	-	-	-
	94_RO_12	-	-	-	-	-	-	-	-	CAA→TAA (Q164stop)
	NCGM_JP_02	-	-	-	-	-	GCC→ACC (A329T)	-	-	-

To be continued

Cluster	Isolate	*ampD*	*ampDh2*	*ampDh3*	*ampR*	*ampG*	*ampO*	*ampP*	*dacB*	*oprD*
	P235_ES_08	-	-	-	-	-	-	-	-	-
	WS394_RO_un	-	-	-	-	-	-	-	-	CAA→TAA (Q164stop)
C11	12_HR_08	-	-	-	-	-	-	-	-	-
	1900_JP_12	-	-	-	-	-	-	-	-	TGG→TGA (W138stop)
	1984_JP_12	-	-	-	-	-	-	-	-	TGG→TGA (W138stop)
	30_IT_12	TTC→TCC (F90S)	-	-	-	-	-	-	-	TCC→CCC (S278P)
	68_HR_09	-	-	-	-	-	-	-	-	-
	87_HR_07	-	-	-	-	-	-	-	-	-
C12	06_FR_92	GTC→GGC (V11G)	-	-	-	-	-	-	-	-
	08_CI_08	GTC→GGC (V11G), ΔC48	-	-	-	-	Δ722-1176	-	-	ΔG631, CAG→TAG (Q327stop)
	48_CH_06	-	-	-	-	-	Δ824-1176	-	-	CAG→TAG (Q340stop)
	53_FR_10	GTC→GGC (V11G)	-	-	-	-	-	-	-	TAC→TAA (Y294stop)
	70_FR_92	GTC→GGC (V11G)	-	-	-	-	-	-	-	ΔG389
	76_TR_97	-	-	-	-	-	-	-	-	CAG→TAG (Q340stop)
	87_FR_91	GTC→GGC (V11G)	-	-	-	-	-	-	-	-
	90_RO_12	GGC→AGC (G157S)	-	-	-	-	-	-	-	-
C13	10_US_12	CCG→CTG (P162L)	-	-	-	-	-	-	-	Insert C1206
	15_DE_11	TGT→TAT (C92Y)	-	-	-	-	-	-	-	-
	32_US_12	-	-	-	-	-	-	-	-	-
	74_US_07	-	-	-	-	-	-	-	-	ΔG179
	SPae_FR_04	-	-	-	-	-	-	-	-	TAC→TAG (Y294stop)
C14	02_AR_08	-	-	-	-	-	-	-	-	Δ1136-1145
	49_AR_09	-	-	-	-	-	-	-	-	Δ1136-1145
	58_FR_12	-	-	-	GAC→AAC (D135N)	-	-	-	-	GCC→TCC (A8S)

To be continued

Cluster	Isolate	*ampD*	*ampDh2*	*ampDh3*	*ampR*	*ampG*	*ampO*	*ampP*	*dacB*	*oprD*
	6077_US_un	ACG→GCG (T139A)	-	-	-	-	-	-	-	-
	78_FR_88	-	-	-	-	-	-	-	-	-
	89_MX_12	-	-	-	-	GCG→GTC (A586V)	-	-	-	-
	BL08_US_un	-	-	-	-	-	-	-	-	-
End of table										

(a) IS*Pa7* was 473 bp and was inserted between nucleotides 71 and 75 of the coding region of the *ampD* gene.

(b) IS*Ppu21* was 475 bp and was inserted between nucleotides 1137 and 1141 of the coding region of the *oprD* gene.

(c) IS*Pa7* was 474 bp and was inserted between nucleotides 71 and 75 of the coding region of the *ampD* gene.

All insertion sequences were identified using ISfinder database (Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 2006;34:D32-6.).

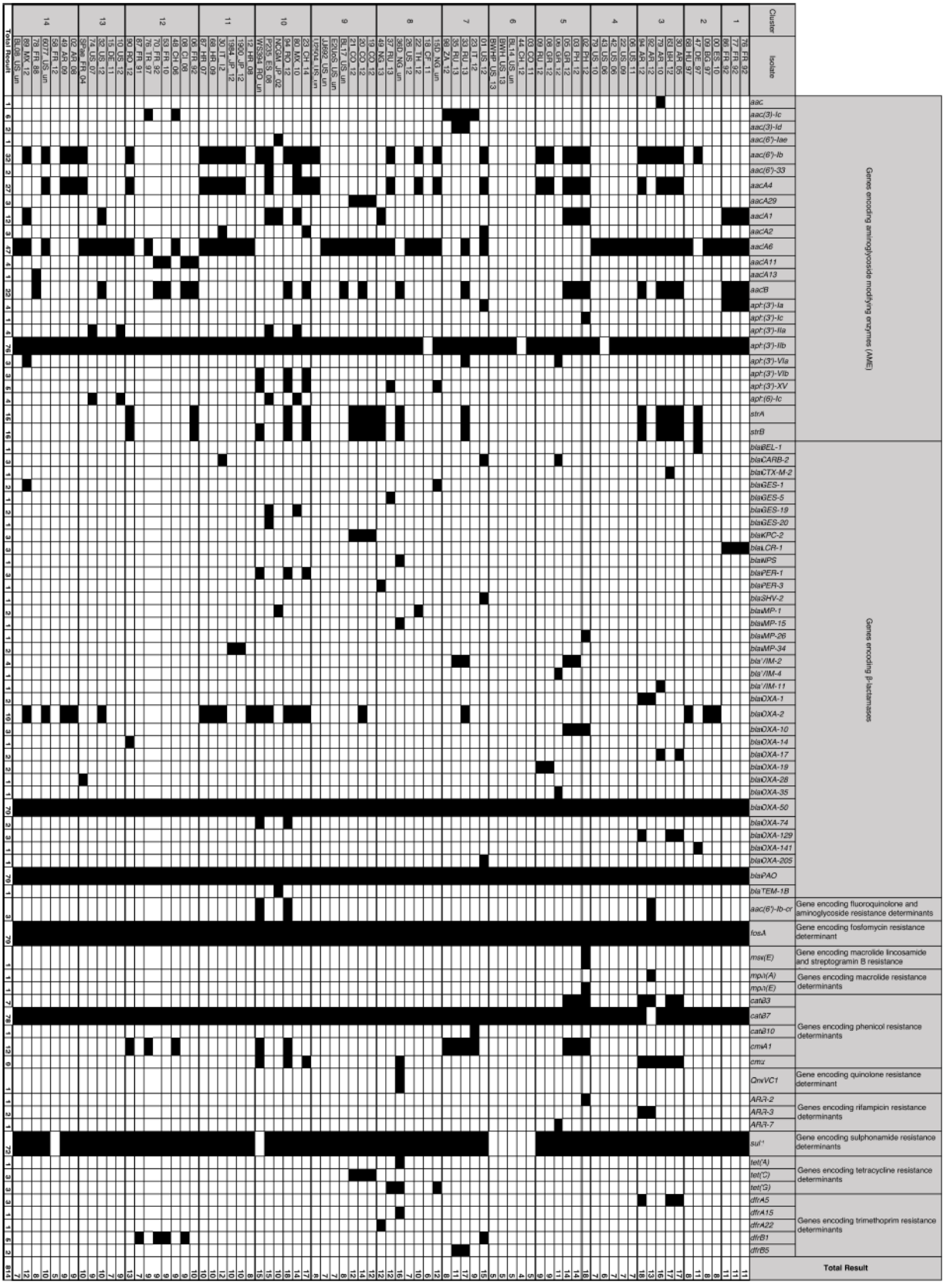


Figure 3.8: **Resistance genes acquired by the 79 *P. aeruginosa* ST235 isolates.**

4

PanISa: a new tool to find insertion sequences on NGS data

4.1/ Introduction

We already have stated, in a previous chapter, that an Insertion Sequence (IS) is a small DNA fragment encoding transposase (Tpase) gene, that has the ability to move within cells and to repeat itself thanks to new insertions. It can be integrated by either plasmids or bacteriophages, to move between genomes as an horizontal gene transfer [Siguier et al., 2014].

The common IS structure consists of 1-2 Tpase open reading frame(s) (orf), with terminal inverted repeats (IRs) for Tpase-mediated cleavage recognition and Tpase binding sites. Direct repeated sequences (DRs), which are short target DNA flanking ISs, are always taken into account when referring to the IS organisation [Siguier et al., 2015]. According to their diversity of organization, their mechanisms and features, insertion sequences are classified within families using the 4 attributes below (see, *e.g.*, [Mahillon and Chandler, 1998]):

1. arrangement of orf;
2. transposition chemistry;
3. features of their ending terminal inverted repeats; and
4. their destination sequences.

At the beginning of their study, 17 families have been found from 443 insertion sequences, and this number has then been increased to reach 29 families using more than 4,000 ISs in ISfinder database previously introduced [Siguier et al., 2015, Mahillon and Chandler, 1998].

Insertion sequences have various impacts on host genomes, like their reduction in size due to IS expansions that may cause mutation, deletion, and rearrangement in the host sequence [Siguier et al., 2014]. In addition, insertion sequences have some effects on genome expression either due to an inactivation caused by an

insertion within a given target gene, or conversely an activation by an upstream insertion.

The repercussions are felt by important phenotype modifications, in terms of virulence, antibiotic resistance, and metabolism [Vandecraen et al., 2017]. For instance, an increasing of fluoroquinolones resistance caused by IS*1* or IS*10* insertion upstream of *acrEF* has been reported in *Salmonella enterica* [Olliver et al., 2005]. Similarly, inactivation of the gene *oprD* by IS insertion induced imipenem-resistance in *Pseudomonas aeruginosa* clinical strain is evoked in [Sun et al., 2016]. Due to this significant role of ISs in the evolution of host microbial genomes, the detection of such sequences is important in evolutionary studies: this detection may help to reveal, for instance, how antibiotic resistance occurs, or the widespread distribution of such resistance genes.

Next-generation sequencing of whole bacterial genomes has become a standard in the analysis of genomics evolution and structure, allowing for instance the detection of mutations by aligning reads against a genome reference. However, IS analysis is more complicated due to the repeated nature of such elements and the limited size of read sequences. Several tools have been developed to overcome this problem [Ewing, 2015]. Some of them require an IS database as template recognition for identifying IS, like ISMapper [Hawkey et al., 2015] and RetroSeq [Keane et al., 2012]: they take benefits from manual feature or function confirmations in the case of known ISs. However, there is no denial in the fact that many unknown ISs are still waiting to be discovered. This is why some detection tools, like DD_DETECTION [Kroon et al., 2015] or Mobster [Thung et al., 2014], do not require any IS database. But, unfortunately, they are only suitable for human genomes.

Hence, to fulfill the study of insertion sequences in bacteria with the ability to discover new kind of ISs, we have designed *panISa*, which is a sensitive and precise ISs detection tool. It has been developed to be easy to use and to necessitate only read mapping sequences (BAM) as input: in other words, no IS library is needed.

The reason to be of panISa has appeared to us during a real bioinformatics investigation, which is reported in the following section.

4.2/ A REAL CASE STUDY AT THE ORIGIN OF PANISA DESIGN

4.2.1/ A FIRST PIPELINE FOR IS STUDY OF ST233 STRAINS

4.2.1.1/ WGS OF *P. aeruginosa* STRAINS

Five strains of ST233 *P. aeruginosa*, isolated from Oct. 2007 to May 2008, have been collected in the Besançon hospital. After DNA extraction, strains were se-

quenced using Illumina HiSeq, with 2x150bp and subsampling of 80x, and a random selection of paired-end reads.

The first isolate was assembled using Ray software set with default values for Illumina data [Boisvert et al., 2010], leading to 118 contigs for a total of 6,997,480 bp. This genome was secondly annotated using Prodigal [Hyatt et al., 2010] (5,086 genes) and homologous proteins were searched with blastP against the proteome of *P. aeruginosa* (Uniprot, 50,812 entries). The 4 next WGS data were finally aligned against the first one using bwa-mem [Li and Durbin, 2009].

4.2.1.2/ GENOME COMPARISON

Genome comparison was performed by variant calling using freebayes [Garrison and Marth, 2012]. This analysis allowed to find 437 mutations (SNP/INDEL) in all strains. Using PAO1 genome as a reference, a phylogenetic analysis was performed using MrBayes [Huelsenbeck and Ronquist, 2001] with a GTR+G+I model on a multiple alignment of 516,277 bp. Mutation rate of each strain was measured experimentally [Oliver et al., 2000].

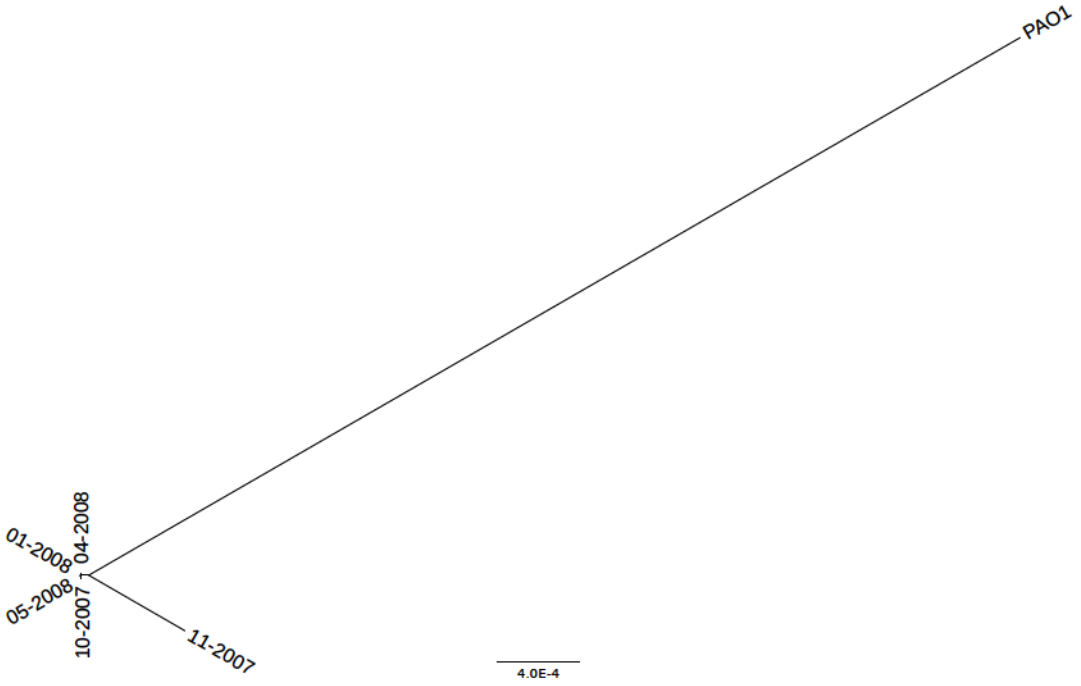


Figure 4.1: **Phylogeny of ST233 *P. aeruginosa*.**

4.2.2/ A CRITICAL ISSUE IN OUR NGS PIPELINE

A large number of mutations has occurred during the spread of the ST233 epidemic clone between October 2007 and May 2008. They are mainly concentrated

in the 11-2007 strain, with about 300 mutations, while they are lower than ten mutations for the other strains (see Figure 4.1). These mutations are widely spread along the genome, which is not compatible with a few horizontal transfer events.

The 11-2007 strain seems to have a high mutation rate (hyper-mutator) that was confirmed experimentally with a value of $2.4E^{-6}$, which should be compared to the other strains, that have a more normal mutation rate ranging from $2E^{-8}$ to $5E^{-9}$. However, no mutation was discovered in the *mutS* and *mutL* DNA reparation genes, when applying a classical NGS analysis, which is surprising in view of the mutation rate of the considered strain. To remove such a contradiction, we assembled the genome of the 11-2007 strain. An investigation of the aforementioned genes revealed that *mutS* is indeed split (cut) in two contigs, with partial sequences at the end of each contig. A sequence alignment with IS*Pa1635* (AY539834) demonstrated that this *mutS* gene was inactivated by an IS insertion.

To address this IS insertion was a difficult task, and phenotype has oriented our detection. We searched for tools that are able to find ISs on our particular kind of data, but we found that none of the available software are dedicated to bacterial insertion sequences on aligned data of draft genomes (*i.e.*, with no IS annotation), thus without knowledge of IS like in Breseq [Barrick et al., 2014] or ISMapper [Hawkey et al., 2015]. This lack has led us to design the PanISa software, which is described below.

4.3/ Our PanISa detection tool: design and evaluation

4.3.1/ Software generalities

PanISa program is a python script that parses a read mapping file (sam or bam) using the pysam library [Heger and contributors, 2009]. Additionally, to detect potential inverted repeat regions, we used the einverted executable program from the well-known EMBOSS package [Rice et al., 2000].

PanISa script is developed under the GPL v3 license, and sources are available at Github (https://github.com/bvalot/panISa).

4.3.2/ Implementation

Looking at alignment data, we discovered a particular pattern at insertion sites: a large number of clip reads (reads that are only partially mapped) ending at one side of the direct repeat region. Most of existing tools uses discordant mate-paired reads combined with either genome annotations of repeat regions or a list of know IS [Ewing, 2015], while our strategy is different, by using only clip reads to find potential IS insertions, that are confirmed in a second stage by reconstructing

Table 4.1: **Organisms considered during simulations.**

Organism	Name	Strain	Accession no.	Gram	Number of experiment IS
Escherichia coli	*Escherichia coli str. K-12 substr. MG1655*	K12	NC_000913	Negative	22
Mycobacterium tuberculosis	*Mycobacterium tuberculosis H37Rv*	H37Rv	NC_000962	Positive	3
Pseudomonas aeruginosa	*Pseudomonas aeruginosa PAO1*	PAO1	NC_002516	Negative	25
Staphylococcus aureus	*Staphylococcus aureus subsp. aureus NCTC 8325*	NCTC8325	NC_007795	Positive	8
Vibrio cholerae	*Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I*	N16961 chromosome I	NC_002505	Negative	9
Vibrio cholerae	*Vibrio cholerae O1 biovar El Tor str. N16961 chromosome II*	N16961 chromosome II	NC_002506	Negative	9

the borders of IS sequence.

As shown in Figure 4.2, panISa program filtered reads in order to only select clip reads, and it grouped them by position and side (left or right) of the clip. If two positions with sufficient number of clip reads in opposite side are close enough, we created a ”potential” IS. From the clip read, panISa program then created consensus sequences of the direct repeat, and of the left and right ends of the inserted sequences. The latter were searched for potential inverted repeats. All potential IS insertions were reported on a text file (tabular format). In addition, Figure 4.3 illustrates the pseudocode of panISa including all cutoff values as the optimal parameters for fitting with real data.

