UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Environmental Sciences

XXIX Cycle

Evolution, comparative genomics and genomic epidemiology of bacteria of public health importance

PhD Thesis

Stefano GAIARSA

R10468

Scientific tutor: Prof. Claudio BANDI

Academic Year: 2016-2017

SSD: BIO/19

Thesis performed at: Department of Biosciences, University of Milan, Italy; Microbiology and Virology Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; Department of Biology and Biotechnology, University of Pavia, Italy.

ABSTRACT IN ENGLISH

The present thesis is focused on genomic epidemiology of bacterial hospital infections. The hospital environment is unique, as it concentrates a high number of bacterial agents, frequent antibiotic use, and patients with weak immune systems. This combination favours the development and selection of antibiotic resistant strains and the spread of opportunistic infections: in general the thriving of nosocomial pathogens. Genomics and evolutionary approaches have emerged as the cutting edge tools for studying this kind of infections, allowing to study the genomic features of bacterial strains and their evolution. Thanks to the possibility to sequence DNA at a constantly cheaper price, research projects are supported by a growing number of genomes and a considerable amount of genomic data is available in the databases, expanding the amount of possible investigations that can be performed.

The first work presented here describes the evolution of the Clonal Complex 258 (CC258) of *Klebsiella pneumoniae*. Single nucleotide polymorphisms (SNPs) allowed to reconstruct the global phylogeny of the entire species and to collocate the CC258 in its evolutionary context. Furthermore, it was possible to detect the presence of a 1.3 Mb recombination in the genomes of the clade in analysis. A molecular clock approach allowed to date this and other previously discovered recombination events. These findings were used to complete the picture of the evolutionary history of CC258, which is characterized by frequent macro-recombination events. A quick evolutive strategy characterized by exchange of high amount of information is a common feature to other nosocomial pathogens, which develop "superbug" phenotypes.

Although common, the macro-recombination evolution model is not shared by all nosocomial infection bacteria. One exception is the SMAL strain of *Acinetobacter baumannii*, presented in another subproject of this thesis. In this work, the genomes of Sequence Type (ST) 78 of *A. baumannii* were analyzed. Phylogeny and comparative genomics revealed the presence of two different clades within the ST, presenting different evolutive "lifestyles". One group (containing the SMAL genomes) was characterized by a lower gene content variability and by the presence of a higher copy number of insertion sequences (ISs). One IS interrupts the *comEC/rec2* gene in all the SMAL genomes. This gene codes for a protein involved in the exogenous DNA importation, thus its inactivation limits the gene exchange, suggesting an explanation for the low genomic plasticity.

In another work presented in this document, genomic epidemiology was applied to reconstruct the spreading routes of a *K. pneumoniae* epidemic event in an hospital intensive care unit. At first, a phylogenetic approach was used to separate the isolates that belonged

to the outbreak from the sporadic ones. Then the isolation dates and genomic SNPs allowed to build a genomic network, which modelled the chain of infection events in the ward. The reconstruction suggested a star-like diffusion of the pathogen from patient zero to the other infected ones, thus revealing a systematic error in the biosafety procedures of the hospital.

This almost-forensic application of genomic epidemiology was also used in two other works presented, both of them concerning the reconstruction of food-borne infections. In one of the works, focused on *Salmonella enterica,* only synonymous SNPs were used as input to a phylogenetic based investigation, in order to filter out pathoadaptative mutations. In the other article, epidemiological data, molecular typing and SNP-based phylogeny were used to investigate the infection of nine *Listeria monocytogenes* isolates, which were believed to be part of the same outbreak and in the end proved to be genomically unrelated.

Lastly, a review paper on genomic epidemiology is also presented. The article is focused on the latest high impact publications analyzing the genome evolution of bacterial pathogens as well as the propagation dynamics of epidemic outbreaks in very short periods of time. The article also describes the latest historical epidemiological studies, which are possible thanks to modern DNA isolation and sequencing technologies.

ABSTRACT IN ITALIANO

La presente tesi è incentrata sull'epidemiologia genomica delle infezioni batteriche ospedaliere. L'ambiente ospedaliero è peculiare, in quanto al suo interno si concentrano un elevato numero di agenti batterici, pazienti con un sistema immunitario debole e un uso massiccio di sostanze antimicrobiche. Questa combinazione favorisce lo sviluppo e la selezione di ceppi resistenti agli antibiotici e la diffusione di infezioni opportunistiche: in generale il prosperare dei patogeni nosocomiali. Alcune tecniche all'avanguardia per lo studio di questo tipo di infezioni sono basate sull'uso della genomica e di approcci evoluzionistici: esse permettono di conoscere le caratteristiche genomiche dei ceppi batterici e di ricostruire la loro storia evolutiva. Grazie alla possibilità di sequenziare il DNA ad un prezzo sempre più economico, i progetti di ricerca sono supportati da un numero sempre crescente di genomi e i dati genomici depositati nelle banche dati sono in crescita esponenziale: questo rende possibile eseguire una varietà sempre maggiore di analisi.

Il primo lavoro qui riportato descrive l'evoluzione del Clonal Complex 258 (CC258) di *Klebsiella pneumoniae*. Le mutazioni puntiformi (single nucleotide polymorphism, SNP) hanno permesso di ricostruire la filogenesi globale di tutta la specie e di collocare il CC258 nel suo contesto evolutivo. Successivamente, è stato possibile rilevare la presenza di una ricombinazione di 1,3 Mb nei genomi del clade in analisi. Un'analisi del molecular clock ha poi consentito di datare sia questo che gli altri eventi di ricombinazione scoperti in lavori precedenti. Questi risultati sono stati usati per completare il quadro della storia evolutiva del CC258, caratterizzata da frequenti eventi di macro-ricombinazione. Un'evoluzione rapida e caratterizzata da scambi di elevate quantità di informazioni genomiche è una caratteristica comune ad altri patogeni nosocomiali che sviluppano fenotipi da "superbatteri".

Sebbene frequente, il modello di evoluzione per macro-ricombinazioni non è comune a tutti i batteri responsabili di infezioni nosocomiali. Un'eccezione è il ceppo SMAL di *Acinetobacter baumannii*, presentato in un altro sottoprogetto di questa tesi. In questo lavoro sono stati analizzati i genomi del sequence type (ST) 78 di *A. baumannii*. La filogenesi e la genomica comparativa hanno rivelato la presenza di due differenti cladi all'interno del ST che presentano differenti "stili" evolutivi. Un gruppo (contenente i genomi SMAL) è caratterizzato da una minore variabilità del contenuto genico e dalla presenza di un numero più elevato di copie di insertion sequence (IS). Una IS interrompe il gene *comEC/rec2* in tutti i genomi SMAL. Questo gene codifica per una proteina coinvolta nell'acquisizione del DNA esogeno, quindi la sua inattivazione limita lo scambio di geni. Questo suggerisce una spiegazione per la bassa plasticità genomica.

In un altro lavoro presentato in questa tesi, l'epidemiologia genomica è stata applicata per ricostruire la diffusione di un focolaio epidemico di *K. pneumoniae* in un'unità di terapia intensiva ospedaliera. In un primo momento, è stato utilizzato un approccio filogenetico per separare gli isolati appartenenti all'epidemia da quelli sporadici. Poi le date di isolamento e gli SNP genomici hanno permesso di costruire una rete genomica che modellasse la propagazione delle infezioni nel reparto. La ricostruzione ha indicato una diffusione radiale del patogeno dal paziente zero a tutti gli altri infetti, rivelando così un errore sistematico nelle procedure di biosicurezza dell'ospedale.

Questa applicazione quasi forense dell'epidemiologia genomica è stata utilizzata anche in altri due lavori qui presentati, entrambi riguardanti la ricostruzione di infezioni alimentari. In uno degli articoli, incentrato su *Salmonella enterica*, l'analisi filogenetica è stata eseguita solamente con gli SNP sinonimi al fine di filtrare le mutazioni patoadattative. Nell'altro lavoro sono stati utilizzati dati epidemiologici, tipizzazione molecolare e filogenesi basata sugli SNP per studiare l'infezione di nove isolati di *Listeria monocytogenes*, che si ritenevano essere parte dello stesso focolaio e alla fine sono risultati genomicamente non correlati.

Infine, viene qui presentato anche un articolo di review riguardante l'epidemiologia genomica. L'articolo è focalizzato sulle ultime pubblicazioni ad alto impatto che analizzano l'evoluzione genomica degli agenti patogeni batterici e le dinamiche di propagazione delle epidemie in brevi periodi di tempo. L'articolo descrive, infine, le ultime ricostruzioni epidemiologiche a livello storico, che sono possibili grazie alle moderne tecnologie di isolamento e sequenza del DNA.

INDEX

Introduction **p. 7**

Article 1: Genomic epidemiology of *Klebsiella pneumoniae*: the Italian scenario, and novel insights into the origin and global evolution of resistance to carbapenem antibiotics **p. 18**

Article 2: Insertion sequences proliferation and limited genomic plasticity in *Acinetobacter baumannii* sequence type 78, a persistent clone in Italian hospitals **p. 41**

Article 3: Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole-genome analysis: small-scale Italian scenario within a single hospital **p. 64**

Article 4: Differential single nucleotide polymorphism-based analysis of an outbreak caused by *Salmonella enterica* serovar Manhattan reveals epidemiological details missed by standard pulsed-field gel electrophoresis **p. 87**

Article 5: Genomic characterization helps dissecting an outbreak of Listeriosis in northern Italy **p. 118**

Article 6: Bacterial genomic epidemiology, from local outbreak characterization to species-history reconstruction **p. 140**

Conclusions **p. 163**

Supplemental material from articles **p. 167**

INTRODUCTION

Nosocomial infections. Prokaryotes have appeared on our planet billions of years before humans and they have always influenced and shaped our life. Indeed, prokaryotes interact with virtually all other life forms, playing multiple important roles, from beneficial symbiosis to harmful pathogenesis. They inhabit a big share of the human body, living on the skin, in the gut, and in the nasal cavity. Moreover, prokaryotes cover every surface we touch and populate every environment we live in. Thus, the interaction with them cannot be avoided, but evolution has shaped our immune system in order to restrict the relationship to some species and only in certain body compartments. For example, our food is almost completely sterilized by the acid environment of the stomach but colonized right afterwards by the gut microbiota. Microbial spillover from the gut to other body compartments is a common phenomenon but is always controlled and restricted by the immune system.

Problems occur when the immune system is weakened, i.e. with old age, during recovery from surgery or in the presence of pathological states such as AIDS or cancer. By this conditions, normally harmless commensals are free to invade and heavily colonize body districts in which they are normally not allowed such as the blood stream, the urinary tract or even the brain. Such infections occur mostly in the hospital environment (in this case they are labelled nosocomial infections).

The Centers for Disease Control and Prevention (CDC) has classified nosocomial infection into 13 types, with 50 infection sites, which are specific on the basis of biological and clinical criteria. The most common types are urinary tract infections (UTI), surgical and soft tissue infections, gastroenteritis, meningitis and respiratory infections (1). In 2011, the number of recorded hospital infections in the USA alone was over 722,000 (2). According to the CDC, the total cost for the treatment of healthcare associated infections is around 30 billion US Dollars (3). In Italy, the number of infections in one year is between 450,000 and 700,000. One percent of these are estimated to be the direct cause of the death of the patient (4).

Bacteria are responsible for about ninety percent of infections, while protozoans, fungi, viruses and mycobacteria take the remaining 10% share. The species that are usually involved in hospital-acquired infections include *Streptococcus* spp., *Acinetobacter* spp., enterococci, *Staphylococcus aureus*, *Bacillus cereus*, *Pseudomonas aeruginosa*, coagulase-negative staphylococci, *Legionella* and Enterobacteriaceae family members including *Proteus mirabilis*, *Klebsiella pneumoniae*, *Escherichia coli*, *Serratia marcescens*. Out of these, *P. aeruginosa*, *S. aureus*, *E. coli* and *Enterococcus* spp. have a major role. *E. coli* is common in the UTI, while *S. aureus* is frequent in other body sites (*S. aureus* is very frequent in blood-borne infections). *Enterococcus* spp. mostly infect surgical-sites while *P. aeruginosa* infections are evenly distributed among all body districts (1).

Resistance to antibiotics. The key factor that turns nosocomial infections in a serious threat for health is resistance to antibiotics. Indeed, infections with resistant bacteria are difficult to treat because identifying what antibiotic is effective requires time-consuming microbiology assays. A delay in the identification of the infective agent and in its characterization is often the cause of death for many patients. Moreover, in a growing number of cases, physicians have to deal with multiresistant microbes or, in some cases, pan-resistant microbes, which are resistant to all known antimicrobial agents.

There are several mechanisms of antibiotic resistance that bacteria can acquire or develop: enzymatic degradation of antibiotics, antibiotic target modification, changes in the bacterial cell wall permeability and the use of pathways alternative to that targeted by the antimicrobial agent.

Enzymatic degradation or inactivation of antibiotics is a very common mechanism of resistance. The most known examples are the β-lactamases; i.e. enzymes hydrolyzing the β-lactam ring of antibiotics such as the cephalosporins. These are mainly of concern in Gram-negative bacteria (5). Additional examples include inactivation of aminoglycosides by enzymatic modification by acetyltransferases, nucleotidyltransferases and phosphotransferases (6). Each of these enzymes has many variants, each active against specific antibiotic molecules (7).

Resistance by target modification consists in modifying the binding site of the antibiotic on its target, thus making the drug ineffective. Examples of this mechanism are mutations in the gyrase and topoisomerase genes, which are the targets of the quinolone and fluoroquinolone antibiotics (8). Instead, in the case of Methicillin resistant Staphylococcus aureus (MRSA), the mecA gene codes for a variant of the penicillin binding protein PBP2A that has a very low affinity for β-lactams (9). One curious case is resistance to colistin (polimixin E) in *Klebsiella pneumoniae*. This phenotype is achieved by changing the composition and thus the charge of the lipopolysaccharide (LPS), which is the target of the antibiotic (10).