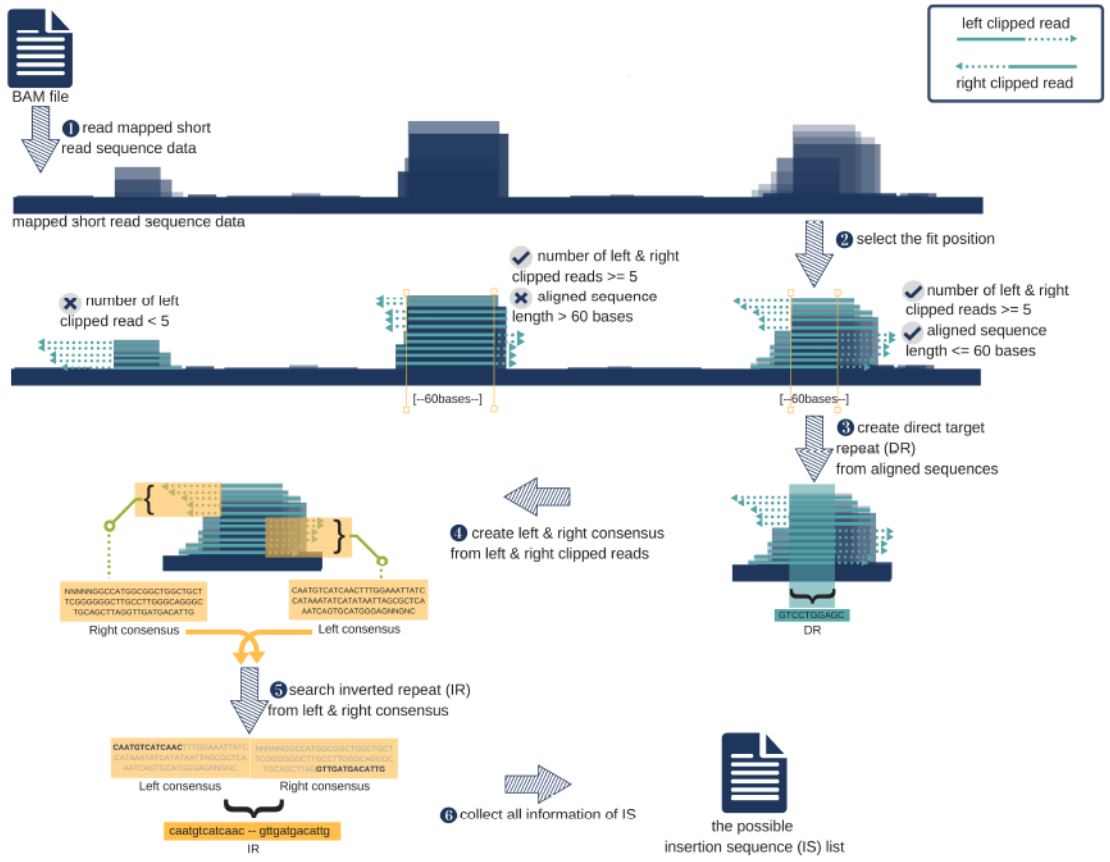


Figure 4.2: **Schema of panISa workflow.**

By nature, all types of insertion or rearrangement events could be detected by our software, including other transposable elements (such as bacteriophage, ICE, and so on). Thus, the list of potential insertion sequences must be manually inspected and validated to conclude to an IS insertion. Most of ISs [Siguier et al., 2014] have an inserted repeat and the detection by panISa of it is a good indication of potential IS. Moreover, reconstruction of left and right borders allows an alignment against the ISFinder database, leading to a potential homology validation to a referenced IS [Siguier et al., 2006a].

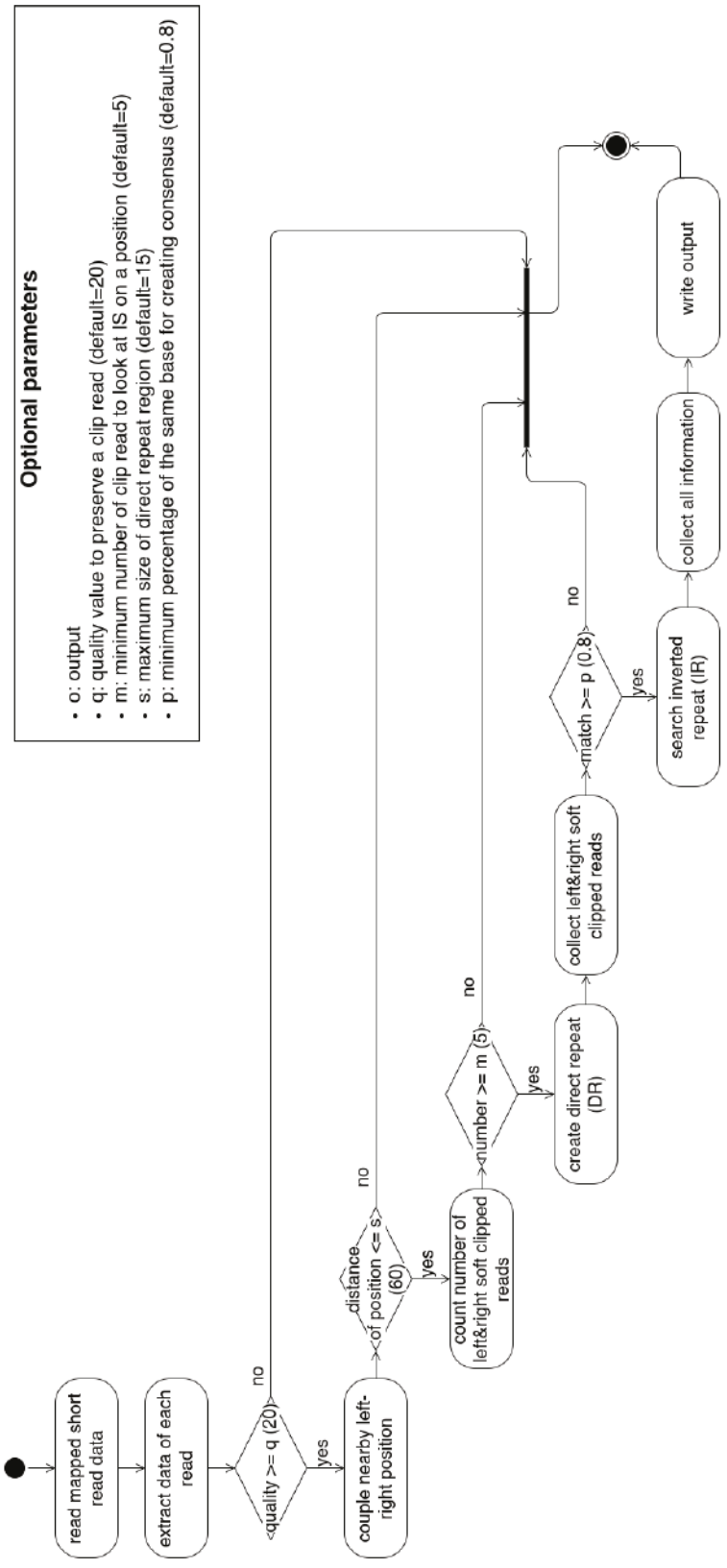


Figure 4.3: **The activity diagram of panISa.**

4.3.3/ EVALUATION PROTOCOL

4.3.3.1/ REFERENCE GENOME AND IS ELEMENT

We have evaluated our panISa proposal on a set of artificially inserted ISs in given bacterial genomes. After such manual insertions, we are then left to measure to which extend our tool is able to recover them.

To simulate IS element insertions, we have used six reference genome sequences of 5 organisms, namely of *Escherichia coli*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Vibrio cholerae*. Sequence data have been obtained from the NCBI database, see Table 4.1 for further references.

Table 4.2: **ISs information in the simulation.**

Organism	IS Name	IS Family	IS Group	DR length	IS length	IR
Escherichia coli	ISEc1	ISAs1	-	8-10	1291	Y
	ISEc10	IS21	-	4-8	2410	Y
	ISEc17	IS3	IS3	3-4	1258	Y
	ISEc18	IS481	-	4-15	1020	Y
	ISEc20	IS110	-	0	1459	N
	ISEc23	IS66	-	8-9	2532	Y
	ISEc29	IS4	IS10	9	1325	Y
	ISEc30	IS1	-	8-9	766	Y
	ISEc35	IS5	IS903	9	1051	Y
	ISEc38	ISL3	-	8	1722	Y
	ISEc41	IS200/IS605	IS605	0	1841	N
	ISEc42	IS200/IS605	IS1341	0	1291	N
	ISEc51	ISKra4	ISAzba1	0	2857	Y
	ISEc58	IS256	-	8-9	1342	Y
	ISEc61	IS5	IS1031	3	954	Y
	ISEc66	IS110	IS1111	0	1441	Y
	ISEc68	IS5	IS5	4	1197	Y
	ISEc70	IS1595	IS1595	8	1018	Y
	ISEc75	IS1595	IS1016	7-9	717	Y
	ISEc9	IS1380	-	4-5	1656	Y
	ISEcB1	IS3	IS150	3-4	1441	Y
	TnEc1	Tn3	-	0	3854	Y
Mycobacterium tuberculosis	ISMt1	IS5	IS427	2-4	969	Y
	ISMt2	IS21	-	4-8	2645	Y
	ISMt3	IS21	-	4-8	2213	Y
Pseudomonas aeruginosa	ISPa12	IS4	ISH8	10	1387	Y
	ISPa1328	IS256	-	8-9	1328	Y
	ISPa14	IS1	-	8-9	788	Y

To be continued

Organism	IS Name	IS Family	IS Group	DR length	IS length	IR
Pseudomonas aeruginosa	ISPa18	IS200/IS605	IS1341	0	1339	N
	ISPa20	IS3	IS3	3-4	1246	Y
	ISPa27	IS256	-	8-9	1361	Y
	ISPa30	IS66	-	8-9	2436	Y
	ISPa33	IS1380	-	4-5	1561	Y
	ISPa35	IS5	IS903	9	1031	Y
	ISPa38	Tn3	-	0	6455	Y
	ISPa39	IS3	IS407	4	1254	Y
	ISPa40	Tn3	-	0	6592	Y
	ISPa45	IS4	IS4	10-13	1637	Y
	ISPa47	IS630	-	2	1107	Y
	ISPa56	IS3	IS2	5	1359	Y
	ISPa59	IS30	-	2-3	1113	Y
	ISPa60	ISAs1	-	8-10	1228	Y
	ISPa61	ISL3	-	8	1219	Y
	ISPa62	IS110	IS1111	0	1382	Y
	ISPa65	IS21		4-8	2520	Y
	ISPa67	IS5	IS427	2-4	852	Y
	ISPa7	IS1182	-	0-60	1669	Y
	ISPa72	IS1595	IS1595	8	1048	Y
	ISPa8	IS5	IS5	4	1324	Y
	ISPa9	IS5	IS5	4	1207	Y
Staphylococcus aureus	ISSau1	IS30	-	2 - 3	1070	Y
	ISSau2	IS3	IS150	3-4	1660	Y
	ISSau3	IS1182	-	0-60	1946	Y
	ISSau4	IS3	IS150	3-4	1261	Y
	ISSau5	IS30	-	2-3	1136	Y
	ISSau6	IS6	-	8	793	Y
	ISSau8	ISL3	-	8	1498	Y
	ISSau9	IS21	-	4-8	2446	Y
Vibrio cholerae	ISAlg	IS3	IS3	3-4	1258	Y
	ISVch1	IS481	-	4-15	1023	Y
	ISVch3	IS21	-	4-8	2593	Y
	ISVch4	IS3	IS3	3-4	1258	Y
	ISVch5	IS5	IS5	4	1197	Y
	ISVch6	IS630	-	2	1088	Y
	ISVch7	IS66	ISBst12	8-9	1588	Y
	ISVch8	IS5	IS903	9	1053	Y
	ISVch9	IS5	IS5	4	1195	Y
End of table						

The IS families that are known to be present within these organisms are listed in Table 4.2. They have been downloaded from the ISfinder database [Siguier et al., 2006b].

4.3.3.2/ SIMULATION DATA

To evaluated panISa program, we have artificially inserted ISs on the reference genome of each species. For each simulation, 30 ISs were randomly picked from our list (Table 4.2), and inserted randomly in the genome with a direct repeat length in agreement with the considered IS family. From each simulated genome, Illumina short reads were created using dwgsim v.0.1.11-3 [Homer, 2014] and aligned against the wild type genome using bwa-mem [Li and Durbin, 2009]. IS detection were performed with our program on the aligned reads and compared to the recorded IS insertion. All described steps are drawn in Figure 4.4

To evaluate the impact of reads data, we performed simulations for each genome with three read lengths (100bp, 150bp, and 300bp), and five different coverage (20x, 40x, 60x, 80x, and 100x). Each step were repeated ten times leading to 27,000 simulated IS insertions when combining all possible genomes, read lengths, and read coverages.

4.3.4/ VALIDATION ON SIMULATED DATA

IS detection As can be seen in Figure 4.5, the sensitivity of IS detection at various read lengths and coverage in all strains indicates some tendencies related to the coverage. According to obtained statistical results, there is no significant differences related to read length variation, while we obtained significant differences when regarding species or coverage (see Table 4.3).

Table 4.3: **The summary results of ANOVA test.**

	Factor	Df	F-value	p-value	Sig.level
Sensitivity	coverage	4	16.796	$1.84e^{-09}$	***
	read length	2	0.423	0.65708	
	species	4	4.808	0.00187	**
Precision	coverage	4	11.545	$4.07e^{-07}$	***
	read length	2	22.348	$4.35e^{-08}$	***
	species	4	3.331	0.0153	*

Our sensitivity is thus assumed to be independent from the chosen sequencing technology of Illumina, as we obtained no significant differences when focusing on read length. Conversely, in the case of species, small differences have been found, and then the Tukey's test shown only that *P. aeruginosa* is much detected than *M. tuberculosis*, corresponding to the two extremes. When considering the various possible coverage rates, we found that 20x is significantly lower from the other ones. This coverage is then too low to have sufficient sensitivity.

When consider all simulations between 40x to 100x coverage, the sensitivity reaches a mean of 98%, with a 95%CI equal to [97.9%-98.2%]. PanISa software is then very sensitive.

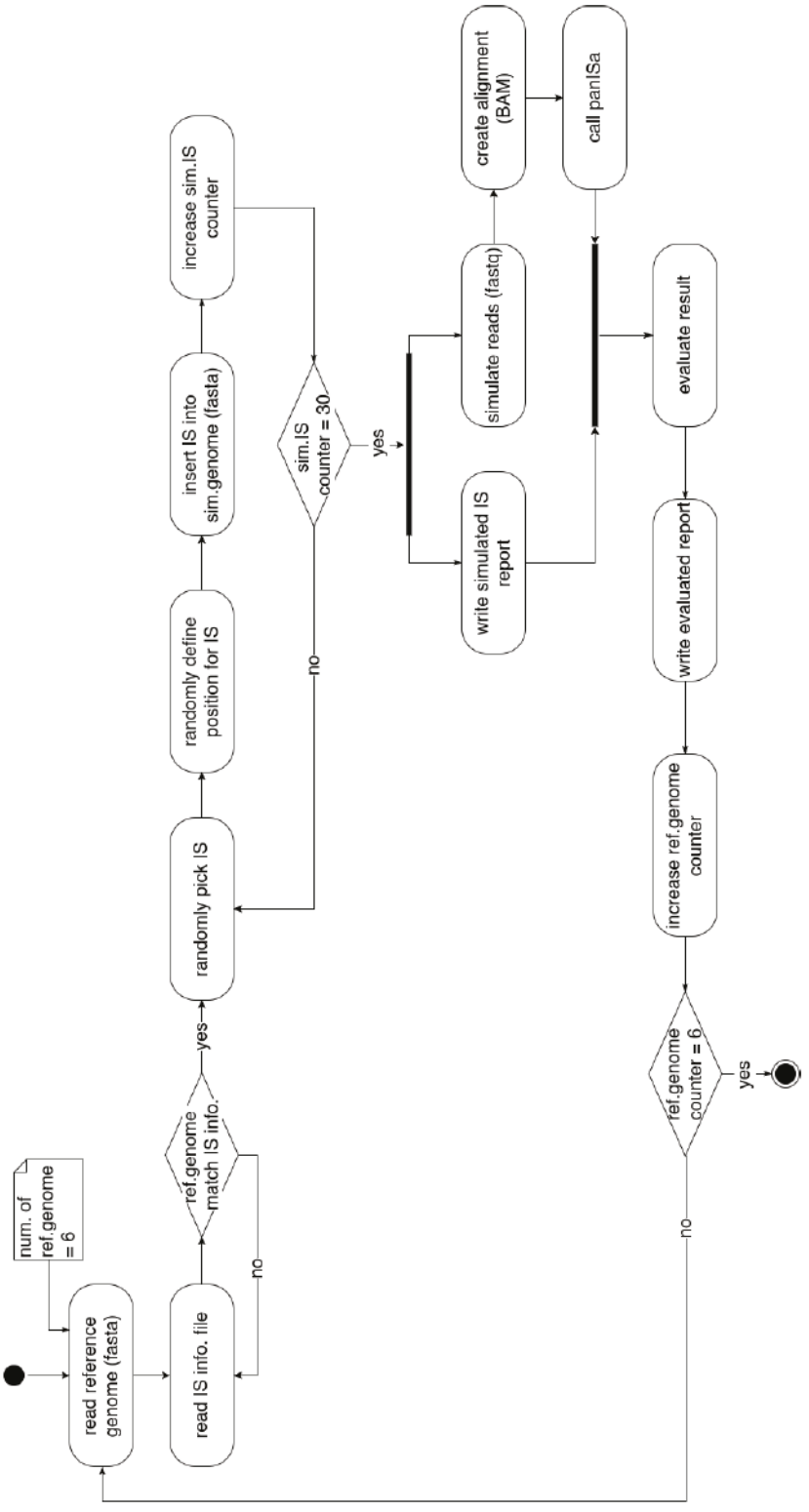


Figure 4.4: **The activity diagram of evaluation on simulated data.**

The precision of IS detection (see Figure 4.6) is very significantly different between read coverage and read length (Table 4.3). According to Tukey results, 100bp read shows a significantly lower precision in comparison to 150bp or