Changing the cell wall or cell envelope permeability implies reducing the entry rate or increasing the efflux of antibiotics. Mutations in pores can limit or completely inhibit the influx of antibiotics into the cell. On the other side, efflux can be increased by synthesizing specific efflux pumps, as in the case of resistance to tetracycline (11), or by over-producing physiologically expressed ones.

Finally, cells can become resistant by deviating from their normal physiological pathway by including a step alternative to that targeted by the antibiotic. An example of this mechanism is the production of an alternative dihydrofolate reductase in trimethoprim resistant *Escherichia coli* and *Citrobacter sp* (12).

Virulence factors in specialized strains. Very often nosocomial pathogenic species comprise strains that evolved specifically to fit into the hospital environment and to invade specific body district of patients. This is achieved through the acquisition of virulence factors, including mechanisms used to attach to surfaces, to migrate between body districts, to defend from the host immune system, to outcompete other microorganism in nutrients uptake and to attack and damage the host.

Bacteria produce a wide variety of surface proteins that allow them to adhere to the host tissues, such as fimbriae, lipoteichoic acid and trimeric autotransporter adhesins. Capsules, instead, are used to evade the immune system, by inhibiting phagocytosis, and to protect the bacteria while outside the host. Another group of virulence factors are destructive enzymes, which cause damage to host tissues in order to invade other body districts or gain nutrients. Enzymes include hyaluronidase, which breaks down the connective tissue component hyaluronic acid, but also a range of proteases, lipases and DNases (13). Other mechanisms to provide nutrients include siderophores and other system used for the uptake of metals and ions. One example of this category of virulence factor is Yersiniabactin, a siderophore used by *Yersinia pestis* (and frequently horizontally transmitted to other species)(14) to monopolize iron in colonized environments (15).

Lastly, a major category of virulence factors are toxins. The lipid A component of LPS (also known as endotoxin) binds to monocytes receptors and stimulates the inflammatory response in the host. An excess of such response can lead to septic shock. Exotoxins, on the other hand, are actively secreted molecules and cause damage to the host by targeting different specific biological processes. The two most potent known exotoxins are tetanospasmin (secreted by *Clostridium tetani*) and the botulinum toxin (*Clostridium botulinum*). Other bacteria that produce exotoxins include: *Escherichia coli*, *Vibrio cholerae*, *Bacillus anthracis*, and *Clostridium difficile*.

Acquisition of antimicrobial resistance factors and virulence factors is also common in bacteria that cause food-borne infections (such as *Salmonella* spp. and *Listeria monocytogenes*), sex transmitted diseases (*Neisseria gonorrhoeae, Treponema pallidum, Chlamydia trachomatis, Haemophilus ducreyi, Mycoplasma genitalium*) (7, 16).

Horizontal gene transfer. Antimicrobial resistance and virulence factors can surely evolve through random mutations and selection, however they can also be exchanged among bacteria using several mechanisms of horizontal gene transfer (HGT). HGT allows such factors to spread quickly in a population and among populations, species and genera, and thus represents a set of mechanisms of great importance for increasing a pathogen virulence of resistance.

Transformation is the direct uptake of DNA and depends on the expression of numerous genes (17). Transformation happens most frequently with genomic material of the same species as the recipient bacterium; Transformed DNA is usually integrated by homologous recombination. Competence for transformation is typically induced by stress conditions during the stationary phase of growth.

Bacterial conjugation, instead, involves the presence and contact of the two living individuals between whom the DNA exchange happens. This process is mediated by pili and consists in the transfer of a conjugative or mobilizable genetic element that is most often a plasmid or transposon. Most conjugative plasmids have systems ensuring that the recipient cell does not already contain a similar element (18). Conjugation was discovered observing the transmission of fertility factor F, an autonomous DNA molecule. F can integrate into the bacterial chromosome to produce Hfr derivatives. Both in the autonomous and in the integrated state, F allows the bacterium to pair with F− recipient bacteria. This allows a copy of the replicating F to be transferred to the partner cell. From in the Hfr state, F can also transfer neighboring parts of the donor chromosome. This process allows the transmission of new functions to the recipient strain (19).

Lastly, transduction is a mechanism of HGT based on bacterial viruses that carry part of the previous host genome when infecting other individuals. Generalized transduction is operated by viral particles that carry a segment of host DNA instead of a replica of the viral genome. On the other hand, specialized transduction is based on the transportation of hybrid molecule with a part of the phage genes and some bacterial genes (20).

The role of genomic epidemiology. Nosocomial bacterial infections are mostly transmitted by contact. The vectors are often the hands of the medical staff, when they are not carefully cleaned in-between the treatment of one patient and another. Other carriers of infections are the so-called fomites: i.e. surfaces like diagnostic tools, catheter tubes, ventilators or simply door handles that are not sterilized. The control of the spread of pathogens might be difficult, due to the many potential sources of contagion and because bacteria can use people with healthy immune system as carriers, by colonizing them (e.g. in their intestines). These category of people, who do not have any symptom, can include the hospital personnel and patients with mild diseases, that are hospitalized in wards where regular checks for pathogens are not run. For this reason, it is fundamental for hospitals to have surveillance programs and strict protocols of behaviour for all workers.

Even the evolution of multidrug resistant pathogens can be contained. This is obtained by regulating and limiting the use of antibiotics. Indeed, resistance is acquired when a selective pressure is applied on the microbial community, i.e. when antibiotics are used. The best environment for the development of resistance is one with a sublethal antimicrobial dose. Thus, low dosages or short treatments must be avoided.

The two provisions above are fundamental in order to solve this problem, but on the other hand, much can be done to improve diagnostic and surveillance systems and to design new strategies and protocols to prevent the development and spread of resistance mechanisms. This can be achieved through the use of genomic approaches and genomic epidemiology.

Indeed, sequencing the DNA of an isolate allows to know what genes are present in its genome and to predict the associated phenotype. Moreover genomic structures can be studied using sequencing: e.g. the presence of plasmids, recombined regions in the chromosome, insertions and deletions of entire portions of the genome, such as pathogenicity islands, i.e. regions where genes coding for virulence factors or determinants for resistance to antimicrobial agents are clustered and usually co-transmitted by HGT. Lastly, shuffling of genomic regions can be recognized using genomics, together with the promoters of such reorganization, e.g. transposons and insertion sequences.

Comparative genomics consists in finding common tracts and differences between two or more genomes or groups of genomes. This allows to find the determinants for a specific phenotype, such as a novel resistance factor. Comparative genomics can be performed by grouping strains based on their phenotypic traits or in light of their evolution (although the two classifications often coincide).

Evolutionary classification is performed using molecular phylogenetics. Ordering bacteria (and their features) on a phylogenetic tree allows to reconstruct the history of a pathogen. Furthermore, using molecular clock approaches it is possible to fit a phylogeny inside a timescale and date evolutionary events. Lastly, phylogeography allows to compare phylogeny with geographical maps: this can be used to locate the area of origin of a strain of interest, as well as to hypothesize the origin of an epidemic event. Performing phylogeny with more isolates, can often lead to a better comprehension of a pathogen. Using a metaphor, genomic sequencing can be compared to taking a picture of a pathogen in a specific moment of its history. The more genomes are sequenced, the more frames can be used to reconstruct a movie about the evolution of the pathogen.

In the present years, sequencing genomes is not a limiting step when performing genomic epidemiology. This is thanks to the high amount of available genomes in the databases, which were sequenced at low cost. This situation is the result of the ongoing revolution represented by the advent of next generation sequencing.

Next generation sequencing. Next generation sequencing (NGS) techniques were introduced in 2005 (21, 22) and completely changed the way many fields of biomedical research were conducted. The new technologies permitted to sequence an extremely high (when compared to the previous technology) amount of nucleotides at the same time, thus making the price of such precious information drop dramatically and starting a revolution, possibly more game-changing than the advent of PCR.

The first technology introduced in 2005 was Roche 454 (21, 22), which was based on the use on a system of beads and wells that permit to separate and simultaneously sequence hundreds of thousands of small fragments of DNA. 454 machines use the energy of the pyrophosphate released by the attached nucleotides to produce a light signal (for this reason, this technology is also known as pyrosequencing). The development of this technology was discontinued in 2013 but its place in the market was taken by the Ion Torrent machines. These sequencers use a similar approach as 454, but replace the base calling system based on pyrophosphate with one based on the positive hydrogen ions released in the DNA synthesis process. The increase of H^+ ions, changes the pH of the solution in which the reaction happens. This difference can be translated into an electric signal (23). This approach makes ion-torrent sequencing cheaper than 454. The policy of Ion Torrent is to produce easy to use sequencing kits and develop analysis software that require very few informatic skills, thus optimizing the technology to perform standard assays.

Among the market leaders in the NGS field, together with Ion Torrent, is Illumina. Machines produced by this company use a technology introduced in 2006 by Solexa, which does not use beads and wells to handle DNA fragments but a flat surface, called the flowcell. DNA fragments are attached to the flowcell thanks to short single strand DNA probes, which bind to the sequencing primers. Each piece of DNA is amplified locally and forms an island of identical sequences on the flowcell. When sequencing is performed, each island emits a light signal which is detectable by a CDC receptor. The use of the flowcells allows Illumina to scale up the throughput of its machines easily. Today the machine with the highest output capabilities, Illumina HiSeq 4000, can sequence up to 1.5 terabases per run. The other main advantage of Illumina technologies is the accuracy in sequencing homopolymers, i.e. repeated stretches of the same base. A system of reversible terminators allows to pause the reaction after the incorporation of each nucleotide, regardless of the nitrogenous base attached. This allows to detect the right length of homopolymers. The very high throughput and versatility of Illumina machines is what makes it the preferred choice for bacterial genomics. In fact, all sequences obtained in the works described in the present thesis were obtained using this technology.

In more recent years, so-called third generation sequencing have been launched on the market. These technologies allow to sequence very long fragments of DNA (~10,000bp) in one read. In 2011 Pacific Bioscience released a sequencing machine based on the Single Molecule Real Time (SMRT) technology, a parallelized single molecule DNA sequencing method. This technology relies on the use of nanoscale wells called zero-mode waveguides (ZMW). At the bottom of each ZMW, a single DNA polymerase enzyme sequences a single molecule of DNA. The structure of ZMW is small enough to observe the light of only one fluorescent dye, released with the incorporation of a single nucleotide (24). SMRT sequencing, also called PacBio, is used today to produce high quality genomic assemblies. Precision and reliability of called bases is achieved through high coverage sequencing or by coupling PacBio and Illumina reads.

A more recent technology called Nanopore, was introduced in 2012. Nanopore sequencing uses electrophoresis to transport DNA through a porin protein of diameter 10^{-9} m. Sequencing is made possible because samples cause specific changes in electric current density across nanopore surfaces, which allow the nucleotide to be recognized (25). At the beginning of 2017, Nick Loman reported having obtained reads of several hundred thousand bases in length (with a maximum of 882 kb) (26). Protocols for the use of this technology are still being developed; thus it is still not as widely used as SMRT.

In the present thesis, all genomes are sequenced with the Illumina technology. In particular, the MiSeq machine was used for most of the projects. The MiSeq is a low output benchtop machine, that allows to obtain the reads of 12-15 bacterial genomes per run. Libraries were prepared using the Nextera XT kit. This technology uses a modified transposases to fragment the DNA and attach the sequencing primers at the same time, in a process called tagmentation. Nextera allows to avoid the use of ultrasonicators and sensibly lower the price of library building. The cost of the reagents for one genome using the approach presented in this thesis is around 100€.

Preface to manuscripts. In the following sections I have included a collection of papers, relative to the research work performed during my period as a doctoral student. Works will be presented in the following order:

a) The first three articles included are works in which I was listed as first author and contain the main results of this Ph.D. project. They are focused on genomic epidemiology of nosocomial pathogens
b) The fourth and fifth articles are focused on the reconstruction of the spreading routes of food-borne pathogens during epidemic events. My contribution in these works was minor
c) The sixth paper is a review article on genomic epidemiology.

REFERENCES

1. Khan HA, Ahmad A, Mehboob R. 2015. Nosocomial infections and their control strategies. Asian Pac J Trop Biomed 5:509–514.

2. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R, Maloney M, McAllister-Hollod L, Nadle J, Ray SM, Thompson DL, Wilson LE, Fridkin SK. 2014. Multistate Point-Prevalence Survey of Health Care–Associated Infections. N Engl J Med 370:1198–1208.

3. Scott RDII. The Direct Medical costs of Healthcare-Associated Infections in U.S. Hospitals and the Benefits of Prevention. Centers for Disease Control and Prevention.

4. Infezioni correlate all’assistenza aspetti epidemiologici. Istituto Superiore di Sanità. http://www.epicentro.iss.it/problemi/infezioni_correlate/epid.asp

5. Livermore DM, Woodford N. 2006. The β-lactamase threat in Enterobacteriaceae, Pseudomonas and Acinetobacter. Trends Microbiol 14:413–420.

6. Wright GD. 1999. Aminoglycoside-modifying enzymes. Curr Opin Microbiol 2:499–503.

7. Verraes C, Van Boxstael S, Van Meervenne E, Van Coillie E, Butaye P, Catry B, de Schaetzen M-A, Van Huffel X, Imberechts H, Dierick K, Daube G, Saegerman C, De Block J, Dewulf J, Herman L. 2013. Antimicrobial resistance in the food chain: a review. Int J Environ Res Public Health 10:2643–2669.

8. Drlica K, Zhao X. 1997. DNA gyrase, topoisomerase IV, and the 4-quinolones. Microbiol Mol Biol Rev 61:377–392.

9. Pinho MG, Filipe SR, de Lencastre H, Tomasz A. 2001. Complementation of the essential peptidoglycan transpeptidase function of penicillin-binding protein 2 (PBP2) by the drug resistance protein PBP2A in Staphylococcus aureus. J Bacteriol 183:6525–6531.