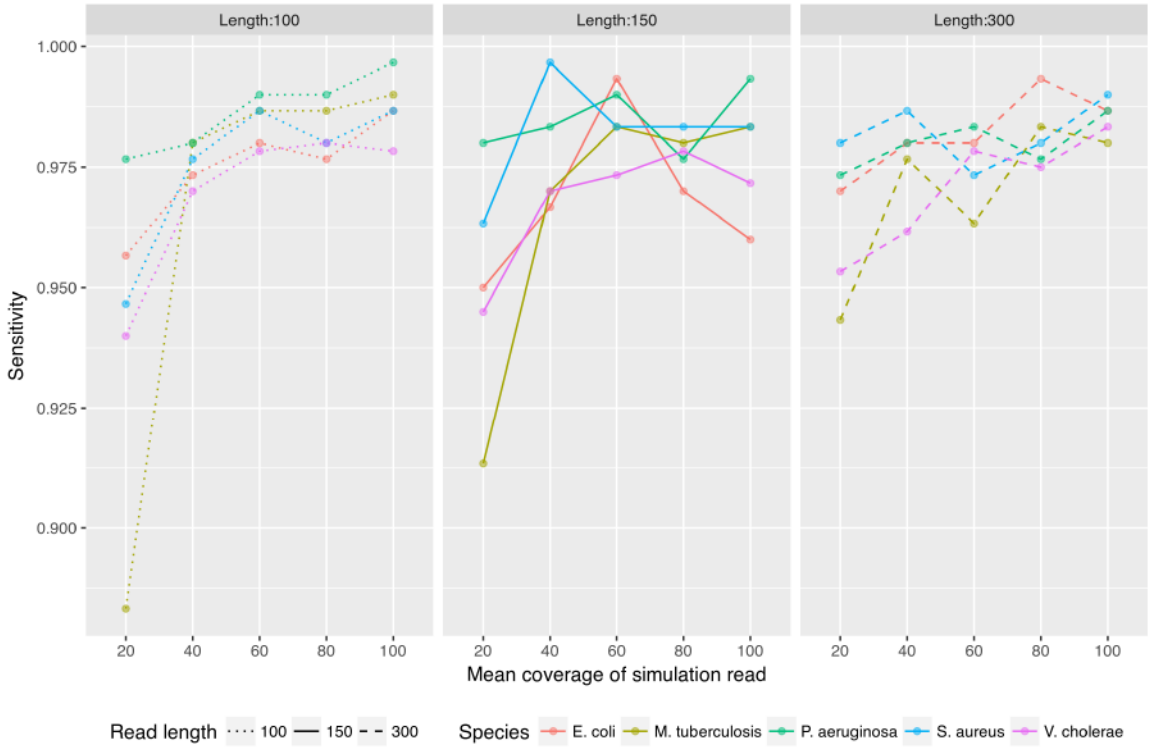


Figure 4.5: **Sensitivity of the proposal.**

300bp. This precision is also affected by a too high coverage (100x and 80x). These precision values are indeed very large in average (larger than 95%), which is very encouraging. Most of false positive positions found correspond to small repeat regions. By increasing the coverage, we obtained more low quality clipped read on these regions up to the threshold. This problem could be overcome by sub-sampling high coverage sequencing data or by increasing the minimum number of clipped read threshold.

In conclusion, the optimum read coverage is around 40x-60x, as they present stable positive values in the two simulation graphs, which should be high for both sensitivity and precision. Similarly, a minimum of 150bp length is necessary to a limited loss of accuracy.

Direct and inverted repeats detection Two other information are provided by panISa, namely: (1) the position and sequence of the direct repeat, and (2) the presence of an inverted repeat between the left and right sequence of the IS.

By collecting all simulation results, we can claim that a correct DR position and length is observed in only 56.8% of the cases (95%CI: [55.5%-57.5%]). Note that incorrect reports correspond in majority (at 99.5%) to a bigger length of 1 to 5 bp. This element is related to the location where the IS is inserted. If the beginning of the IS corresponds to the genome sequence, the read mapping overlaps a region

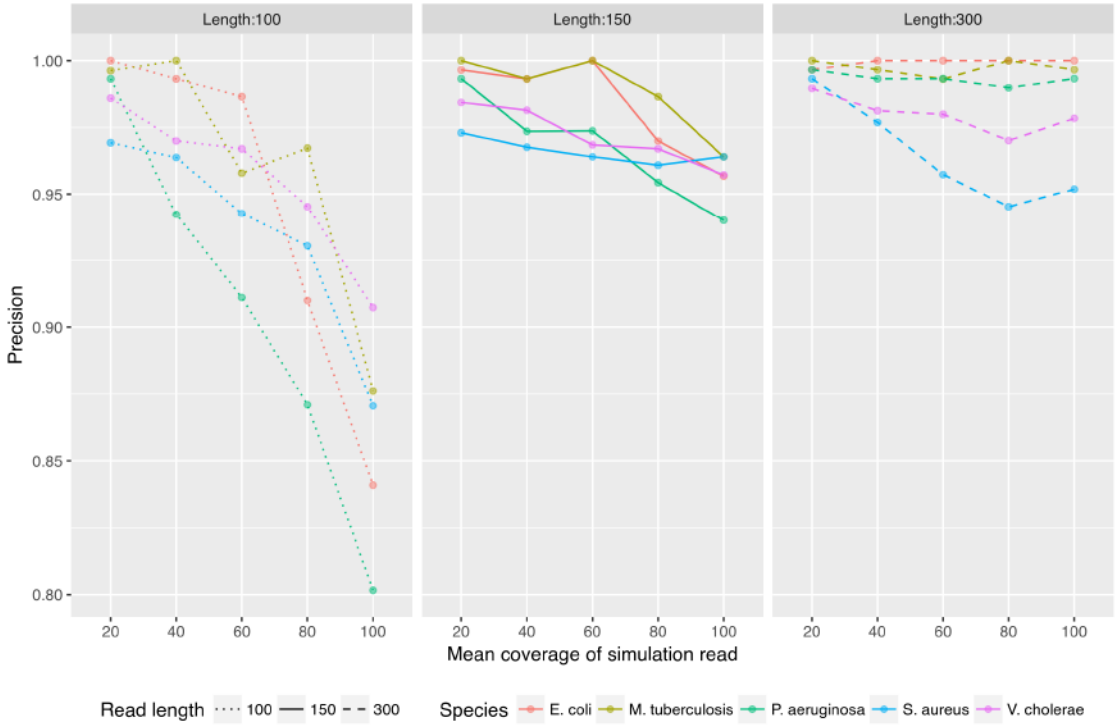


Figure 4.6: **Precision of the proposal.**

larger than the original DR and then the clip position is shifted.

Concerning the detection of an IR, the sensitivity has a value of 74% (95%CI: [71.2%-76.8%]) and the precision a value of 99.9% (95%CI: [99.87%-100%]). This low sensitivity value of IR detection is directly related to which IS is inserted. For instance, the IR of IS*Vch7* and IS*Pa61* were never detected because their IRs are too much divergent.

4.4/ SCREENING ISS IN *P. aeruginosa* STRAINS USING PANISA

PanISa were run with its default parameters on each aligned ST233 strain. Left and right sequences of potential ISs were clustered using sumaclust [Mercier et al., 2013] with an identity cutoff set to 95%. Each representative has then been searched for homology in the ISFinder database [Siguier et al., 2015, Mahillon and Chandler, 1998], and complete IS sequences were amplified by PCR and then sequenced (see Table 4.4).

Table 4.4: **Primer list used for IS amplification.**

Homologous IS	Forward	Reverse
IS*Pa1635*	TGTAGGTGGATTTACCGCCC	GCTACTTCATCGAGCTGCCA
IS*Pa45*	AGGAAATCCGCCTGGAACTG	CTTTTCGTTGGCACGCTACC
IS*UnCu3*/IS*Ppu7*	GTGCGCGAATCCGAAAATGA	CGTAAAAAGGCGCGACACAA
IS*Pst3*	CCTGTCCCGGTGTCTTTAGG	TTACCTATCAAGGCGCACCA
IS*Ppu7* partial	TGACGAATGCAGGAGAGGTG	TACGCGATGCTGCTGATACC
IS*Pst5*	GGTAGATGGACGGTTCCCAG	TAGCGCTATGGCCTTGGTTG

We found 46 possible insertions during the spread. After clustering of left and right insertion sequences, only 8 different sequences were found. After validation with alignments against the ISFinder database, 6 sequences shown a high homology to a know IS, corresponding to the 44 insertions that are listed in Table 4.5. The two possible inserted sequences that have been removed correspond to a small insertion of 18bp on the one hand, and a false positive detection due to a palindromic region on the other hand.

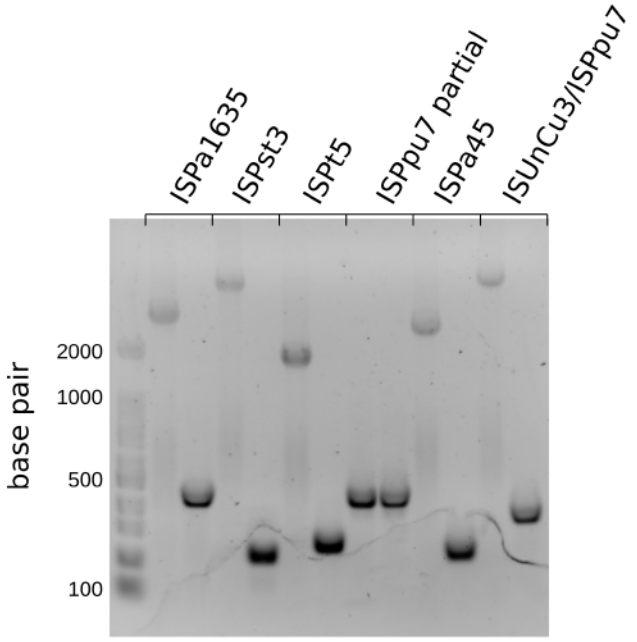


Figure 4.7: **Validation of IS insertion by PCR.**

To validate the insertions *in vivo*, we designed primers surrounding the positions detected by our software and we amplified them using PCR. Considering the 10-2007 strain as control, five of the six ISs shown a band with an increase of 1500 to 2500 bp (see Figure 4.7). One amplification with similar size (ISPpu7 partial) indicated a false positive detection. A sequencing of these amplifications has finally validated the insertion sequences. In conclusion, IS insertions have been confirmed for 43 positions in the 46 ones detected by our software, resulting on a precision of 93.5%, a somewhat lower value than obtained in simulation.

A search of validated ISs in the first strain shown that these ISs were already present at the beginning of the epidemic. During the spread, the ISs were 1 to 3 times copied in new locations, except IS*Pa1635* that has 36 new insertions (*cf.* Table 4.5). Thirty-four were present in the hypermutator strain 11-2007. This result seems to indicate that the loose of DNA repair mechanism has induced a proliferation cycle of IS*Pa1635* in this strain [Siguier et al., 2014].

Table 4.5: **Detected of new IS insertion on *P. aeruginosa* strain using panISa.**

Homologous IS	Validated by PCR	Present in 10-2007	11-2007	01-2008	04-2008	05-2008
IS*Pa1635*	Yes	Yes	34			2
IS*Pa45*	Yes	Yes				1
IS*UnCu3*/ IS*Ppu7*	Yes	Yes		2		1
IS*Pst3*	Yes	Yes		2		
IS*Ppu7* partial	No	NA		1		
IS*Pst5*	Yes	Yes		1		

4.5/ CONCLUSION

Due to an obvious issue appeared in a real case study, we developed a tool called panISa to search new IS insertions on re-sequencing data. This tool has shown to be able to detect ISs on simulated data, with a good sensibility and precision.

Its application on a epidemic clone of *P. aeruginosa* has allowed to identify 5 different types of ISs with 43 insertions on 4 strains. Three of them correspond to newly described ISs in *Pseudominas aeruginosa*. One strain has an hypermutator phenotype due to an IS*Pa1635* insertion on the *mutS* gene, and this has induced an expansion phase (34 new insertions) of this IS.

Reanalysis of published sequencing dataset of clonal epidemic strains with PanISa could reveal some missing resistance or virulence variations due to IS insertions. We intend to investigate systematically this possibility in a further work.

III

CONCLUSION

5

CONCLUSION AND FUTURE WORK

5.1/ CONCLUSION

Pseudomonas aeruginosa is a major nosocomial pathogen with ST235 being the most prevalent of the so-called 'international' or 'high-risk' clones. This clone is associated with poor clinical outcomes in part due to multi- and high-level antibiotic resistance. Despite its clinical importance, the molecular basis for the success of the ST235 clone is poorly understood. Thus this thesis aimed to understand the origin of ST235 and the molecular basis for its success, including the design of bioinformatics tools for finding insertion sequences (IS) of bacterial genomes.

To fulfill these objectives, this thesis was divided into 2 parts.

First, the genomes of 79 *P. aeruginosa* ST235 isolates collected worldwide over a 27-year period were examined. A phylogenetic network was built using Hamming distance-based method, namely the NeighborNet. Then we have found the Time to the Most Recent Common Ancestor (TMRCA) by applying a Bayesian approach. Additionally, we have identified antibiotic resistance determinants, CRISPR-Cas systems, and ST235-specific genes profiles. The results suggested that the ST235 sublineage emerged in Europe around 1984, coinciding with the introduction of fluoroquinolones as an antipseudomonal treatment. The ST235 sublineage seemingly spreads from Europe via two independent clones. ST235 isolates then appeared to acquire resistance determinants to aminoglycosides, β-lactams, and carbapenems locally. Additionally, all the ST235 genomes contained the exoU-encoded exotoxin and identified 22 ST235-specific genes clustering in blocks and implicated in transmembrane efflux, DNA processing and bacterial transformation. These unique genes may have contributed to the poor outcome associated with *P. aeruginosa* ST235 infections and increased the ability of this international clone to acquire mobile resistance elements.

The second part was to design a new Insertion Sequence (IS) searching tool on next-generation sequencing data, named panISa. This tool identifies the IS position, direct target repeats (DR) and inverted repeats (IR) from short read data (.bam/.sam) by investigating only the reference genome (without any IS database). To validate our proposal, we used simulated reads from 5 major bac-

terial pathogenic species: *Escherichia coli*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Vibrio cholerae*. The experiment set is constituted by reads of various lengths (100, 150, and 300 nucleotides) and coverage of simulated reads at 20x, 40x, 60x, 80x, and 100x. We performed sensitivity and precision analyses to evaluate panISa and found that the sensitivity of IS position is not significantly different when the read length is changed, while the modifications become significant depending on species and read coverage.

When focusing on the different read coverage, we found a significant difference only at 20x. In the other situations (40x-100x) we obtained a very good mean of sensitivity, which is equal to 98% (95%CI: [97.9%-98.2%]). Similarly, the mean of precision for IS position is larger than 95%, which is also a good score. In terms of IR evaluation, although the mean of sensitivity should be improved (74%), the mean of precision is extremely good (99.9%). In case of DR assessment, the accuracy is only equal to 56.8%, which is quite low since the majority of errors contains a larger size of approximately 1-5 bp. However, panISa has been validated on real data of *Pseudomonas aeruginosa* ST233 and also confirmed by PCR. And the result of IS detection, which is equal to 93.5%, is really satisfactory.

In conclusion, *P. aeruginosa* ST235 (i) has become prevalent across the globe potentially due to the selective pressure of fluoroquinolones and (ii) readily became resistant to aminoglycosides, β-lactams, and carbapenems through mutation and acquisition of resistance elements among local populations. Concerning the second point, our panISa proposal is a sensitive and highly precise tool for identifying ISs from short reads of bacterial data, which will be useful to study the epidemiology and bacterial evolution.

5.2/ FUTURE WORK

WGS pipeline analysis is a powerful tool to decipher bacterial epidemiology at a global or a local scale. Having such tools on hand, we will identify the factors (antibiotic prescription, lack of hygiene, transmission vector, human behavior, bacterial virulence factors) that could have favored the spread of bacterial pathogens. It would also be interesting to compare the geographical progress of different pathogens over time, and the occurrence of their multi- and high-level antibiotic resistance. We will wonder, at each time, if the success of each clone has a molecular basis, or if other explanations must be found.

We have already pointed out the limitation of existing data for the global clone ST235 analysis, and that these data are biased in their geographical and temporal distributions. For instance, we have only a few number of strains from Asia and Africa, and conversely a large number of sequences from North America and Europe. And, in particular, the majority of early isolates are from Europe which may influence the TMRCA estimation. Thus, in future works, we will expand the analysis to other high-risk clones of *Pseudomonas aeruginosa* (*e.g.* ST111,

ST175, ST244, and ST395). As in the ST235 case, we will collect new genomic data representing the worldwide diversity of each of these clones, sequencing *de novo* new genomes if needed. Then, after having located and dated well each sequence of each type, we will perform a similar study than in Chapter 3, encompassing a phylogenetic analysis and a time to most recent common ancestor evaluation, in order to obtain, for each strain, the most supported evolutionary scenario.

This study may be applied, *mutatis mutandis*, to other kinds of bacteria with their high-risk clones, in which some strains have presented a similar way of recent emergence and global widespread coupled with multi- and high-level antibiotic resistance. We will investigate in particular whether a generic pipeline may be of interest, that will receive a collection of sequences of a given ST, with both location and date, and that will provide first a phylegenetic study, and then various possible molecular and environmental scenario explaining the clone evolution with its acquisition of resistance genes.

The expected outputs are a better characterization of the spread of bacterial pathogens at different scale (from a hospital-scale to worldwide) to inform public health authorities about interventions that should be prioritized to control the emergence, the transmission and the diffusion of these organisms.

According to the results of ST235-specific determinants, we should do validation by focusing on the functionality of the transformation pathway and also perform phenotype test.

Concerning our second contribution, an important issue should be pointed out. We have emphasized the ability of panISa to reach a good performance for close mapping input, or for any reads that have been mapped with close reference strains. Conversely, if aligned reads have been mapped with distant reference strains, the result from panISa will contain other mutations than IS regions. Although this issue is quite complicated to solve, panISa can be improved by the following future works.

First of all, we want to improve the inverted repeat sensitivity, whose mean must be at least equal to 95%. Other indicators of good performance like the sensitivity of DR identification must be improved too if possible, and an intensive comparison with state-of-the-art tools must be performed on a larger set of well annotated genomes, to place panISa among the most accurate tools for insertion sequence detection.