10. Velkov T, Deris ZZ, Huang JX, Azad MAK, Butler M, Sivanesan S, Kaminskas LM, Dong Y-D, Boyd B, Baker MA, Cooper MA, Nation RL, Li J. 2014. Surface changes and polymyxin interactions with a resistant strain of Klebsiella pneumoniae. Innate Immun 20:350–363.

11. McMurry L, Petrucci RE Jr, Levy SB. 1980. Active efflux of tetracycline encoded by four genetically different tetracycline resistance determinants in Escherichia coli. Proc Natl Acad Sci U S A 77:3974–3977.

12. Pattishall KH, Acar J, Burchall JJ, Goldstein FW, Harvey RJ. 1977. Two distinct types of trimethoprim-resistant dihydrofolate reductase specified by R-plasmids of different compatibility groups. J Biol Chem 252:2319–2323.

13. Levinson W. 2016. Review of Medical Microbiology and Immunology 14E. McGraw Hill Professional.

14. Lawlor MS, O'connor C, Miller VL. 2007. Yersiniabactin is a virulence factor for Klebsiella pneumoniae during pulmonary infection. Infect Immun 75:1463–1472.

15. Perry RD, Balbo PB, Jones HA, Fetherston JD, DeMoll E. 1999. Yersiniabactin from Yersinia pestis: biochemical characterization of the siderophore and its role in iron transport and regulation. Microbiology 145 (Pt 5):1181–1190.

16. Krupp K, Madhivanan P. 2015. Antibiotic resistance in prevalent bacterial and protozoan sexually transmitted infections. Indian J Sex Transm Dis 36:3–8.

17. Chen I, Dubnau D. 2004. DNA uptake during bacterial transformation. Nat Rev Microbiol 2:241–249.

18. Ryan K, George Ray C, Ahmad N, Lawrence Drew W, Plorde J. 2014. Sherris Medical Microbiology, Sixth Edition. McGraw Hill Professional.

19. Hayes W. 1984. The Genetics of Bacteria and Their Viruses: Studies in Basic Genetics and Molecular Biology. Wiley, New York, NY, USA.

20. Arber W. 2014. Horizontal Gene Transfer among Bacteria and its Role in Biological Evolution. Life 4:217–224.

21. Rothberg JM, Leamon JH. 2008. The development and impact of 454 sequencing. Nat Biotechnol 26:1117–1124.

22. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF,

Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

23. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352.

24. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323:133–138.

25. Feng Y, Zhang Y, Ying C, Wang D, Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics 13:4–16.

26. Loman N. 2017. Thar she blows! Ultra long read method for nanopore sequencing. Loman Labs.

ARTICLE 1

Genomic epidemiology of *Klebsiella pneumoniae*: the Italian scenario, and novel insights into the origin and global evolution of resistance to carbapenem antibiotics

AMERICAN SOCIETY FOR MICROBIOLOGY | Antimicrobial Agents and Chemotherapy

2015. 59:389–396.
doi:10.1128/AAC.04224-14

Genomic Epidemiology of *Klebsiella pneumoniae* in Italy and Novel Insights into the Origin and Global Evolution of Its Resistance to Carbapenem Antibiotics

Stefano Gaiarsa[a,b], Francesco Comandatore[b,c], Paolo Gaibani[d], Marta Corbella[a,e], Claudia Dalla Valle[a], Sara Epis[b], Erika Scaltriti[f], Edoardo Carretto[g],Claudio Farina[h], Maria Labonia[i], Maria Paola Landini[d], Stefano Pongolini[f], Vittorio Sambri[j], Claudio Bandi[b], Piero Marone[a] and Davide Sassera[c]

Author Affiliations

[a] Microbiology and Virology Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy.
[b] Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milan, Italy
[c] Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy
[d] Unit of Clinical Microbiology, St. Orsola-Malpighi University Hospital, Bologna, Italy
[e] Biometric and Medical Statistics Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy
[f] Sezione Diagnostica di Parma, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Parma, Italy
[g] Clinical Microbiology Laboratory, IRCCS Arcispedale S. Maria Nuova, Reggio Emilia, Italy
[h] Microbiology Institute, AO Papa Giovanni XXIII, Bergamo, Italy
[i] Dipartimento di Diagnostica di Laboratorio e Trasfusionale, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy
[j] Unit of Clinical Microbiology, The Greater Romagna Area-Hub Laboratory, Pievesestina, Italy
S.G. and F.C. contributed equally

Received 4 September 2014. Returned for modification 26 September 2014. Accepted 26 October 2014. Accepted manuscript posted online 3 November 2014.
Address correspondence to Davide Sassera, **davide.sassera@unipv.it**.

Supplemental material is available at the end of the thesis

This article can be found at **http://dx.doi.org/10.1128/AAC.04224-14**
Please scan this QR code to access the website from the printed version

ACKNOWLEDGMENTS This work was supported by Ricerca Corrente 2013 funding from Fondazione IRCCS Policlinico S. Matteo to P.M. We thank Simone Ambretti for providing samples and Rosa Visiello for her assistance in correcting the manuscript.

ABSTRACT

Klebsiella pneumoniae is at the forefront of antimicrobial resistance for Gram-negative pathogenic bacteria, as strains resistant to third-generation cephalosporins and carbapenems are widely reported. The worldwide diffusion of these strains is of great concern due to the high morbidity and mortality often associated with *K. pneumoniae* infections in nosocomial environments. We sequenced the genomes of 89 *K. pneumoniae* strains isolated in six Italian hospitals. Strains were selected based on antibiotypes, regardless of multilocus sequence type, to obtain a picture of the epidemiology of *K. pneumoniae* in Italy. Thirty-one strains were carbapenem-resistant *K. pneumoniae* carbapenemase producers, 29 were resistant to third-generation cephalosporins, and 29 were susceptible to the aforementioned antibiotics. The genomes were compared to all of the sequences available in the databases, obtaining a data set of 319 genomes spanning the known diversity of *K. pneumoniae* worldwide. Bioinformatic analyses of this global data set allowed us to construct a whole-species phylogeny, to detect patterns of antibiotic resistance distribution, and to date the differentiation between specific clades of interest. Finally, we detected an ~1.3-Mb recombination that characterizes all of the isolates of clonal complex 258, the most widespread carbapenem-resistant group of *K. pneumoniae*. The evolution of this complex was modeled, dating the newly detected and the previously reported recombination events. The present study contributes to the understanding of *K. pneumoniae* evolution, providing novel insights into its global genomic characteristics and drawing a dated epidemiological scenario for this pathogen in Italy.

INTRODUCTION

Multidrug resistance is currently a matter of concern worldwide. At the end of the 1970s, most *Escherichia coli* and *Klebsiella pneumoniae* strains encoded ampicillin-hydrolyzing β-lactamases, making it necessary to use third-generation cephalosporins. In the early 1980s, the first cases of resistance to these novel antibiotics were reported in *Enterobacteriaceae* (1) and were caused by genes classified as ESBL (extended-spectrum beta-lactamases). In 1985, the United States Food and Drug Administration approved the commercialization of imipenem, a molecule that showed activity against ESBL producers. This drug, and similar compounds that quickly followed (i.e., carbapenems), then were introduced into clinical practice and widely used.

In 2001, Yigit and colleagues reported a *K. pneumoniae* strain isolated in 1996 that exhibited resistance to the carbapenems imipenem and meropenem (2). The gene responsible for the resistance was identified as a group 2f, class A, carbapenem-hydrolyzing beta-lactamase, named *Klebsiella pneumoniae* carbapenemase 1 (KPC-1). Since its discovery, carbapenem resistance caused by the *bla*$_{\text{KPC}}$ gene has been reported increasingly in *K. pneumoniae* isolates, initially moving through the northeastern states (3, 4) and quickly becoming the most frequently found carbapenemase in the United States (5). The spread of KPC then continued, with reports from different countries appearing ceaselessly, to the point that today this is regarded as a worldwide issue (6).

The *bla*$_{\text{KPC}}$ gene is carried by a plasmid; thus, horizontal transfer between various *K. pneumoniae* strains, as well as other bacterial species, could be expected and was extensively reported (7–9). Nevertheless, most of the clinical reports to date have been caused by *K. pneumoniae* isolates belonging to clonal complex 258 (CC258) (10). This complex comprises sequence type 258 (ST258) and single-allele mutant STs based on multilocus sequence typing (MLST), such as ST11 and ST512. These epidemiological data suggest a dissemination starting from a single ancestor and that CC258 presents a genomic background that is favorable both to the acquisition of plasmids bearing the *bla*$_{\text{KPC}}$ gene and to the clonal spread in nosocomial environments. In 2014, Deleo and colleagues (11) presented a phylogenomic study on 85 *K. pneumoniae* isolates belonging to CC258, detecting two subclades and concluding that an ~215-kb recombination event was at the origin of the differentiation between the two. A second comparative genomic analysis, presented by Chen and colleagues (12), detected an ~1.1-Mb recombination between an ST11 recipient and an ST442 donor as the event that originated the present ST258 strain.

Since the first finding of circulation of ESBL-producing *K. pneumoniae* in Italy in 1994, a rapid and extensive dissemination of different types of ESBLs has been reported (13–15). More recently, the first Italian KPC-positive *K. pneumoniae* strain, belonging to ST258, was isolated in a hospital in Florence in 2008 from an inpatient with a complicated intra-abdominal infection (16). Since then, the diffusion of carbapenemase-producing *K. pneumoniae* in Italy has been extremely rapid and characterized mainly by isolates of CC258 (i.e., ST258 and ST512) (17–19). ST512 in particular, first reported in Israel in 2006 (20), has been spreading in southern Europe and South America (11, 19). The sporadic detection of isolates belonging to other STs (e.g., ST101 and ST147) also have characterized the epidemiology of KPC *K. pneumoniae* in Italy (19).

The aim of this study was to evaluate the geographic and phylogenetic distribution of *K. pneumoniae* isolates of different antibiotypes, both at a national and a global scale. Thus, we sequenced and analyzed the genomes from 89 *K. pneumoniae* strains, collected in six Italian hospitals from 2006 to 2013, without any *a priori* knowledge of the sequence type. We compared this national collection to all of the *K. pneumoniae* genomes available from worldwide isolations to obtain insights into both the Italian epidemiology and the global structure of the species.

MATERIALS AND METHODS

Strain sampling. Eighty-nine non-duplicate *K. pneumoniae* strains, collected from six different Italian hospitals, were included in this study without prior knowledge of the sequence type. Thirty-one were KPC producers, as demonstrated using phenotypical tests (positivity with disk diffusion synergy testing using a meropenem disk alone and in combination with aminophenylboronic acid) (21) and/or genotypical analysis (in-house methods based on reference 22); 29 were ESBL producers, as demonstrated using the procedure recommended by the CLSI (23), while 29 were susceptible to third-generation cephalosporins and carbapenems. Throughout this work, we refer to this last group of isolates as susceptible. Antimicrobial susceptibility testing was performed using a Vitek2 automated system (bioMérieux), and MICs were interpreted by following the European Committee on Antimicrobial Susceptibility Testing guidelines (24). The list of isolates, year, location of isolation, sequence type, and presence of selected antibiotic resistance genes are reported in Table S1 in the supplemental material.

Genome sequences. DNA was extracted using a QIAamp DNA minikit (Qiagen) by following the manufacturer's instructions. Whole genomic DNA was sequenced using an

Illumina Miseq platform with a 2 by 250 paired-end run after Nextera XT paired-end library preparation. On 24 March 2014, sequences of draft and complete genomes of *K. pneumoniae* were retrieved from the NCBI ftp site, while sequencing reads of the isolates sequenced by Deleo and coworkers (11) were retrieved from the sequence read archive (SRA) database (accession no. SRP036874).

Genome assembly and retrieval. Sequencing reads from the isolates obtained in this study were assembled using MIRA 4.0 software (25) with accurate *de novo* settings. Assembled genomes are now publicly available under Bioproject (EMBL project B6543). Reads retrieved from the SRA database were checked and filtered for sequencing quality using an in-house script and then assembled using Velvet (26) with a K-mer length of 35 and automatic detection of average expected coverage and low coverage threshold.

Resistance profile and MLST determination. The MLST profile was obtained *in silico* by searching the characterizing gene variants on each genome, using an in-house Python script. The antibiotic resistance profile was determined using a BLAST search on a gene database comprising all of the most common resistance genes associated with resistance to beta-lactams, including ESBL- and KPC-producing phenotypes.

Core SNP detection and phylogeny. Single-nucleotide polymorphisms (SNPs) were detected using an in-house pipeline based on Mauve software (27), using the NJST258_1 complete genome as a reference. Each genome was individually aligned to the reference, and alignments were merged with in-house scripts. Core SNPs were defined as single-nucleotide mutations flanked by identical bases present in all of the analyzed genomes. The core SNP alignment was used to perform a phylogenetic analysis using the software RAxML (28) with a generalized time-reversible (GTR) model and 100 bootstraps. The same phylogenetic approach was used to perform the analysis on three core SNP sub-data sets (i.e., non-recombined regions and two distinct putatively recombined regions).

Recombination. We divided the genome alignment in 5,264 windows of 1,000 nucleotides (nt) each and calculated core SNP frequency in each window for each genome, generating a matrix. The software R then was used to generate a heatmap of SNP frequency. The newly characterized strain 46AVR was used as a reference for plotting SNPs, being a member of the sister group to CC258. In parallel, we created a sub-data set of 174 CC258 genomes and 13 closely related *K. pneumoniae* genomes, removing genomes of isolates distant from the CC258 clade (*n* = 103) and the genomes within CC258 that exhibited extremely limited variability (*n* = 29), such as all but one of those obtained from single outbreaks. The choice of using a relatively large number of non-CC258 genomes (*n* = 13) was made in order to allow the detection of recombination events common to the whole clonal complex. We used this sub-data set of core SNPs in 187 genomes to perform a recombination detection analysis using the software BRATnextgen (29) with 100-iteration analysis, using 100 replicates for statistical significance.