Secondly, technical improvements are possible too, to make panISa a widely used state-of-the-art detection tool. We first intend to design an easy-to-use service which will be integrated to “Galaxy ToolShed”, so that panISa can be easily integrated in any homemade bioinformatics pipeline, such as BLAST to search for identical or homologous IS. Furthermore, we want to investigate other kinds of mobile elements like eucarkyotic transposons and retrotransposons, reflecting on the ways to extend our second contribution to such kind of elements.

Finally, we can apply panISa on the major pathogenic dataset released in our pre-

viously published research. This will be done not only as additional validation of this tool, but also to discover hidden ISs, and possibly to explore how the evolution of those bacteria has been influenced by such ISs.

IV

APPENDIX

Clinical Microbiology and Infection xxx (2017)

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Original article

Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone

P. Treepong [1, 2], V.N. Kos [3], C. Guyeux [1], D.S. Blanc [4], X. Bertrand [5, 6], B. Valot [6], D. Hocquet [5, 6, 7, *]

[1] *UMR CNRS 6174, Institut FEMTO-ST, Département DISC, Université de Bourgogne Franche-Comté, Besançon, France*
[2] *Faculty of Technology and Environment, Prince of Songkla University, Phuket, Thailand*
[3] *Infection Innovative Medicines Unit, AstraZeneca R&D Boston, Waltham, MA, USA*
[4] *Service of Hospital Preventive Medicine and Institute of Microbiology, Lausanne University Hospital, Lausanne, Switzerland*
[5] *Laboratoire d'Hygiène Hospitalière, Centre Hospitalier Régional Universitaire, Besançon, France*
[6] *UMR CNRS 6249, Chrono-environnement, Université de Bourgogne Franche-Comté, Besançon, France*
[7] *Centre de Ressources Biologiques—Filière Microbiologique de Besançon, Centre Hospitalier Régional Universitaire, Besançon, France*

ARTICLE INFO

Article history:
Received 12 January 2017
Received in revised form
15 June 2017
Accepted 17 June 2017
Available online xxx

Editor: P.T. Tassios

Keywords:
Bacterial resistance
Epidemic
Fluoroquinolones
High-risk clones
International clones
Pathogen
Phylogeny

ABSTRACT

Objectives: Despite the non-clonal epidemic population structure of *Pseudomonas aeruginosa*, several multi-locus sequence types are distributed worldwide and are frequently associated with epidemics where multidrug resistance confounds treatment. ST235 is the most prevalent of these widespread clones. In this study we aimed to understand the origin of ST235 and the molecular basis for its success.
Methods: The genomes of 79 *P. aeruginosa* ST235 isolates collected worldwide over a 27-year period were examined. A phylogenetic network was built, using a Bayesian approach to find the Most Recent Common Ancestor, and we identified antibiotic resistance determinants and ST235-specific genes.
Results: Our data suggested that the ST235 sublineage emerged in Europe around 1984, coinciding with the introduction of fluoroquinolones as an antipseudomonal treatment. The ST235 sublineage seemingly spread from Europe via two independent clones. ST235 isolates then appeared to acquire resistance determinants to aminoglycosides, β-lactams and carbapenems locally. Additionally, we found that all the ST235 genomes contained the *exoU*-encoded exotoxin and identified 22 ST235-specific genes clustering in blocks and implicated in transmembrane efflux, DNA processing and bacterial transformation. These unique combinations of genes may have contributed to the poor outcome associated with *P. aeruginosa* ST235 infections and increased the ability of this international clone to acquire mobile resistance elements.
Conclusion: Our data suggest that *P. aeruginosa* ST235 (a) has become prevalent across the globe potentially due to the selective pressure of fluoroquinolones and (b) readily became resistant to aminoglycosides, β-lactams and carbapenems through mutation and acquisition of resistance elements among local populations. **P. Treepong, Clin Microbiol Infect 2017;■:1**

Introduction

Pseudomonas aeruginosa is a major opportunistic pathogen responsible for nosocomial infections in humans and for morbidity in individuals afflicted with cystic fibrosis [1]. This ubiquitous Gram-negative bacillus has a non-clonal epidemic population structure but with several sequence types (ST111, ST175, ST235, ST244 and ST395) distributed worldwide and frequently associated with outbreaks. ST235 is the most prevalent of these so-called 'international', 'high-risk', or 'widespread' clones associated with poor clinical outcomes in part due to multi-level and high-level antibiotic resistance [2–4]. Treatment of *P. aeruginosa* infections relies on three major antibiotic families: the β-lactams, aminoglycosides and fluoroquinolones. High levels of resistance to these

* Corresponding author. D. Hocquet, Laboratoire d'Hygiène Hospitalière, Centre Hospitalier Régional Universitaire, 3 boulevard Fleming, Besançon, Cedex 25030, France.
E-mail address: dhocquet@chu-besancon.fr (D. Hocquet).

http://dx.doi.org/10.1016/j.cmi.2017.06.018

compounds can be readily rendered by chromosomal changes. Furthermore, acquisition of resistance genes borne by specific genomic islands and associated transposons is particularly common among *P. aeruginosa* clinical isolates, with nearly 100 different horizontally acquired resistance elements currently reported in ST235 isolates [4,5]. The expression of these acquired genes, together with the intrinsic resistance mechanisms, considerably reduces the therapeutic options for the treatment of infections caused by ST235 *P. aeruginosa*.

Additionally, *P. aeruginosa* has the ability to cause severe infections due to its many virulence factors. During acute disease, this pathogen uses the toxins of the type III secretion system to circumvent the host immune system and establish infection. Of the four exotoxins (ExoS, ExoT, ExoU, ExoY), ExoU, a potent phospholipase that disrupts the plasma membrane and leads to rapid cell death, is the most virulent [6].

Despite its clinical importance, the molecular basis for the success of the ST235 clone is poorly understood. In addition, although the spread of ST235 has been documented in many geographic locations, little is known about the evolution and emergence of this clone on a global scale.

To better understand the history of the *P. aeruginosa* ST235 lineage as a high-risk international clone, we carried out a Bayesian phylogenetic reconstruction using 79 *P. aeruginosa* ST235 isolates collected from five continents over a 27-year period. Genome comparison of these isolates identified antibiotic-resistance determinants, virulence genes and ST235-specific genes that may have contributed to the success of this clone.

Materials and methods

Sequence type 235 genome collection

The genomes of 79 isolates representing sequence type (ST) 235 *P. aeruginosa* clonal cluster were obtained from various sources collected over a 27-year period (Fig. 1, Table 1). The 17 newly sequenced isolates used in this study were deposited at DDBJ/EMBL/GenBank (see Supplementary material, Table S1). The isolates came from Africa ($n = 6$), Asia ($n = 7$), Europe ($n = 34$), North America ($n = 21$) and South America ($n = 11$) (Fig. 1).

Core genome determination and ST235-specific gene identification

To define the 'core' genome of ST235 sublineage, the 79 genomes were aligned to *P. aeruginosa* reference isolate NCGM2.S1 using MUMmer [7]. The 'core' genome consists of regions that shared at least 98% identity and were present among all 79 ST235 genomes. The regions were then aligned with T-Coffee [8]. ST235-specific genes were defined by first searching for all possible ST235-similar genes from alignment to the 6358 genes present in NCGM2.S1 by GMAP with 95% identity and 60% coverage [9–11]. Then ST235-specific genes were defined as those present in ≥95% of the genomes ST235 isolates (i.e. ≥75 out of the 79 tested) and absent from all other isolates of the genus *Pseudomonas* with completely closed genomes available via the NCBI reference sequence database in October 2016 (i.e. 65 non-ST235 *P. aeruginosa* isolates and 177 *Pseudomonas* sp. isolates).

Phylogenetic network

The evolutionary relationship between the 79 ST235 isolates was investigated using an unrooted phylogenetic network constructed from the sequence alignment generated by SplitTree4 and clades were identified with the affinity propagation clustering (APcluster) R-package [12,13]. The split network was obtained from a Hamming distance-based method, namely NeighborNet [14]. Bootstrapping techniques were applied to generate 1000 networks.

Most Recent Common Ancestor analysis

To determine the most likely evolutionary scenario for the ST235 clone, we calculated the time of the Most Recent Common Ancestor (MRCA) of the 69 isolates with known collection dates using the BEAST2 package on a multi-TypeTree template with a strict clock model and 5 000 000 Markov chain Monte Carlo iterations, using the core genome determined as described above [15].

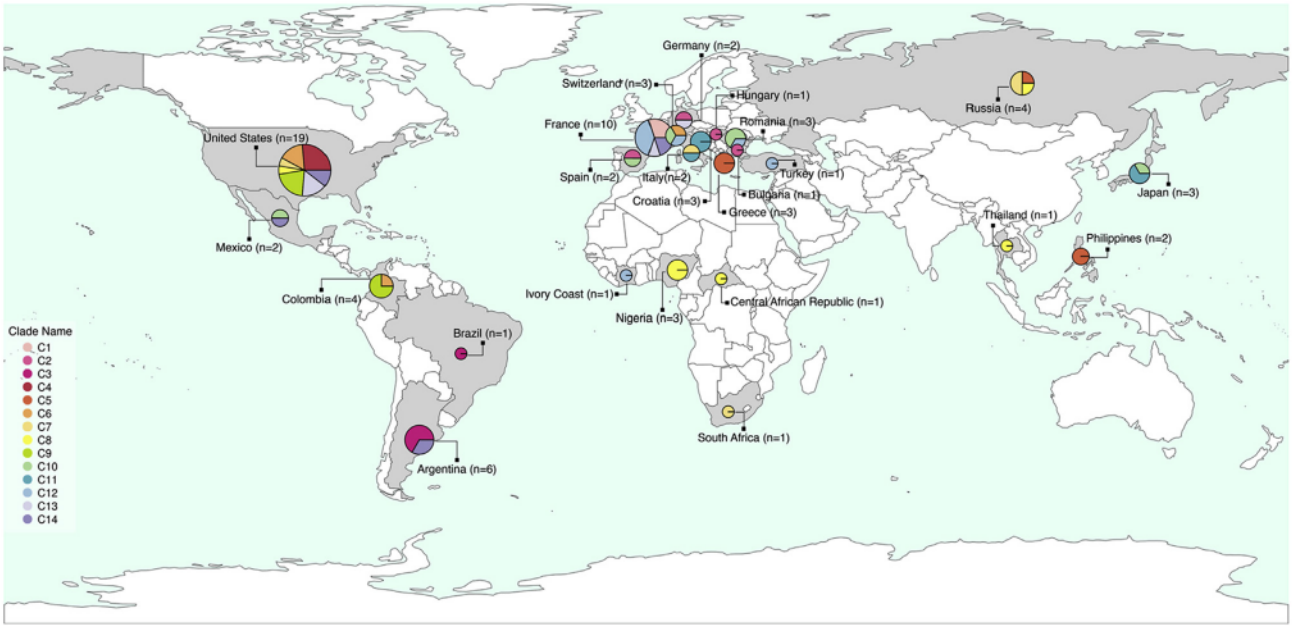


Fig. 1. Worldwide distribution of the 79 ST235 isolates of *Pseudomonas aeruginosa* which genomes were used in this study. The countries of origin of the isolates are shaded in grey. Pie chart diameters are proportional to the number of isolates collected from each country. The area of each slice is proportional to the quantity of isolates of each clade.

Please cite this article in press as: Treepong P, et al., Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone, Clinical Microbiology and Infection (2017), http://dx.doi.org/10.1016/j.cmi.2017.06.018

Table 1
Genes acquired by ST235 isolates that confer resistance to aminoglycosides, to extended-spectrum β-lactams, to carbapenems and fluoroquinolones

Isolate	Country	Genes encoding aminoglycoside-modifying enzymes																				Genes encoding β-lactamase with extended-spectrum																FA[a]	Q[b]	
		aac(3)-Ic	*aac(3)-Id*	*aac(6')-Iae*	*aac(6')-Ib*	*aac(6')-IIb*	*aac(6')-33*	*aacA4*	*aacA29*	*aadA1*	*aadA2*	*aadA6*	*aadA11*	*aadB*	*aph(3')-Ia*	*aph(3')-Ic*	*aph(3')-IIa*	*aph(3')-VIa*	*aph(3')-VIb*	*aph(3')-XV*	*aph(6)-Ic*	*strA*	*strB*	*bla* BEL-1	*bla* CTX-M-2	*bla* GES-1	*bla* GES-5	*bla* GES-19	*bla* GES-20	*bla* KPC-2	*bla* PER-1	*bla* PER-3	*bla* SHV-2	*bla* IMP-1	*bla* IMP-15	*bla* IMP-26	*bla* IMP-34	*bla* VIM-2	*bla* VIM-4	*bla* VIM-11	*bla* OXA-17	*bla* OXA-19	*bla* OXA-28	*bla* OXA-35	*bla* OXA-74	*bla* OXA-129	*aac(6')-Ib-cr*	*QnrVC1*	Clade
76_FR_92	France																																											1
77_FR_92	France																																											
86_FR_92	France																																										
00_ES_10	Spain																																										2
09_BG_97	Bulgaria																																										
47_DE_97	Germany																																										
68_HU_97	Hungary																																										
30_AR_05	Argentina																																										3
63_BR_12	Brazil																																										
79_AR_10	Argentina																																										
92_AR_12	Argentina																																										
94_AR_12	Argentina																																										
06_US_11	USA																																										4
22_US_09	USA																																										
42_US_06	USA																																										
43_US_06	USA																																										
79_US_10	USA																																										
02_PH_12	Philippines																																										5
03_PH_12	Philippines																																										
05_GR_12	Greece																																										
06_GR_12	Greece																																										
08_GR_12	Greece																																										
09_RU_12	Russian Federation																																										
03_CO_11	Colombia																																										6
44_CH_12	Switzerland																																										
BL14_US_un	USA																																										
BWH_US_13	USA																																										
BWHP_US_13	USA																																										
01_US_12	USA																																										7
23_IT_12	Italy																																										
33_RU_13	Russian Federation																																										
35_RU_13	Russian Federation																																										
98_ZA_12	South Africa																																										
15D_NG_un	Nigeria																																										8
18_CF_11	Central African Republic																																										
22_TH_12	Thailand																																										
26_US_12	USA																																										
36D_NG_un	Nigeria																																										
37_RU_13	Russian Federation																																										
49_NG_13	Nigeria																																										
19_CO_12	Colombia																																										9
20_CO_12	Colombia																																										
21_CO_12	Colombia																																										
BL17_US_un	USA																																										
E2UoS_US_un	USA																																										
JJ692_US_un	USA																																										
U2504_US_un	USA																																										
23_CH_14	Switzerland																																										10
80_MX_10	Mexico																																										

(*continued on next page*)

Table 1 (*continued*)

Isolate	Country	Genes encoding aminoglycoside-modifying enzymes																				Genes encoding β-lactamase with extended-spectrum												FA[a]	Q[b]	
		aac(3)-Ic	*aac(3)-Id*	*aac(6′)-Iae*	*aac(6′)-Ib*	*aac(6′)-IIb*	*aac(6′)-33*	*aacA4*	*aacA29*	*aadA1*	*aadA2*	*aadA6*	*aadA11*	*aadB*	*aph(3′)-Ia*	*aph(3′)-Ic*	*aph(3′)-IIa*	*aph(3′)-VIa*	*aph(3′)-VIb*	*aph(3′)-XV*	*aph(6)-Ic*	*strA*	*strB*	*bla* BEL-1	*bla* CTX-M-2	*bla* GES-1	*bla* GES-5	*bla* GES-19	*bla* GES-20	*bla* KPC-2	*bla* PER-1	*bla* PER-3	*bla* SHV-2	*bla* IMP-1	*bla* IMP-15	*bla* IMP-26	*bla* IMP-34	*bla* VIM-2	*bla* VIM-4	*bla* VIM-11	*bla* OXA-17	*bla* OXA-19	*bla* OXA-28	*bla* OXA-35	*bla* OXA-74	*bla* OXA-129	*aac(6′)-Ib-cr*	*QnrVC1*	Clade
94_RO_12	Romania				■									■					■	■		■	■							■															■		■		
NCGM_JP_02	Japan			■						■											■											■															
P235_ES_08	Spain				■		■	■		■							■				■							■	■																		
WS394_RO_un	Romania				■														■	■		■								■														■		■		
12_HR_08	Croatia											■																																				11
1900_JP_12	Japan				■			■				■																									■										
1984_JP_12	Japan				■			■			■	■																									■										
30_IT_12	Italy				■			■				■																																			
68_HR_09	Croatia				■			■				■																																			
87_HR_07	Croatia				■			■				■																																			
06_FR_92	France												■	■								■																									12
08_CI_08	Ivory Coast												■																																		
48_CH_06	Switzerland	■										■																																		
53_FR_10	France												■	■																																
70_FR_92	France												■																																		
76_TR_97	Turkey	■										■																																		
87_FR_91	France																																													
90_RO_12	Romania				■			■				■										■	■																							
10_US_12	USA											■					■				■																									13
15_DE_11	Germany					■				■		■		■																																
32_US_12	USA											■					■				■																									
74_US_07	USA											■																																		
SPae_FR_04	France				■			■				■																																		
02_AR_08	Argentina				■			■																																					14
49_AR_09	Argentina				■			■																																					
58_FR_12	France																																												
6077_US_un	USA				■			■				■																																		
78_FR_88	France													■																																
89_MX_12	Mexico				■					■		■						■								■																				
BL08_US_un	USA											■																																		

[a] Genes encoding resistance determinants to fluoroquinolones and to aminoglycosides.

[b] Genes encoding resistance determinants to fluoroquinolones. The gene *aph(3′)-IIb* was present in all the isolates (except 44_CH_12, 43_US_06, 18_CF_11). The genes bla_{OXA-50}, bla_{AmpC}, and *fosA* were present in all the isolates tested.

Resistance and virulence gene identification

Genomes of all ST235 isolates were scaffolded using Ragout with the *P. aeruginosa* NCGM2.S1 genome as a reference [16]. Sequences sharing ≥95% identity over ≥60% of the gene length with resistance and virulence determinants of the ResFinder and VFDB databases were identified using GMAP [9–11]. Virulence factors were compared to those of the core genome of the species [17]. To detect mutations that confer high-level resistance to β-lactams and fluoroquinolones, we looked for non-synonymous mutations and insertion sequences in the genes encoding the cephalosporinase AmpC and its regulators, in *oprD*, and in the quinolone-determining regions (QRDR).