Analysis of the recombined region. A database was created collecting protein sequences of factors previously reported to be involved in virulence and antibiotic resistance. We collected sequences from the Comprehensive Antibiotic Resistance Database (CARD) (30) and from the Antibiotic Resistance Genes Database (ARDB) (31), from proteins involved in the biosynthesis of lipopolysaccharides (LPS) and polymyxin resistance, and from the most common virulence factors and siderophores found in Gram-negative bacteria (obtained from the NCBI site). Finally, we added to our manually designed database all *K. pneumoniae* proteins described as potential virulence or resistance factors in the work by Lery and colleagues (32). Gene sequences present in the novel putative recombined region were extracted from the genome of strain NJST258_1 using an in-house Python script. Correspondence between proteins in our database and genes in the recombined region was tested using a TBLASTN search, selecting genes covering at least 75% of the database sequence with a minimum of 75% identity. Results then were manually checked (see Table S2 in the supplemental material for a complete list).

Molecular clock. We created a sub-data set of 174 CC258 genomes and 3 closely related *K. pneumoniae* genomes (used as outgroups), removing genomes of isolates distant from the CC258 clade (*n* = 113) and the genomes within CC258 that exhibited extremely limited variability (*n* = 29), such as all but one of those obtained from single outbreaks. We used the software BEAST (33) on the core SNP alignment of the 177-genome sub-data set after removing SNPs located in the potentially recombined regions. BEAST parameters used were the following: uncorrelated log-normal relaxed clock with the GTR model, with no correction for site rate heterogeneity according to analyses performed in similar scenarios (34). The analysis was run for 1,000,000,000 steps, and at every 10,000 steps samples were taken. We discarded 250,000,000 steps as burn-in. The program TRACER (http://beast.bio.ed.ac.uk/tracer/) was used to evaluate the convergence of the analysis.

RESULTS

Sampling and genome sequencing. Eighty-nine *K. pneumoniae* strains were collected in six Italian hospitals, chosen based on antibiotypes regardless of sequence type, which was determined only afterwards. The data set was composed of 31 KPC producers, 29 ESBL producers, and 29 strains susceptible to carbapenems and third-generation cephalosporins, here referred to as susceptible. The genome of each of the 89 isolates was sequenced and assembled (average genome size, 5,551,959 nt; average N50, 154,414 nt; average coverage, 76.46×). All of the available *K. pneumoniae* genome sequences and reads then were retrieved from the databases (n = 230) to create a global data set of 319 *K. pneumoniae* genomes. All genomes in the data set were screened for genes responsible for KPC and beta-lactam resistance phenotypes, as well as for all MLST genes. A total of 55 different MLST profiles were detected, eight of which were novel; thus, they were submitted to the curators of the *K. pneumoniae* MLST database (35). Each of the eight new profiles was represented by a single newly sequenced Italian isolate (7 susceptible, 1 ESBL producer). Two of these isolates also presented a single novel allele, one for the gene *rpoB* and one for the gene *infB*. See Table S1 in the supplemental material for a list of all of the isolates sequenced in this study and their main characteristics.

Global SNP phylogeny. We used a maximum likelihood phylogenomic approach based on core SNPs to elucidate the relationships within the global genome data set comprising the newly sequenced isolates and the *K. pneumoniae* genome sequences available in the database. The presence of antibiotic resistance genes was mapped on the resulting phylogenetic tree, obtained from an alignment of 94,812 core SNPs (Fig. 1). This revealed that 97% of all KPC *K. pneumoniae* strains sequenced to date, regardless of the location of isolation, belong to a well-supported clade, corresponding to the complex CC258. On the other hand, the phylogenomic analysis showed that the isolates encoding common beta-lactam resistance genes (*bla*$_{\mathrm{SHV}}$ family, *bla*$_{\mathrm{TEM}}$ family, *bla*$_{\mathrm{OXA}}$ family, and *bla*$_{\mathrm{CTX\text{-}M}}$family) are widespread along the tree and belong to various STs (both inside and outside CC258), with no sign of clustering. In fact, the 141 isolates encoding *bla*$_{\mathrm{TEM}}$ belong to 24 different STs, the 26 isolates encoding *bla*$_{\mathrm{OXA}}$ belong to 11 different STs, and the 37 isolates encoding *bla*$_{\mathrm{CTX\text{-}M}}$ belong to 16 different STs.

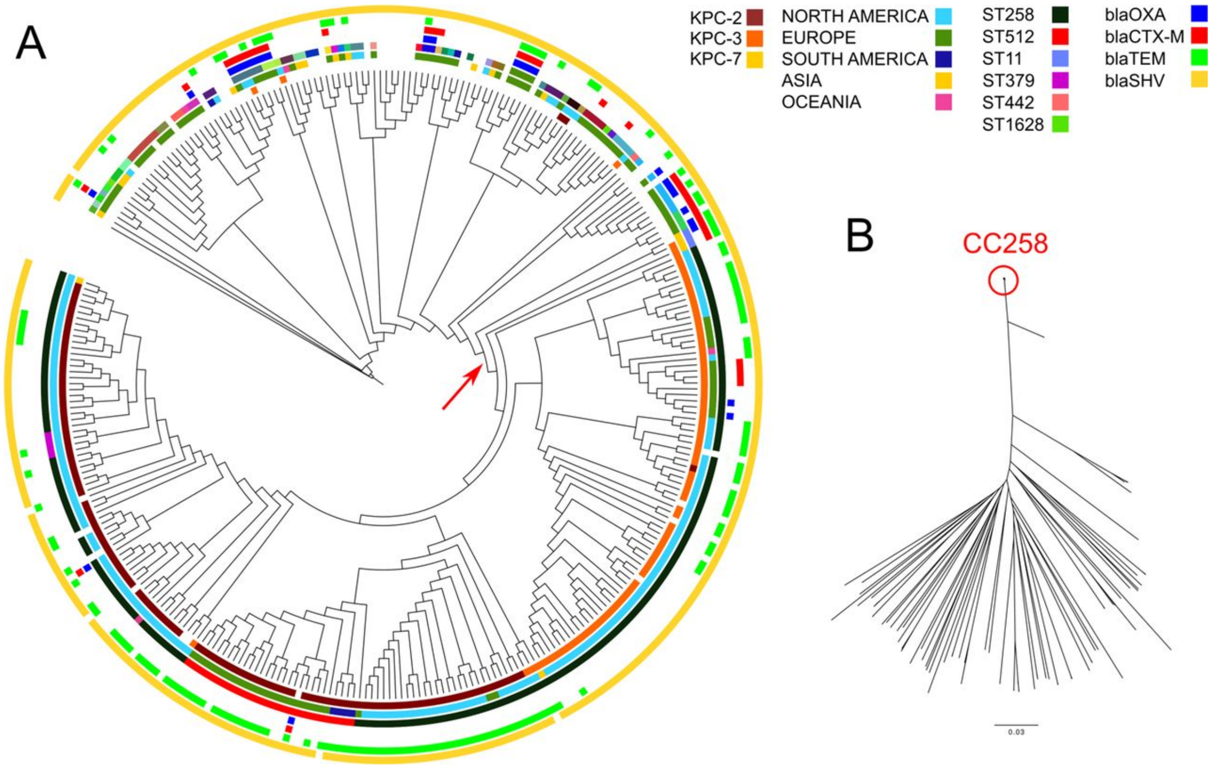


FIG 1. Maximum likelihood phylogeny of *Klebsiella pneumoniae*, based on 319 genomes. The phylogeny was reconstructed starting from an alignment of 94,812 core SNPs, using the software RAxML with a generalized time-reversible (GTR) model and 100 bootstraps, which are not shown for the sake of figure clarity. (A) Circular representation of the phylogeny, obtained using iTOL (itol.embl.de), ignoring branch length. Color circles indicate, from the innermost to the outermost, presence/absence of KPC variants, geographic location in terms of continents, ST based on multilocus sequence typing, and presence in the genome of genes from four beta-lactamase families. The red arrow indicates the origin of the clonal complex 258 clade. (B) Unrooted representation of the phylogeny showing the branch lengths, highlighting the genetic uniformity of clonal complex 258.