Results

ST235 population structure by phylogenetic network analysis

A phylogenetic network built from the core genome alignment depicted 14 distinct clades (C1 to C14; Fig. 2). The largest sampling of isolates within this collection came from the USA (19 isolates) and France (ten isolates). These isolates were scattered throughout the phylogenetic network, with isolates collected in the USA appearing in seven clades and those from France in four clades. Focusing on the spatiotemporal distribution of isolates within each clade, four clades were found to be comprised of isolates from a single country or continent of origin: C1 and C2 included three and four isolates, respectively, from France (C1) and Europe (Spain, Germany, Hungary and Bulgaria; C2); C3 included five isolates from South America; and C4 included five isolates from North America (Fig. 2). Although a majority of the ST235 clades suggests a global dispersion, countrywide spreads are confirmed (Fig. 1) [18,19]. Isolates from C1, C2 and C3 shared a geographical origin (France for C1, Spain/Germany/Bulgaria/Hungary for C2, Argentina/Brazil for C3; Fig. 2, and see Supplementary material, Table S1). In contrast, the remaining clades included isolates obtained from two to four continents. Most representative of the worldwide spread of the clone ST235 was C8—seven isolates collected within a 2-year time frame (2011–2013) from four distinct continents (Africa, North America, Asia and Europe).

Spatiotemporal origin of the ST235 clone

The MRCA appeared approximately 32 years ago (~1984) with a very high confidence interval (95% CI 29.12–30.82 years from the date of isolation of the latest isolate in 2014) (Fig. 3a). The oldest isolate within the current collection was isolated in France in 1988 (78_FR_88), 4 years after the MRCA. The clade of French isolates in

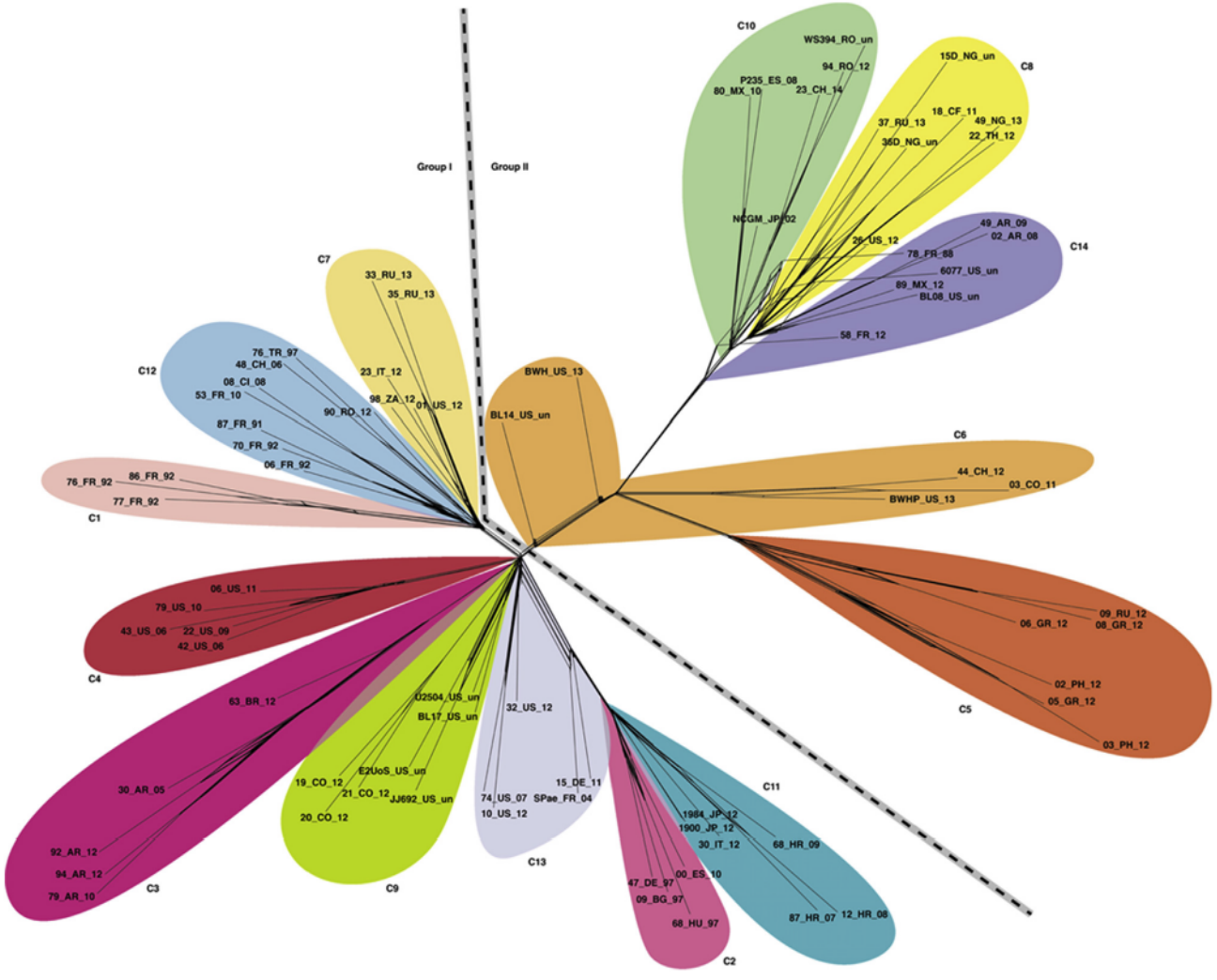


Fig. 2. Phylogenetic network of ST235 *Pseudomonas aeruginosa*. The clades, labelled from C1 to C14 and shaded with different colours for clarity, formed two groups (Group I and Group II) separated by a dashed line.

Please cite this article in press as: Treepong P, et al., Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone, Clinical Microbiology and Infection (2017), http://dx.doi.org/10.1016/j.cmi.2017.06.018

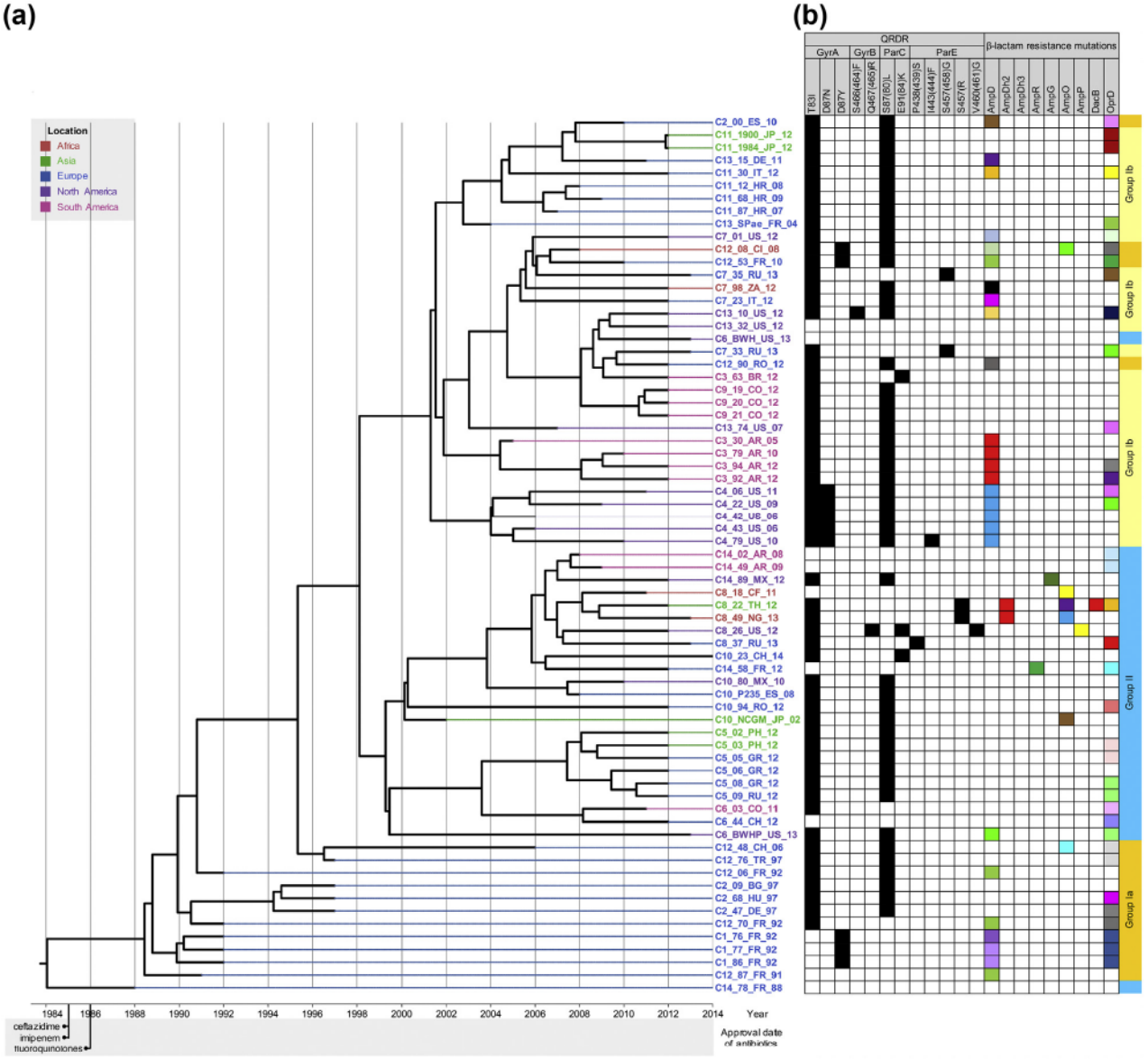


Fig. 3. Time of Most Recent Common Ancestor (MRCA) of *Pseudomonas aeruginosa* ST235 and chromosomal mutations conferring high-level resistance to fluoroquinolones, extended-spectrum cephalosporinases, and carbapenems. (a) Phylogenetic tree with time scale was calculated from the 69 genomes of isolates with known isolation date with BEAST2 using continent origin as co-variable. The estimated mutation rate was 4.85×10^{-6} (95% CI 4.59×10^{-6} to 5.15×10^{-6}) per site per year. The time of MRCA is ~30 years ago from 2014. The tips are labelled with the isolate name and are coloured by continent of origin (see insert). Names of the isolates are prefixed with the clade to which they belong. (b) Mutations in the quinolone-determining regions (QRDRs) of GyrA, GyrB, ParC and ParE, in the regulators of the cephalosporinase AmpC and in OprD. For the QRDR, the numbers of the corresponding codons in *Escherichia coli* are in parentheses. Black cells and white cells indicate the presence or absence of a given mutation, respectively. For mutation in each regulator of the cephalosporinase AmpC and in OprD, every single mutation is represented by a single colour. White cells indicate an intact protein. The detail of the mutations is given in the Supplementary material (Table S3). Every protein was compared with its closest homologue born by a β-lactam-susceptible isolate of *P. aeruginosa* (strain M18 for AmpD, AmpP, DacB; strain PA14 for AmpDh2, AmpG, AmpO, OprD; strain MBT-1 for AmpDh3, AmpR). AmpO polymorphism for all the isolates of the collection: S125T, D256E. AmpP polymorphism for all the isolates of the collection: L74F and L98F.

the time-scale analysis suggests that the clone emerged in this country (Fig. 3a). A split of more recent isolates in two sub-lineages (Groups I and II, Fig. 2) was retrieved by both phylogenetic analyses. Hence, 100% of the 1000 phylogenetic trees that were aggregated to build the network (Fig. 2) identified the split into the two groups, I and II. This split was confirmed independently with a Bayesian approach (Fig. 3) with a posterior probability score of 1.0. These data suggested that the international spread of ST235 implies at least two independent clades. The single nucleotide polymorphisms that discriminated these two clades were neither in resistance genes nor in virulence genes. Group I consisted of clades that shared a common ancestor that emerged in ~2001 (C3, C4, C7, C9, C11 and C13) with older isolates that were identified in Europe (C1, C2 and C12; Figs. 2 and 3a). Group II (C5, C6, C8, C10 and C14) shared a common ancestor that emerged in ~1999 and then appeared to spread worldwide.

Cumulative resistance to antibiotics by chromosomal mutations

We specifically searched for non-synonymous mutations in the QRDRs of each isolate and correlated them with their date of isolation. The two earliest isolates (78_FR_88 and 87_FR_91,

retrieved in France in 1988 and 1991, respectively) had wild-type QRDR sequences compatible with a full susceptibility to fluoroquinolones (Fig. 3b). Among the five isolates retrieved in 1992, four displayed a T83I or a D87Y change in GyrA. Mutations in the QRDR of *parC* first appeared in 1992 in isolate 06_FR_92. Analysis of isolates collected post-1992 showed that the number of QRDR mutations per isolate seemingly grew over time with the late isolates (2011 to 2014) accumulating mostly T83I change in GyrA (33 of 38 isolates) and S80L change in ParC (24 of 38 isolates), with occasional additional substitution in position 87 of GyrA, in GyrB, and in ParE (Fig. 3b).

Mutation-dependent overproduction of intrinsic β-lactamase AmpC is the main cause of resistance of clinical strains of *P. aeruginosa* to antipseudomonal penicillins and cephalosporins. Additionally, loss or alteration of the outer membrane porin protein OprD is by far the most common mechanism of resistance to the carbapenems in *P. aeruginosa* [20]. Hence, we searched for non-synonymous mutations and insertion sequences in genes whose inactivation up-regulates AmpC cephalosporinase production and in *oprD* (Fig. 3b). Out of the 79 isolates, 36 had acquired one or more AmpC regulator mutations (Fig. 3b, and see Supplementary material, Table S2). Of these 36 isolates, 27 produced a mutated transcriptional regulator AmpD. Sequence analysis of the OprD porin showed that 39 isolates had acquired mutations that presumably affect the porin activity. The genes *ampD* and *oprD* had 15 and 29 different types of non-synonymous mutations, respectively. Mutations among the AmpC regulators and OprD were mostly unique in contrast to the relative uniformity in the QRDR mutations (T83I in GyrA, S80L in ParC) in the late isolates.

High diversity of foreign antibiotic resistant determinants among the ST235 isolates

Table 1 details the antibiotic resistance genes acquired by the isolates (see Supplementary material, Table S3). Analysis of the aminoglycoside-modifying enzyme content among the collection revealed that most (72 out of 79) of the ST235 isolates harboured at least one aminoglycoside-modifying enzyme potentially conferring a decrease in aminoglycoside susceptibility (see Supplementary material, Table S3).

Approximately 40% (32 of 79) of the isolates possessed at least one of 23 β-lactamases characterized as having extended spectrum activity (Table 1). Seven isolates accumulated two β-lactamases with an extended spectrum, mostly combining the production of an extended-spectrum oxacillinase with that of an extended-spectrum β-lactamase or a metallo-β-lactamase. Extended-spectrum enzymes were generally accumulated within specific sub-lineages. For example, bla_{OXA-17} and $bla_{OXA-129}$ were uniquely identified in isolates belonging to C3, bla_{OXA-19} to C5, bla_{KPC-2} to C9, bla_{GES-19}, bla_{PER-1} and bla_{OXA-74} to C10, and bla_{IMP-34} to C11 (Table 1). Within C5, the three Greek isolates harboured either bla_{VIM-2}, bla_{VIM-4}, bla_{OXA-19}, or bla_{OXA-35} whereas the two isolates from the Philippines had bla_{IMP-26} or bla_{VIM-2}. Other rare extended-spectrum enzymes were shared by clades with bla_{IMP-1} harboured by isolates from C8 (one of seven isolates) and C10 (one of six isolates), and bla_{VIM-2} present in isolates of C5 (two of six isolates) and C7 (two of five isolates), which were most likely acquired independently. Acquired resistance determinants were not detected within six of the isolates of the collection (C6_03_CO_11, C6_44_CH_12, C6_BL14_US_un, C6_BWH_US_13, C6_BWHP_US_13 and C14_58_FR_12) with five of these isolates originating from diverse geographic locations but clustering in C6 (see Supplementary material, Table S3). The number of acquired resistance genes was significantly lower ($p < 0.05$) in the isolates collected at an earlier time-point (1988–1997, 11 isolates) than those collected between 2011 and 2014 (38 isolates).

ST235-specific determinants

One could imagine that the global success of *P. aeruginosa* international clones (e.g. ST111, ST175, ST235, ST244 and ST395) could rely on specific determinants. Unfortunately, we did not retrieve any gene signature shared among these widespread clones (data not shown). We then identified genes present in the genome of all ST235 isolates but absent from a majority of the non-ST235 isolates. The type III secretion system exotoxin encoded by the *exoU* gene was present in all of the ST235 genomes, although not specific for this lineage. Hence, the analysis of 65 non-ST235 genomes of *P. aeruginosa* also detected the presence of the *exoU* gene in 18 strains representative of ST253 (e.g. strain PA14), ST313, ST316, ST357, ST823, ST1024, ST1047, ST1971, and an unassigned ST (strain MTB-1). We further identified 22 ST235-specific genes which clustered into three blocks (Table 2). Block 1 contained nine genes and was part of ExoU island A. Three genes (NCGM2_1830 to NCGM2_1832) encoded homologues of the components (TolC, EmrA, EmrB) of a tripartite efflux pump. Two other contiguous genes (NCGM2_1836 and NCGM2_1837) encoded a putative transporter and a periplasmic adaptor, respectively, that could act together as another efflux pump. The production of these two efflux pumps could specifically enhance the resistance of ST235 isolates to antibiotics. Block 2 included ten genes with nine encoding proteins implicated in DNA processing (a P-loop NTPase, two type-1 restriction endonucleases—HsdR and HsdS, a UvrD/REP helicase, an SMC domain-containing protein and two DNA methylases, DprA and RecQ).

Discussion

ST235-specific determinants

We identified *dprA* as a specific determinant of ST235 sublineage. DprA is required for the protection of incoming single-stranded DNA and interacts with the ubiquitous recombinase RecA to integrate the acquired DNA into the host chromosome of naturally transformable bacterial species such as *Streptococcus pneumoniae*. RecQ is a DNA helicase that affects DNA transformation in Gram-negative and Gram-positive bacteria [21]. Of note, the rest of the transformation machinery (i.e. secretion channel PilQ, DNA receptor ComE, transmembrane channel ComA, translocase ComFA, and the competence activator Tfox) was conserved in the species *P. aeruginosa* [17]. ST235 displays a very high diversity of acquired resistance genes (Table 1, see also Supplementary material, Table S3) [4]. DprA universality among transformable species and its demonstrated role in homologous recombination provide evidence that its presence in ST235 possibly increases the ability of this widespread clone to acquire and maintain foreign resistance elements at a greater rate than other *P. aeruginosa* clones. Overall, 22 genes were found to be unique to the ST235 sublineage and encoded proteins implicated in DNA processing, transport through the membrane, and bacterial transformation. The role of these ST235-specific proteins in the success of the clone (via higher intrinsic resistance to antibiotics or easier acquisition of foreign resistance determinants) is speculative and deserves experimental verification.

Although not fully specific for the ST235 clone, *exoU*-encoded exotoxin was retrieved in all ST235 isolates. The presence and production of ExoU is a marker for early mortality associated with *P. aeruginosa* infections [22]. The virulence in *P. aeruginosa* is multifactorial and combinatorial; however, these data suggest that

Please cite this article in press as: Treepong P, et al., Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone, Clinical Microbiology and Infection (2017), http://dx.doi.org/10.1016/j.cmi.2017.06.018

Table 2
Description of the 22 genes highly conserved in and specific to *Pseudomonas aeruginosa* ST235 lineage

Block number	Gene symbol in strain NCGM2.S1	Domain	Additional description	Accession no. of the closest homologue (name, bacterial species, % identity)
1	NCGM2_1826	Transposase	—	—
	NCGM2_1828	α/β hydrolase family protein	—	—
	NCGM2_1829	Pirin-related protein	Putative transcriptional regulation	—
	NCGM2_1830	Putative RND outer membrane protein (TolC family)	—	NP_417507 (TolC, *Escherichia coli*, 22%)
	NCGM2_1831	Putative RND membrane fusion protein (ErmA family)	—	NP_417170.1 (EmrA, *E. coli*, 50%)
	NCGM2_1832	Putative MFS multidrug efflux transporter (EmrB family)	—	NP_418166.1 (EmrB, *E. coli*, 24%)
	NCGM2_1836	Putative transporter membrane protein	—	WP_058142560.1 (*P. aeruginosa*, 99%)
	NCGM2_1837	Putative RND membrane fusion protein (HlyD/EmrA family)	—	—
	NCGM2_1838	PucR C-terminal helix-turn-helix	Probable transcriptional regulator	NP_391122.1 (*Bacillus subtilis*, 45%)
2	NCGM2_3761	Hypothetical protein	—	—
	NCGM2_3762	P-loop NTPase	Involved in replication	WP_025991883.1 (*P. aeruginosa*, 99%)
	NCGM2_3765	Type-I restriction endonuclease HsdR	Restriction-modification system	—
	NCGM2_3766	Type-I restriction endonuclease HsdS	Restriction-modification system	—
	NCGM2_3767	UvrD/REP helicase	—	—
	NCGM2_3768	SMC domain-containing protein	Replication, recombination and DNA repair	—
	NCGM2_3769	N-6 DNA methylase	—	—
	NCGM2_3770	N-7 DNA methylase	—	—
	dprA	DNA protection protein	Dedicated to natural bacterial transformation	—
	recQ	ATP-dependent DNA helicase	Involved in genome maintenance	—
	leuS	Leucyl-tRNA synthetase	—	—
3	NCGM2_6332	Hypothetical protein	—	—
	NCGM2_6333	DEAD/DEAH box helicase	—	—

ExoU production could participate in the poor outcome of infections due to ST235 [4,23].

High diversity of resistance determinants to aminoglycosides and to β-lactams in ST235 isolates

The variety of both chromosomal and acquired determinants of resistance to aminoglycosides and β-lactams in the clone ST235 (Fig. 3, Table 1) indicates independent and local acquisition in line with previous observations [4,24]. Hence, one can hypothesize that the global spread of these lineages does not fully rely on resistance to aminoglycosides and to β-lactams. This high diversity of foreign resistance determinants in ST235 contrasts with that observed in another widespread clone, ST175, in which the resistance to antibiotics mainly occurs via chromosomal mutations [25].

Altogether, the mosaic of either acquired or mutational resistance determinants could argue for a limited role of the selective pressure of aminoglycosides and β-lactams in the international spread of ST235 clone (Table 1, Fig. 3b).

Role of the fluoroquinolones in the spread and emergence of ST235

Interestingly, the emergence of the ST235 sublineage in 1984 coincides with the beginning of the use of antipseudomonal fluoroquinolones (pefloxacin, ofloxacin and ciprofloxacin) between 1984 and 1987. The emergence of fluoroquinolone-resistant mutants of *P. aeruginosa* after treatment with these compounds has been well documented [26]. The expansion of ST235 may have been favoured by the extensive use of fluoroquinolones, selecting for QRDR mutations. Similarly, the worldwide expansion of the major pathogens *Escherichia coli* ST131 H30-Rx, methicillin-resistant *Staphylococcus aureus* EMRSA-15 in the mid-1980s, and that of *Clostridium difficile* 027 in the early 1990s also incriminated the clinical use of fluoroquinolones [27–29]. Although QRDR mutations have been observed in many other clones of *P. aeruginosa*, it has been experimentally demonstrated that the fitness cost of mutations in the QRDR of *P. aeruginosa* depends on the genetic background of the strains [30]. Hence, it has been recently shown that mutations in *parC* increase or have no effect on the fitness of *exoU* strains, whereas they decrease that of *exoS* strains [31]. Overall, although other features of ST235 may have contributed to its spread, fluoroquinolone use could have highly favoured the spread of *exoU* strains (including ST235) in which the fitness burden of the resistance to these major antibiotics is lower.

Weaknesses and limitations of the study

Although a thorough genotypic analysis was completed on these isolates, the corresponding phenotypic antibiotic resistance level of a majority of the isolates was unknown. However, previous studies have illustrated significant correlation between the genotype and phenotype for β-lactams, fluoroquinolones and cephalosporins [32,33]. Despite the particular efforts we made to collect 'historic' ST235 genomes collected pre-1992 and the high degree of confidence of the MRCA age, the relatively low number of early isolates could have biased the result. Additionally, the isolate collection explored here ($n = 79$) is of limited size and suffers from a bias in geographical distribution, with a low representation of isolates from Asia and Africa compared with isolates from North America and Europe.

Conclusion

Analysis of the genome sequences of 79 isolates of *P. aeruginosa* ST235 obtained over a 27-year period from diverse regions of the

Please cite this article in press as: Treepong P, et al., Global emergence of the widespread *Pseudomonas aeruginosa* ST235 clone, Clinical Microbiology and Infection (2017), http://dx.doi.org/10.1016/j.cmi.2017.06.018

globe was used to gain an understanding of the epidemiology of this international high-risk clone. Clustering analysis confirmed that clonal spreads occur on a country or regional scale but also revealed that ST235 subclones can spread across continents. Comparative genomic analysis suggested that the ST235 most likely emerged as a global clone due to a unique combination of chromosomal genes (*exoU*, *dprA* related gene) and the ability of the clone to readily acquire antibiotic resistance genes that limit antibiotic treatment options.

Transparency declaration

The authors have declared that no competing interests exist.

Acknowledgements

We thank Damien Fournier for helpful discussion. Computations have been performed at the 'Mésocentre de Calculs de Franche-Comté'. We also thank Fabrice Poncet from the sequencing facility of the SFR FED 4234 (University of Franche-Comté, Besançon).

Funding

PT was supported by a grant from Prince of Songkla University, Thailand. The project was partially granted by the Région Franche-Comté (Grant 'Environnement, Homme, Territoire' of the Réseau franco-suisse). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

BV and DH conceived and designed the experiments; PT, CG and BV performed the experiments; PT, VNK, DB, XB, BV and DH analysed the data; and PT, BV and DH wrote the paper.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cmi.2017.06.018.

References

[1] Gellatly SL, Hancock REW. *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. Pathog Dis 2013;67:159–73.
[2] Pirnay J-P, De Vos D, Cochez C, Bilocq F, Vanderkelen A, Zizi M, et al. *Pseudomonas aeruginosa* displays an epidemic population structure. Environ Microbiol 2002;4:898–911.
[3] Woodford N, Turton JF, Livermore DM. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. FEMS Microbiol Rev 2011;35:736–55.
[4] Oliver A, Mulet X, López-Causapé C, Juan C. The increasing threat of *Pseudomonas aeruginosa* high-risk clones. Drug Resist Updat 2015;21–22:41–59.
[5] Roy Chowdhury P, Scott MJ, Djordjevic SP. Genomic islands 1 and 2 carry multiple antibiotic resistance genes in *Pseudomonas aeruginosa* ST235, ST253, ST111 and ST175 and are globally dispersed. J Antimicrob Chemother 2017;72:620–2.
[6] Sato H, Frank DW, Hillard CJ, Feix JB, Pankhaniya RR, Moriyama K, et al. The mechanism of action of the *Pseudomonas aeruginosa*-encoded type III cytotoxin, ExoU. EMBO J 2003;22:2959–69.
[7] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5:R12.
[8] Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–17.
[9] Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012;67:2640–4.
[10] Chen L. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 2004;33:D325–8.
[11] Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–75.
[12] Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 2006;23:254–67.
[13] Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. Bioinformatics 2011;27:2463–4.
[14] Bryant D, Moulton V. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. Lect Notes Comput Sci 2002:375–91.
[15] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10, e1003537.
[16] Kolmogorov M, Raney B, Paten B, Pham S. Ragout-a reference-assisted assembly tool for bacterial genomes. Bioinformatics 2014;30:i302–9.
[17] Valot B, Guyeux C, Rolland JY, Mazouzi K, Bertrand X, Hocquet D. What it takes to be a *Pseudomonas aeruginosa*? The core genome of the opportunistic pathogen updated. PLoS One 2015;10, e0126468.
[18] Edelstein MV, Skleenova EN, Shevchenko OV, D'souza JW, Tapalski DV, Azizov IS, et al. Spread of extensively resistant VIM-2-positive ST235 *Pseudomonas aeruginosa* in Belarus, Kazakhstan, and Russia: a longitudinal epidemiological and clinical study. Lancet Infect Dis 2013;13:867–76.
[19] Koutsogiannou M, Drougka E, Liakopoulos A, Jelastopulu E, Petinaki E, Anastassiou ED, et al. Spread of multidrug-resistant *Pseudomonas aeruginosa* clones in a university hospital. J Clin Microbiol 2013;51:665–8.
[20] Fournier D, Richardot C, Müller E, Robert-Nicoud M, Llanes C, Plésiat P, et al. Complexity of resistance mechanisms to imipenem in intensive care unit strains of *Pseudomonas aeruginosa*. J Antimicrob Chemother 2013;68:1772–80.
[21] Johnston C, Martin B, Fichant G, Polard P, Claverys J-P. Bacterial transformation: distribution, shared mechanisms and divergent control. Nat Rev Microbiol 2014;12:181–96.
[22] Peña C, Cabot G, Gómez-Zorrilla S, Zamorano L, Ocampo-Sosa A, Murillas J, et al. Influence of virulence genotype and resistance profile in the mortality of *Pseudomonas aeruginosa* bloodstream infections. Clin Infect Dis 2015;60:539–48.
[23] Lee DG, Urbach JM, Wu G, Liberati NT, Feinbaum RL, Miyata S, et al. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol 2006;7:R90.
[24] Juan C, Zamorano L, Mena A, Albertí S, Pérez JL, Oliver A. Metallo-β-lactamase-producing *Pseudomonas putida* as a reservoir of multidrug resistance elements that can be transferred to successful *Pseudomonas aeruginosa* clones. J Antimicrob Chemother 2010;65:474.
[25] Cabot G, López-Causapé C, Ocampo-Sosa AA, Sommer LM, Domínguez MÁ, Zamorano L, et al. Deciphering the resistome of the widespread *Pseudomonas aeruginosa* sequence type 175 international high-risk clone through whole-genome sequencing. Antimicrob Agents Chemother 2016;60:7415–23.
[26] Ball P. Emergent resistance to ciprofloxacin amongst *Pseudomonas aeruginosa* and *Staphylococcus aureus*: clinical significance and therapeutic approaches. J Antimicrob Chemother 1990;26(Suppl. F):165–79.
[27] Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, et al. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. MBio 2016;7.
[28] Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res 2013;23:653–64.
[29] He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. Nat Genet 2013;45:109–13.
[30] Kugelberg E, Löfmark S, Wretlind B, Andersson DI. Reduction of the fitness burden of quinolone resistance in *Pseudomonas aeruginosa*. J Antimicrob Chemother 2005;55:22–30.
[31] Agnello M, Finkel SE, Wong-Beringer A. Fitness cost of fluoroquinolone resistance in clinical isolates of *Pseudomonas aeruginosa* differs by type III secretion genotype. Front Microbiol 2016;7:1591.
[32] Kos VN, McLaughlin RE, Gardner HA. Elucidation of mechanisms of ceftazidime resistance among clinical isolates of *Pseudomonas aeruginosa* by using genomic data. Antimicrob Agents Chemother 2016;60:3856–61.
[33] Kos VN, Déraspe M, McLaughlin RE, Whiteaker JD, Roy PH, Alm RA, et al. The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. Antimicrob Agents Chemother 2015;59:427–36.

BIBLIOGRAPHY

[Addario-Berry et al., 2004] Addario-Berry, L., Chor, B., Hallett, M., Lagergren, J., Panconesi, A., and Wareham, T. (2004). Ancestral maximum likelihood of evolutionary trees is hard. *Journal of Bioinformatics and Computational Biology*, 2(02):257–271.

[Agnello et al., 2016] Agnello, M., Finkel, S. E., and Wong-Beringer, A. (2016). Fitness cost of fluoroquinolone resistance in clinical isolates of *Pseudomonas aeruginosa* differs by type III secretion genotype. *Front. Microbiol.*, 7:1591.

[Aguilar-Rodea et al., 2017] Aguilar-Rodea, P., Zúñiga, G., Rodríguez-Espino, A. B., Olivares Cervantes, L. A., Gamiño Arroyo, E. A., Moreno-Espinosa, S., de la Rosa Zamboni, D., López Martínez, B., Castellanos-Cruz, C. M. D., Parra-Ortega, I., Jiménez Rojas, L. V., Vigueras Galindo, C. J., and Velázquez-Guadarrama, N. (2017). Identification of extensive drug resistant *Pseudomonas aeruginosa* strains: New clone st1725 and high-risk clone st233. *PloS one*, 12(3).

[Altschul and Gish, 1996] Altschul, S. F. and Gish, W. (1996). [27] local alignment statistics. *Methods in enzymology*, 266:460–480.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[Altschul et al., 1997] 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 research*, 25(17):3389–3402.

[Andersson and Hughes, 2010] Andersson, D. and Hughes, D. (2010). Antibiotic resistance and its cost: Is it possible to reverse resistance? *Nature Reviews Microbiology*, 8(4):260–271. cited By 625.

[Balasubramanian et al., 2013] Balasubramanian, D., Schneper, L., Kumari, H., and Mathee, K. (2013). A dynamic and intricate regulatory network determines *Pseudomonas aeruginosa* virulence. *Nucleic Acids Research*, 41(1):1.

[Ball, 1990] Ball, P. (1990). Emergent resistance to ciprofloxacin amongst *Pseudomonas aeruginosa* and *Staphylococcus aureus*: clinical significance and therapeutic approaches. *J. Antimicrob. Chemother.*, 26 Suppl F:165–179.

[Bankevich et al., 2012] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, S. A., Lesin, M. V., Nikolenko, I. S., Pham, S., Prjibelski, D. A., Pyshkin, V. A., Sirotkin, V. A., Vyahhi, N., Tesler, G., Alekseyev, A. M., and Pevzner, A. P. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology: a journal of computational molecular cell biology*, 19(5):455–477.

[Bao et al., 2011] Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y.-Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *Journal of human genetics*, 56(6):406–414.

[Baquero, 2004] Baquero, F. (2004). From pieces to patterns: evolutionary engineering in bacterial pathogens. *Nature Reviews Microbiology*, 2(6):510–518.

[Barrick et al., 2014] Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., Knoester, D. B., Reba, A., and Meyer, A. G. (2014). Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC genomics*, 15:1039.

[Battle et al., 2008] Battle, S. E., Meyer, F., Rello, J., Kung, V. L., and Hauser, A. R. (2008). Hybrid pathogenicity island pagi-5 contributes to the highly virulent phenotype of a *Pseudomonas aeruginosa* isolate in mammals. *Journal of bacteriology*, 190(21):7130–7140.

[Beceiro et al., 2013] Beceiro, A., Tomás, M., and Bou, G. (2013). Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clinical Microbiology Reviews*, 26(2):185–230.

[Ben Zakour et al., 2016] Ben Zakour, N. L., Alsheikh-Hussain, A. S., Ashcroft, M. M., Khanh Nhu, N. T., Roberts, L. W., Stanton-Cook, M., Schembri, M. A., and Beatson, S. A. (2016). Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of escherichia coli ST131. *MBio*, 7(2):e00347–16.

[Bennett, 2008] Bennett, M. P. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, 153 Suppl 1:S347–S357.

[Berrazeg et al., 2015] Berrazeg, M., Jeannot, K., Ntsogo Enguéné, V. Y., Broutin, I., Loeffert, S., Fournier, D., and Plésiat, P. (2015). Mutations in β-Lactamase AmpC increase resistance of *Pseudomonas aeruginosa* isolates to antipseudomonal cephalosporins. *Antimicrob. Agents Chemother.*, 59(10):6248–6255.

[Blasdel et al., 2017] Blasdel, G. B., Smet, D. J., Hendrix, H., Lavigne, R., and Danis-Wlodarczyk, K. (2017). *Pseudomonas* predators: understanding and exploiting phage-host interactions. *Nature Reviews Microbiology*.

[Bodenhofer et al., 2011] Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. *Bioinformatics*, 27(17):2463–2464.

[Boissy et al., 2011] Boissy, R., Ahmed, A., Janto, B., Earl, J., Hall, B. G., Hogg, J. S., Pusch, G. D., Hiller, L. N., Powell, E., Hayes, J., et al. (2011). Comparative supragenomic analyses among the pathogens staphylococcus aureus, streptococcus pneumoniae, and haemophilus influenzae using a modification of the finite supragenome model. *BMC genomics*, 12(1):1.

[Boisvert et al., 2010] Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. *Journal of Computational Biology*, 17(11):1519–1533.

[Bolotin et al., 2001] Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S. D., and Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. *Genome Res.*, 11(5):731–753.

[Bouckaert et al., 2014] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537.

[Breidenstein et al., 2011] Breidenstein, M. E. B., de la Fuente-Núñez, C., and Hancock, W. R. E. (2011). *Pseudomonas aeruginosa*: all roads lead to resistance. *Trends in microbiology*, 19(8):419–426.

[Brown, 2010] Brown, T. (2010). *Gene cloning and DNA analysis: an introduction*. John Wiley & Sons.

[Bruno et al., 2000] Bruno, W. J., Socci, N. D., and Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1):189–197.

[Bryant and Moulton, 2002] Bryant, D. and Moulton, V. (2002). NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In Guigó, R. and Gusfield, D., editors, *Algorithms in Bioinformatics*, volume 2452 of *Lecture Notes in Computer Science*, pages 375–391. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Cady et al., 2012] Cady, C. K., Bondy-Denomy, J., Heussler, E. G., Davidson, R. A., and O'Toole, A. G. (2012). The crispr/cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. *Journal of bacteriology*, 194(21):5728–5738.

[Carmeli et al., 1999] Carmeli, Y., Troillet, N., Eliopoulos, G. M., and Samore, M. H. (1999). Emergence of antibiotic-resistant *Pseudomonas aeruginosa*: comparison of risks associated with different antipseudomonal agents. *Antimicrob. Agents Chemother.*, 43(6):1379–1382.

[Chandler and Mahillon, 2002] Chandler, M. and Mahillon, J. (2002). Insertion sequences revisited. *Mobile DNA II*, pages 305–366.

[Chen et al., 2005] Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., and Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, 33(Database issue):D325–8.

[Cholley et al., 2014] Cholley, P., Ka, R., Guyeux, C., Thouverez, M., Guessennd, N., Ghebremedhin, B., Frank, T., Bertrand, X., and Hocquet, D. (2014). Population structure of clinical *Pseudomonas aeruginosa* from west and central african countries. *PLoS One*, 9(9):e107008.

[Craig et al., 2002] Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M. (2002). Mobile dna ii. american society for microbiology press. *Washington, DC*.

[Cramer et al., 2011] Cramer, N., Klockgether, J., Wrasman, K., Schmidt, M., Davenport, C. F., and Tümmler, B. (2011). Microevolution of the major common *Pseudomonas aeruginosa* clones c and pa14 in cystic fibrosis lungs. *Environmental microbiology*, 13(7):1690–1704.

[Curran et al., 2004] Curran, B., Jonas, D., Grundmann, H., Pitt, T., and Dowson, C. G. (2004). Development of a multilocus sequence typing scheme for the opportunistic pathogen *Pseudomonas aeruginosa*. *J. Clin. Microbiol.*, 42(12):5644–5649.

[Darriba et al., 2012] Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, 9(8):772.

[Daubin and Szöllősi, 2016] Daubin, V. and Szöllősi, J. G. (2016). Horizontal gene transfer and the history of life. *Cold Spring Harbor perspectives in biology*, 8(4).

[Delcher et al., 1999a] Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999a). Improved microbial gene identification with glimmer. *Nucleic acids research*, 27(23):4636–4641.

[Delcher et al., 1999b] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999b). Alignment of whole genomes. *Nucleic acids research*, 27(11):2369–2376.

[Denamur et al., 1993] Denamur, E., Picard, B., Decoux, G., Denis, J.-B., and Elion, J. (1993). The absence of correlation between allozyme and rrn rflp analysis indicates a high gene flow rate within human clinical *Pseudomonas aeruginosa* isolates. *FEMS Microbiology Letters*, 110(3):275.

[Dettman et al., 2013] Dettman, J. R., Rodrigue, N., Aaron, S. D., and Kassen, R. (2013). Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences*, 110(52):21065–21070.

[Eaton and Ree, 2013] Eaton, D. A. and Ree, R. H. (2013). Inferring phylogeny and introgression using radseq data: an example from flowering plants (pedicularis: Orobanchaceae). *Systematic Biology*, 62(5):689–706.

[Edelstein et al., 2013] Edelstein, M. V., Skleenova, E. N., Shevchenko, O. V., D'souza, J. W., Tapalski, D. V., Azizov, I. S., Sukhorukova, M. V., Pavlukov, R. A., Kozlov, R. S., Toleman, M. A., and Walsh, T. R. (2013). Spread of extensively resistant vim-2-positive st235 *Pseudomonas aeruginosa* in belarus, kazakhstan, and russia: a longitudinal epidemiological and clinical study. *The Lancet Infectious Diseases*, 13(10):867 – 876.

[Edgar, 2007] Edgar, R. C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, 8:18.

[Efron et al., 1996] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences*, 93(23):13429–13429.

[Ewing, 2015] Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. *Mobile DNA*, 6.

[Felsenstein, 1985] Felsenstein, J. (1985). Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol*, 39:783–791.

[Felsenstein, 1988] Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual review of genetics*, 22(1):521–565.

[Flicek and Birney, 2009] Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*. Thesis-NGS.

[Florek et al., 2014] Florek, M. C., Gilbert, D. P., and Plague, G. R. (2014). Insertion sequence distribution bias in archaea. *Mobile Genetic Elements*, 4(1):e27829.

[Fonseca et al., 2012] Fonseca, A. N., Rung, J., Brazma, A., and Marioni, C. J. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24).

[Fournier et al., 2013] Fournier, D., Richardot, C., Müller, E., Robert-Nicoud, M., Llanes, C., Plésiat, P., and Jeannot, K. (2013). Complexity of resistance mechanisms to imipenem in intensive care unit strains of *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.*, 68(8):1772–1780.

[Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315(5814):972–976.

[Frost et al., 2005] Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732.

[García-Castillo et al., 2012] García-Castillo, M., Máiz, L., Morosini, M.-I., Rodríguez-Baños, M., Suarez, L., Fernández-Olmos, A., Baquero, F., Cantón,

R., and del Campo, R. (2012). Emergence of a mutl mutation causing multilocus sequence typing–pulsed-field gel electrophoresis discrepancy among *Pseudomonas aeruginosa* isolates from a cystic fibrosis patient. *Journal of Clinical Microbiology*, 50(5):1777–1778.

[Garrison and Marth, 2012] Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*. arXiv: 1207.3907.

[Gellatly and Hancock, 2013] Gellatly, S. L. and Hancock, R. E. W. (2013). *Pseudomonas aeruginosa*: new insights into pathogenesis and host defenses. *Pathog. Dis.*, 67(3):159–173.

[Godfrey-Smith and Sterelny, 2008] Godfrey-Smith, P. and Sterelny, K. (2008). Biological information. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

[Gooderham and Hancock, 2009] Gooderham, W. J. and Hancock, R. E. W. (2009). Regulation of virulence and antibiotic resistance by two-component regulatory systems in *Pseudomonas aeruginosa*. *FEMS Microbiology Reviews*, 33(2):279.

[Grosso-Becerra et al., 2014] Grosso-Becerra, M.-V., Santos-Medellín, C., González-Valdez, A., Méndez, J.-L., Delgado, G., Morales-Espinosa, R., Servín-González, L., Alcaraz, L.-D., and Soberón-Chávez, G. (2014). *Pseudomonas aeruginosa* clinical and environmental isolates constitute a single population with high phenotypic diversity. *BMC genomics*, 15(1):318.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696–704.

[Hauser, 2009] Hauser, A. R. (2009). The type iii secretion system of *Pseudomonas aeruginosa*: infection by injection. *Nature Reviews Microbiology*, 7(9):654–665.

[Hawkey et al., 2015] Hawkey, J., Hamidian, M., Wick, R. R., Edwards, D. J., Billman-Jacobe, H., Hall, R. M., and Holt, K. E. (2015). Ismapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics*, 16(1):667.

[He et al., 2013] He, M., Miyajima, F., Roberts, P., Ellison, L., Pickard, D. J., Martin, M. J., Connor, T. R., Harris, S. R., Fairley, D., Bamford, K. B., D'Arc, S., Brazier, J., Brown, D., Coia, J. E., Douce, G., Gerding, D., Kim, H. J., Koh, T. H., Kato, H., Senoh, M., Louie, T., Michell, S., Butt, E., Peacock, S. J., Brown, N. M., Riley, T., Songer, G., Wilcox, M., Pirmohamed, M., Kuijper, E., Hawkey, P., Wren, B. W., Dougan, G., Parkhill, J., and Lawley, T. D. (2013). Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.*, 45(1):109–113.

[Heger and contributors, 2009] Heger, Andreas, J.-K. and contributors (2009). pysam: htslib interface for python. https://github.com/pysam-developers/pysam.

[Henson et al., 2012] Henson, J., Tischler, G., and Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, 13(8):901–915.

[Hickman et al., 2010] Hickman, A. B., Chandler, M., and Dyda, F. (2010). Integrating prokaryotes and eukaryotes: Dna transposases in light of structure. *Critical reviews in biochemistry and molecular biology*, 45(1):50–69.

[Higgins and Sharp, 1988] Higgins, D. G. and Sharp, P. M. (1988). Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244.

[Hilker et al., 2015] Hilker, R., Munder, A., Klockgether, J., Losada, P. M., Chouvarine, P., Cramer, N., Davenport, C. F., Dethlefsen, S., Fischer, S., Peng, H., Schönfelder, T., Türk, O., Wiehlmann, L., Wölbeling, F., Gulbins, E., Goesmann, A., and Tümmler, B. (2015). Interclonal gradient of virulence in the *Pseudomonas aeruginosa* pangenome from disease and environment. *Environmental Microbiology*, 17(1):29–46.

[Ho and Phillips, 2009] Ho, W. S. Y. and Phillips, J. M. (2009). Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic biology*, 58(3):367–380. Thesis-TMRCA.

[Hocquet et al., 2012] Hocquet, D., Llanes, C., Thouverez, M., Kulasekara, H. D., Bertrand, X., Plésiat, P., Mazel, D., and Miller, S. I. (2012). Evidence for induction of integron-based antibiotic resistance by the SOS response in a clinical setting. *PLoS Pathog.*, 8(6):e1002778.

[Holden et al., 2013] Holden, M. T. G., Hsu, L.-Y., Kurt, K., Weinert, L. A., Mather, A. E., Harris, S. R., Strommenger, B., Layer, F., Witte, W., de Lencastre, H., Skov, R., Westh, H., Zemlicková, H., Coombs, G., Kearns, A. M., Hill, R. L. R., Edgeworth, J., Gould, I., Gant, V., Cooke, J., Edwards, G. F., McAdam, P. R., Templeton, K. E., McCann, A., Zhou, Z., Castillo-Ramírez, S., Feil, E. J., Hudson, L. O., Enright, M. C., Balloux, F., Aanensen, D. M., Spratt, B. G., Fitzgerald, J. R., Parkhill, J., Achtman, M., Bentley, S. D., and Nübel, U. (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Res.*, 23(4):653–664.

[Homer, 2014] Homer, N. (2014). Dwgsim - whole genome simulator for next-generation sequencing. https://github.com/nh13/DWGSIM.

[Huelsenbeck and Ronquist, 2001] Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

[Huelsenbeck et al., 2001] Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314.

[Huson and Bryant, 2006] Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23(2):254–267.

[Huson et al., 2011] Huson, D. H., Rupp, R., and Scornavacca, C. (2011). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge Univ Pr.

[Hyatt et al., 2010] 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(1):119.

[Jeukens et al., 2014] Jeukens, J., Boyle, B., Kukavica-Ibrulj, I., Ouellet, M. M., Aaron, S. D., Charette, S. J., Fothergill, J. L., Tucker, N. P., Winstanley, C., and Levesque, R. C. (2014). Comparative genomics of isolates of a *Pseudomonas aeruginosa* epidemic strain associated with chronic lung infections of cystic fibrosis patients. *PloS one*, 9(2):e87611.

[Johnston et al., 2014] Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.-P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.*, 12(3):181–196.

[Keane et al., 2012] Keane, T. M., Wong, K., and Adams, D. J. (2012). Retroseq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, 29(3):389–390.

[Kent, 2002] Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664.

[Kidd et al., 2012] Kidd, T. J., Ritchie, S. R., Ramsay, K. A., Grimwood, K., Bell, S. C., and Rainey, P. B. (2012). *Pseudomonas aeruginosa* exhibits frequent recombination, but only a limited association between genotype and ecological setting. *PLoS One*, 7(9):e44199.

[Kiewitz and Tümmler, 2000] Kiewitz, C. and Tümmler, B. (2000). Sequence diversity of *Pseudomonas aeruginosa*: Impact on population structure and genome evolution. *Journal of Bacteriology*, 182(11):3125–3135.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

[Klockgether et al., 2011] Klockgether, J., Cramer, N., Wiehlmann, L., Davenport, C. F., and Tümmler, B. (2011). *Pseudomonas aeruginosa* genomic structure and diversity. *Frontiers in microbiology*, 2.

[Kolmogorov et al., 2014] Kolmogorov, M., Raney, B., Paten, B., and Pham, S. (2014). Ragout-a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, 30(12):i302–9.

[Kos et al., 2015] Kos, V. N., Déraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., Corbeil, J., and Gardner, H. (2015). The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrob. Agents Chemother.*, 59(1):427–436.

[Kroon et al., 2015] Kroon, M., Lameijer, E.-W., Lakenberg, N., Hehir-Kwa, J. Y., Thung, D., Slagboom, P. E., Kok, J. N., and Ye, K. (2015). Detecting dispersed duplications in high-throughput sequencing data using a database-free approach. *Bioinformatics*, 32(4):505–510.

[Kugelberg et al., 2005] Kugelberg, E., Löfmark, S., Wretlind, B., and Andersson, D. I. (2005). Reduction of the fitness burden of quinolone resistance in *Pseudomonas aeruginosa*. *J. Antimicrob. Chemother.*, 55(1):22–30.

[Kung et al., 2010] Kung, L. V., Ozer, A. E., and Hauser, R. A. (2010). The accessory genome of *Pseudomonas aeruginosa*. *Microbiology and molecular biology reviews: MMBR*, 74(4):621–641.

[Kurtz et al., 2004] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2):R12.

[Larsen et al., 2012] Larsen, V. M., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, L. R., Jelsbak, L., Sicheritz-Pontén, T., Ussery, W. D., Aarestrup, M. F., and Lund, O. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of clinical microbiology*, 50(4):1355–1361.

[Lawe-Davies and Bennet, 2017] Lawe-Davies, O. and Bennet, S. (2017). World health organization: Who. "http://www.who.int/mediacentre/news/releases/2017/bacteria-antibiotics-needed/en/".

[Lemey et al., 2009] Lemey, P., Salemi, M., and Vandamme, A.-M., editors (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li and Homer, 2010] Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483.

[Linares et al., 2005] Linares, J. F., López, J. A., Camafeita, E., Albar, J. P., Rojo, F., and Martínez, J. L. (2005). Overexpression of the multidrug efflux pumps mexcd-oprj and mexef-oprn is associated with a reduction of type iii secretion in *Pseudomonas aeruginosa*. *Journal of Bacteriology*, 187(4):1384–1391.

[Lister et al., 2009] Lister, D. P., Wolter, J. D., and Hanson, D. N. (2009). Antibacterial-resistant *Pseudomonas aeruginosa*: clinical impact and complex regulation of chromosomally encoded resistance mechanisms. *Clinical microbiology reviews*, 22(4):582–610.

[Livermore, 2002] Livermore, D. M. (2002). Multiple mechanisms of antimicrobial resistance in *Pseudomonas aeruginosa*: our worst nightmare? *Clin. Infect. Dis.*, 34(5):634–640.

[Livermore, 2009] Livermore, D. M. (2009). Has the era of untreatable infections arrived? *Journal of Antimicrobial Chemotherapy*, 64(1):i29.

[Loman et al., 2012] Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., and Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Micro*, 10(9):599–606.

[Loman and Pallen, 2015] Loman, N. J. and Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. *Nat Rev Micro*, 13(12):787–794.

[Lyczak et al., 2002] Lyczak, J. B., Cannon, C. L., and Pier, G. B. (2002). Lung infections associated with cystic fibrosis. *Clin. Microbiol. Rev.*, 15(2):194–222.

[Maatallah et al., 2011] Maatallah, M., Cheriaa, J., Backhrouf, A., Iversen, A., Grundmann, H., Do, T., Lanotte, P., Mastouri, M., Elghmati, M. S., Rojo, F., Mejdi, S., and Giske, C. G. (2011). Population structure of *Pseudomonas aeruginosa* from five mediterranean countries: evidence for frequent recombination and epidemic occurrence of CC235. *PLoS One*, 6(10):e25617.

[Maciá et al., 2005] Maciá, M. D., Blanquer, D., Togores, B., Sauleda, J., Pérez, J. L., and Oliver, A. (2005). Hypermutation is a key factor in development of multiple-antimicrobial resistance in *Pseudomonas aeruginosa* strains causing chronic lung infections. *Antimicrobial Agents and Chemotherapy*, 49(8):3382–3386.

[Magoc et al., 2013] Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, J. L., and Salzberg, L. S. (2013). Gage-b: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29(14).

[Mahillon and Chandler, 1998] Mahillon, J. and Chandler, M. (1998). Insertion sequences. *Microbiology and Molecular Biology Reviews*, 62(3):725–774.

[Martínez-Ramos et al., 2014] Martínez-Ramos, I., Mulet, X., Moyá, B., Barbier, M., Oliver, A., and Albertí, S. (2014). Overexpression of mexcd-oprj reduces *Pseudomonas aeruginosa* virulence by increasing its susceptibility to complement-mediated killing. *Antimicrobial agents and chemotherapy*, 58(4):2426–2429.

[Martínez and Baquero, 2002] Martínez, J. L. and Baquero, F. (2002). Interactions among strategies associated with bacterial infection: Pathogenicity, epidemicity, and antibiotic resistance. *Clinical Microbiology Reviews*, 15(4):647–679.

[Mathee et al., 2008] Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., Rokas, A., Yandava, C. N., Engels, R., Zeng, E., et al. (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proceedings of the National Academy of Sciences*, 105(8):3100–3105.

[Mena et al., 2008] Mena, A., Smith, E. E., Burns, J. L., Speert, D. P., Moskowitz, S. M., Perez, J. L., and Oliver, A. (2008). Genetic adaptation of *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients is catalyzed by hypermutation. *Journal of Bacteriology*, 190(24):7910–7917.

[Mercier et al., 2013] Mercier, C., Boyer, F., Kopylova, E., Taberlet, P., Bonin, A., and Coissac, E. (2013). sumaclust: Fast and exact clustering of sequences. https://git.metabarcoding.org/obitools/sumaclust.

[Mesaros et al., 2007] Mesaros, N., Nordmann, P., Plésiat, P., Roussel-Delvallez, M., Eldere, J. V., Glupczynski, Y., Laethem, Y. V., Jacobs, F., Lebecque, P., Malfroot, A., Tulkens, P., and Bambeke, F. V. (2007). *Pseudomonas aeruginosa*: resistance and therapeutic options at the turn of the new millennium. *Clinical Microbiology and Infection*, 13(6):560 – 578.

[Moradali et al., 2017] Moradali, F. M., Ghods, S., and Rehm, A. B. H. (2017). Lifestyle: A paradigm for adaptation, survival, and persistence. *Frontiers in cellular and infection microbiology*, 7.

[Müller and Vaughan, 2017] Müller, N. F. and Vaughan, T. (2017). *Structured coalescent: Population structure using MultiTypeTree*.

[Naas et al., 2013] Naas, T., Bonnin, A. R., Cuzon, G., Villegas, M.-V., and Nordmann, P. (2013). Complete sequence of two kpc-harbouring plasmids from *Pseudomonas aeruginosa*. *The Journal of antimicrobial chemotherapy*, 68(8):1757–1762.

[Nei and Gojobori, 1986] Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418–426.

[Notredame et al., 2000] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217.

[Ochiai et al., 2005] Ochiai, H., Inoue, Y., Takeya, M., Sasaki, A., and Kaku, H. (2005). Genome sequence of xanthomonas oryzae pv. oryzae suggests contribution of large numbers of effector genes and insertion sequences to its race diversity. *Japan Agricultural Research Quarterly*, 39(4):275.

[Oliver et al., 2002] Oliver, A., Baquero, F., and Blázquez, J. (2002). The mismatch repair system (muts, mutl and uvrd genes) in *Pseudomonas aeruginosa*: molecular characterization of naturally occurring mutants. *Molecular Microbiology*, 43(6):1641–1650.

[Oliver et al., 2000] Oliver, A., Cantón, R., Campo, P., Baquero, F., and Blázquez, J. (2000). High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science*, 288(5469):1251–1253.

[Oliver et al., 2008] Oliver, A., Mena, A., and Maciá, M. D. (2008). Evolution of *Pseudomonas aeruginosa* pathogenicity: From acute to chronic infections. In *Evolutionary Biology of Bacterial and Fungal Pathogens*, pages 433–444. American Society of Microbiology.

[Oliver et al., 2015] Oliver, A., Mulet, X., López-Causapé, C., and Juan, C. (2015). The increasing threat of *Pseudomonas aeruginosa* high-risk clones. *Drug Resist. Updat.*, 21-22:41–59.

[Olliver et al., 2005] Olliver, A., Vallé, M., Chaslus-Dancla, E., and Cloeckaert, A. (2005). Overexpression of the multidrug efflux operon acref by insertional activation with is1 or is10 elements in salmonella enterica serovar typhimurium dt204 acrb mutants selected with fluoroquinolones. *Antimicrobial agents and chemotherapy*, 49(1):289–301.

[Ooka et al., 2009] Ooka, T., Ogura, Y., Asadulghani, M., Ohnishi, M., Nakayama, K., Terajima, J., Watanabe, H., and Hayashi, T. (2009). Inference of the impact of insertion sequence (is) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in escherichia coli o157 genomes. *Genome research*, 19(10):1809–1816.

[Parkhill et al., 2003] Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., Holden, M. T., Churcher, C. M., Bentley, S. D., Mungall, K. L., et al. (2003). Comparative analysis of the genome sequences of bordetella pertussis, bordetella parapertussis and bordetella bronchiseptica. *Nature genetics*, 35(1):32–40.

[Peña et al., 2012] Peña, C., Suarez, C., Gozalo, M., Murillas, J., Almirante, B., Pomar, V., Aguilar, M., Granados, A., Calbo, E., Rodríguez-Baño, J., et al. (2012). Prospective multicenter study of the impact of carbapenem resistance on mortality in *Pseudomonas aeruginosa* bloodstream infections. *Antimicrobial agents and chemotherapy*, 56(3):1265–1272.

[Pevsner, 2005] Pevsner, J. (2005). *Bioinformatics and functional genomics*. John Wiley & Sons.

[Picard et al., 1994] Picard, B., Denamur, E., Barakat, A., Elion, J., and Goullet, P. (1994). Genetic heterogeneity of *Pseudomonas aeruginosa* clinical isolates revealed by esterase electrophoretic polymorphism and restriction fragment length polymorphism of the ribosomal rna gene region. *Journal of Medical Microbiology*, 40(5):313–322.

[Pirnay et al., 2002] Pirnay, J.-P., De Vos, D., Cochez, C., Bilocq, F., Vanderkelen, A., Zizi, M., Ghysels, B., and Cornelis, P. (2002). *Pseudomonas aeruginosa* displays an epidemic population structure. *Environ. Microbiol.*, 4(12):898–911.

[Plague, 2010] Plague, G. R. (2010). Intergenic transposable elements are not randomly distributed in bacteria. *Genome biology and evolution*, 2:584–590.

[Poole, 2011] Poole, K. (2011). *Pseudomonas aeruginosa*: resistance to the max. *Front. Microbiol.*, 2:65.

[Posada, 2008] Posada, D. (2008). jmodeltest: phylogenetic model averaging. *Molecular biology and evolution*, 25(7):1253–1256.

[Potron et al., 2015] Potron, A., Poirel, L., and Nordmann, P. (2015). Emerging broad-spectrum resistance in *Pseudomonas aeruginosa* and *Acinetobacter baumannii*: Mechanisms and epidemiology. *Int. J. Antimicrob. Agents*, 45(6):568–585.

[Rambaut, 2014] Rambaut, A. (2014). FigTree, a graphical viewer of phylogenetic trees. http://tree.bio.ed.ac.uk/software/figtree/. Accessed: 2015-NA-NA.

[Ray et al., 2007] Ray, D. A., Pagan, H. J., Thompson, M. L., and Stevens, R. D. (2007). Bats with hats: evidence for recent dna transposon activity in genus myotis. *Molecular biology and evolution*, 24(3):632–639.

[Rice et al., 2000] Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–7.

[Robinson et al., 2012] Robinson, D. G., Lee, M.-C., and Marx, C. J. (2012). Oasis: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic acids research*, 40(22):e174–e174.

[Roy et al., 2010] Roy, P. H., Tetu, S. G., Larouche, A., Elbourne, L., Tremblay, S., Ren, Q., Dodson, R., Harkins, D., Shay, R., Watkins, K., et al. (2010). Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* pa7. *PloS one*, 5(1):e8842.

[Rubinstein, 2001] Rubinstein, E. (2001). History of quinolones and their side effects. *Chemotherapy*, 47 Suppl 3:3–8; discussion 44–8.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

[Samson et al., 2013] Samson, E. J., Magadán, H. A., Sabri, M., and Moineau, S. (2013). Revenge of the phages: defeating bacterial defences. *Nature reviews. Microbiology*, 11(10):675–687.

[Sato et al., 2003] Sato, H., Frank, D. W., Hillard, C. J., Feix, J. B., Pankhaniya, R. R., Moriyama, K., Finck-Barbançon, V., Buchaklian, A., Lei, M., Long, R. M., Wiener-Kronish, J., and Sawa, T. (2003). The mechanism of action of the *Pseudomonas aeruginosa*-encoded type III cytotoxin, ExoU. *EMBO J.*, 22(12):2959–2969.

[Sawa et al., 2014] Sawa, T., Shimizu, M., Moriyama, K., and Wiener-Kronish, J. P. (2014). Association between *Pseudomonas aeruginosa* type iii secretion, antibiotic resistance, and clinical outcome: a review. *Critical Care*, 18(6):668.

[Schbath et al., 2012] Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of computational biology: a journal of computational molecular cell biology*, 19(6):796–813.

[Schmidtke and Hanson, 2008] Schmidtke, A. J. and Hanson, N. D. (2008). Role of *ampD* homologs in overproduction of AmpC in clinical isolates of *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.*, 52(11):3922–3927.

[Siguier et al., 2006a] Siguier, P., Filée, J., and Chandler, M. (2006a). Insertion sequences in prokaryotic genomes. *Current opinion in microbiology*, 9(5):526–531.

[Siguier et al., 2014] Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, 38(5):865–891.

[Siguier et al., 2015] Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., and Chandler, M. (2015). Everyman's guide to bacterial insertion sequences. *Microbiology Spectrum*, 3(2).

[Siguier et al., 2006b] Siguier, P., Pérochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006b). Isfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl 1):D32–D36.

[Siguier et al., 2012] Siguier, P., Varani, A., Perochon, J., and Chandler, M. (2012). Exploring bacterial insertion sequences with isfinder: objectives, uses, and future developments. In *Mobile Genetic Elements*, pages 91–103. Springer.

[Silby et al., 2011] Silby, M. W., Winstanley, C., Godfrey, S. A., Levy, S. B., and Jackson, R. W. (2011). *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiology Reviews*, 35(4):652.

[Simpson and Pop, 2015] Simpson, T. J. and Pop, M. (2015). The theory and practice of genome sequence assembly. *Annual review of genomics and human genetics*, 16:153–172.

[Skurnik et al., 2013] Skurnik, D., Roux, D., Cattoir, V., Danilchanka, O., Lu, X., Yoder-Himes, D. R., Han, K., Guillard, T., Jiang, D., Gaultier, C., Guerin, F., Aschard, H., Leclercq, R., Mekalanos, J. J., Lory, S., and Pier, G. B. (2013). Enhanced in vivo fitness of carbapenem-resistant oprd mutants of *Pseudomonas aeruginosa* revealed through high-throughput sequencing. *Proceedings of the National Academy of Sciences*, 110(51):20747–20752.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.

[Soltis et al., 1999] Soltis, P. S., Soltis, D. E., and Chase, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*, 402(6760):402–404.

[Soltis et al., 2003] Soltis, P. S., Soltis, D. E., et al. (2003). Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18(2):256–267.

[Summers et al., 2005] Summers, O. A., Leplae, R., Frost, S. L., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9).

[Sun et al., 2016] Sun, Q., Ba, Z., Wu, G., Wang, W., Lin, S., and Yang, H. (2016). Insertion sequence *ISRP*10 inactivation of the *oprD* gene in imipenem-resistant *Pseudomonas aeruginosa* clinical isolates. *International Journal of Antimicrobial Agents*, 47(5):375–379.

[Suzuki and Shimodaira, 2006] Suzuki, R. and Shimodaira, H. (2006). Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12). Thesis-resistance clustering.

[Tamura and Nei, 1993] Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526.

[Tateno et al., 1994] Tateno, Y., Takezaki, N., and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution*, 11(2):261–277.

[Tavaré, 1984] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, 26(2):119–164.

[Thung et al., 2014] Thung, D. T., de Ligt, J., Vissers, L. E., Steehouwer, M., Kroon, M., de Vries, P., Slagboom, E. P., Ye, K., Veltman, J. A., and Hehir-Kwa, J. Y. (2014). Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome biology*, 15(10):488.

[Touchon et al., 2012] Touchon, M., Charpentier, S., Pognard, D., Picard, B., Arlet, G., Rocha, E. P. C., Denamur, E., and Branger, C. (2012). Antibiotic resistance plasmids spread among natural isolates of *Escherichia coli* in spite of CRISPR elements. *Microbiology*, 158(Pt 12):2997–3004.

[Touchon et al., 2009] Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. *PLoS genet*, 5(1):e1000344.

[Touchon and Rocha, 2007] Touchon, M. and Rocha, E. P. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Molecular biology and evolution*, 24(4):969–981.

[Turton et al., 2015] Turton, J. F., Wright, L., Underwood, A., Witney, A. A., Chan, Y.-T., Al-Shahib, A., Arnold, C., Doumith, M., Patel, B., Planche, T. D., Green, J., Holliman, R., and Woodford, N. (2015). High-Resolution analysis by Whole-Genome sequencing of an international lineage (sequence type 111) of *Pseudomonas aeruginosa* associated with Metallo-Carbapenemases in the united kingdom. *J. Clin. Microbiol.*, 53(8):2622–2631.

[Valot et al., 2015] Valot, B., Guyeux, C., Rolland, J. Y., Mazouzi, K., Bertrand, X., and Hocquet, D. (2015). What it takes to be a *Pseudomonas aeruginosa*? the core genome of the opportunistic pathogen updated. *PloS one*, 10(5):e0126468.

[van Belkum et al., 2015] van Belkum, A., Soriaga, L. B., LaFave, M. C., Akella, S., Veyrieras, J.-B., Barbu, E. M., Shortridge, D., Blanc, B., Hannum, G., Zambardi, G., Miller, K., Enright, M. C., Mugnier, N., Brami, D., Schicklin, S., Felderman, M., Schwartz, A. S., Richardson, T. H., Peterson, T. C., Hubby, B., and Cady, K. C. (2015). Phylogenetic distribution of CRISPR-Cas systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *MBio*, 6(6):e01796–15.

[Vandecraen et al., 2017] Vandecraen, J., Chandler, M., Aertsen, A., and Van Houdt, R. (2017). The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*, pages 1–22.

[Varani et al., 2011] Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., and Chandler, M. (2011). Issaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol*, 12(3):R30.

[Vincent, 2003] Vincent, J.-L. (2003). Nosocomial infections in adult intensive-care units. *The Lancet*, 361(9374):2068 – 2077.

[von Wintersdorff et al., 2016] von Wintersdorff, H. C. J., Penders, J., van Niekerk, M. J., Mills, D. N., Majumder, S., van Alphen, B. L., Savelkoul, M. P. H., and Wolffs, G. P. F. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in microbiology*, 7.

[Wagner et al., 2007] Wagner, A., Lewis, C., and Bichsel, M. (2007). A survey of bacterial insertion sequences using iscan. *Nucleic Acids Research*, 35(16):5284–5293.

[Wilson and Sarich, 1969] Wilson, A. C. and Sarich, V. M. (1969). A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences*, 63(4):1088–1093.

[Wolfgang et al., 2003] Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C. G., and Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proceedings of the National Academy of Sciences*, 100(14):8484–8489.

[Woodford et al., 2011] Woodford, N., Turton, J. F., and Livermore, D. M. (2011). Multiresistant gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.*, 35(5):736–755.

[Wozniak and Waldor, 2010] Wozniak, F. R. A. and Waldor, K. M. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature reviews. Microbiology*, 8(8):552–563.

[Wu and Watanabe, 2005] Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.

[Yada and Hirosawa, 1996] Yada, T. and Hirosawa, M. (1996). Detection of short protein coding regions within the cyanobacterium genome: application of the hidden markov model. *DNA Research*, 3(6):355–361.

[Yang, 1994] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39(3):306–314.

[Zankari et al., 2012] Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., and Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, 67(11):2640–2644.

[Zeng and Jin, 2003] Zeng, L. and Jin, S. (2003). *aph(3')-IIb*, a gene encoding an aminoglycoside-modifying enzyme, is under the positive control of surrogate regulator HpaA. *Antimicrob. Agents Chemother.*, 47(12):3867–3876.

[Zhang et al., 2011] Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one*, 6(3).

[Zhou et al., 2008] Zhou, F., Olman, V., and Xu, Y. (2008). Insertion sequences show diverse recent activities in cyanobacteria and archaea. *BMC genomics*, 9(1):36.

[Zimin et al., 2013] Zimin, V. A., Marçais, G., Puiu, D., Roberts, M., Salzberg, L. S., and Yorke, A. J. (2013). The masurca genome assembler. *Bioinformatics (Oxford, England)*, 29(21):2669–2677.

List of Figures

List of Tables

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale sciences pour l'ingénieur et microtechniques

Title: Bioinformatic analysis of the genomes of epidemic *Pseudomonas aeruginosa*

Keywords: Genomics, Bioinformatics, Bacteriology, *Pseudomonas aeruginosa*

Abstract:

Pseudomonas aeruginosa is a major nosocomial pathogen with ST235 being the most prevalent of the so-called 'international' or 'high-risk' clones. This clone is associated with poor clinical outcomes in part due to multi- and high-level antibiotic resistance. Despite its clinical importance, the molecular basis for the success of the ST235 clone is poorly understood. This study aimed at understanding the spatiotemporal origin of the clone and the molecular basis of its success. Using an analysis pipeline of WGS data, we found that ST235 clone emerged in Europe around 1984, and that all the ST235 isolates produced the ExoU exotoxin. We also identified 22 ST235-specific genes clustering in blocks and implicated in transmembrane efflux, DNA processing and bacterial transformation. This unique combination of genes may have contributed to the poor outcome associated with *P. aeruginosa* ST235 infections and increased the ability of this international clone to acquire mobile resistance elements. ST235 has presumably become prevalent across the globe potentially due to the selective pressure of fluoroquinolones and readily became resistant to aminoglycosides, β-lactams, and carbapenems through mutation and acquisition of resistance elements among local populations. For this analysis, we mostly used existing tools but found that programs dedicated to the detection of insertion sequences (IS) - that are important drivers of bacterial evolution - were not adapted to bacteria. We then developed and optimized the *panISa* program, a sensitive and highly precise tool for detection IS from raw sequencing data of bacterial genomes.

Titre : Analyse bioinformatique des génomes d'une souche épidémique de *Pseudomonas aeruginosa*

Mots-clés : Génomique, Bio-informatique, Bactériologie, *Pseudomonas aeruginosa*

Résumé :

Le *Pseudomonas aeruginosa* est un pathogène nosocomial majeur. Le clone ST235 est le plus prévalent des clones internationaux dits à haut risque. Ce clone est très fréquemment multi-résistant aux antibiotiques, ce qui complique la prise en charge des infections dont il est à l'origine. Malgré son importance clinique, la base moléculaire du succès du clone ST235 n'est pas comprise. Dans ce travail, nous avons cherché à comprendre l'origine spaciotemporelle de ce clone et les bases moléculaires de son succès. A l'aide d'outils bioinformatiques existants, nous avons trouvé que le clone ST235 a émergé en Europe en 1984 et que tous les isolates ST235 produisent l'exotoxine ExoU. Nous avons également identifié 22 gènes contigus spécifiques de ce clones et impliqués dans l'efflux transmembranaire, dans le traitement de l'ADN et dans la transformation bactérienne. Cette combinaison unique de gènes a pu contribuer à la gravité des infections dues à ce clone et à sa capacité à acquérir des gènes de résistance aux antibiotiques. Ainsi, la diffusion mondiale de ce clone a probablement été favorisée par l'utilisation extensive des fluoroquinolones, puis il est devenu localement résistant aux aminoglycosides, aux β-lactamines, et aux carbapénèmes par mutation et acquisition d'éléments de résistance. Nous avons majoritairement utilisé des outils existants, mais avons découvert que les programmes de détection des séquences d'insertions (IS, ayant un rôle important dans l'évolution des génomes bactériens) ne sont pas adaptés aux données dont nous disposions. Nous avons ainsi mis au point un outil (appelé *panISa*) qui détecte de façon précise et sensible les IS à partir de données brutes de séquençage de génomes bactériens.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besanon, France