Phylogeny excluding potentially recombined regions. In a recent work by Castillo-Ramirez and coworkers (34), high-density SNPs clusters with a low ratio of nonsynonymous to synonymous evolutionary changes (*dN*/*dS*) in closely related bacterial genomes were suggested to be indicators of recombination events. Thus, we evaluated the distribution of SNPs on the genome data set, detecting a highly uneven distribution in the genomes of CC258 isolates, as most core SNPs clustered into two main regions. The first region is located between positions 1,675,550 and 2,740,033, while the second comprises the origin of replication and spans from 4,554,906 to 629,621 in strain NJST258_1 (Fig. 2) (for the distribution of core SNPs on the whole data set of 319 genomes, see Fig. S1 in the supplemental material). To further analyze the possible presence of recombination events in CC258, we used the software BRATnextgen (29), specifically intended for this purpose, on a reduced data set of 187 genomes of CC258 and closely related strains. This analysis (see Fig. S2) confirmed the presence of the two main recombination events, additionally indicating in what position of the phylogeny they could have occurred. The first event was placed between the entire CC258 clade and the non-KPC external isolates of different STs, while the second was between the outermost strains of ST11 and the inner CC258 clade. Details on these recombined regions are presented in the following paragraph.

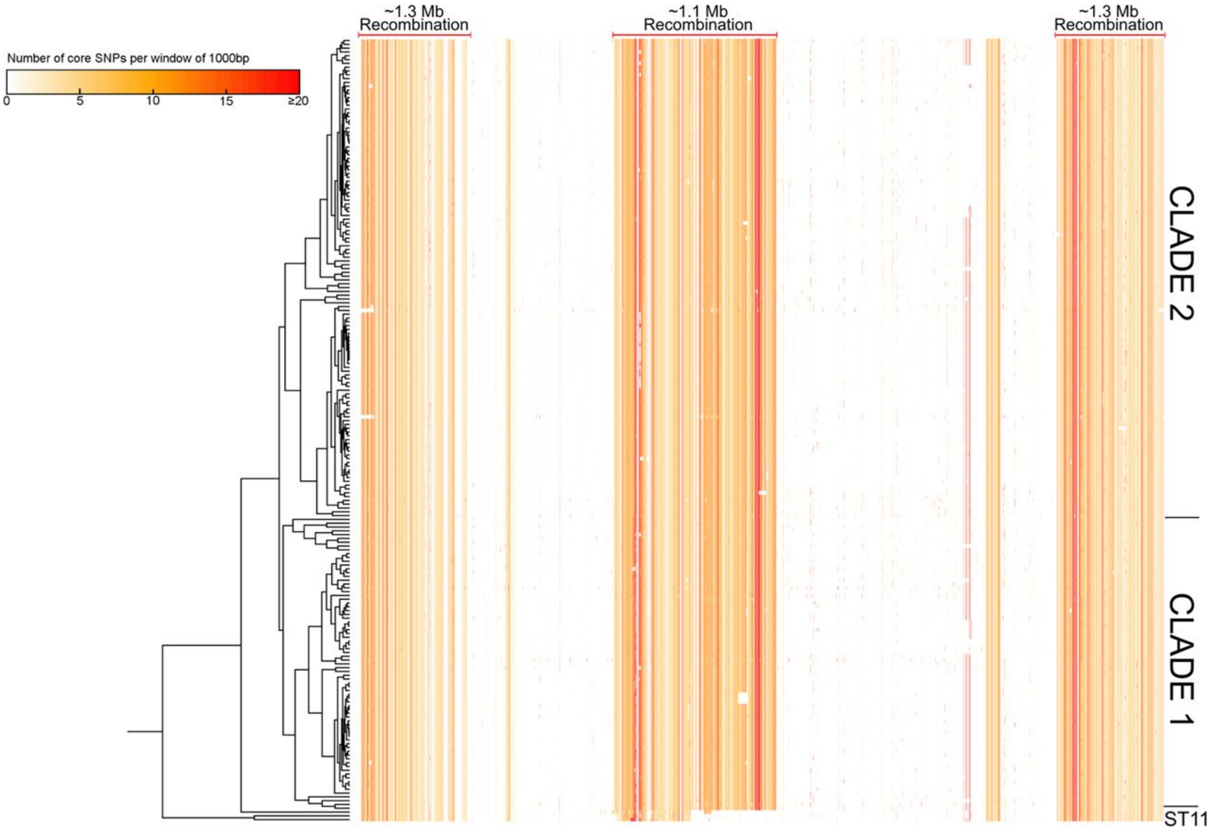


FIG 2. Uneven clustering of core SNPs in the clonal complex 258 clade. The phylogenetic reconstruction of the 206 representatives of the clonal complex 258 clade is shown on the left, while the core SNP frequency is shown on the right in shades of red, representing the number of core SNPs per 1,000-bp window for each genome. Detected recombinations are indicated at the top of the figure, and main clades of the clonal complex are indicated on the right side of the figure.

We removed the two putative recombined regions from the core SNPs data set of 319 *K. pneumoniae* genomes and performed a phylogenetic analysis on the remaining 55,368 core SNPs. The resulting tree (see Fig. S3 in the supplemental material) is largely consistent with the one generated from the initial data set, confirming the widespread distribution of susceptible and ESBL isolates and the presence of the highly supported KPC CC258 clade. Indeed, both the analysis on all core SNPs and the one performed by removing recombining sites agree in clustering 97% of all KPC *K. pneumoniae* isolates sequenced in a well-supported clade (Fig. 1; also see Fig. S3). This monophyletic clade comprises 203 strains from Asia, Europe, Oceania, and North and South America, with isolation dates ranging from 2002 to 2013; 193 of these (95%) present the bla_{KPC} gene. Most isolates of this clade belong to ST258 (*n* = 167), but 4 other sequence types are present (i.e., ST11, SST379, ST418, and ST512), all single-nucleotide variants of ST258; thus, they belong to CC258. The second most common sequence type in the CC258 clade is 512, represented by 28 isolates that form a single monophyletic subgroup, located within the ST258 diversity. Interestingly, 24 of these 28 have been isolated in Italy, mostly in this study (*n* = 19) but also in previous works (18, 36). Within the CC258 clade, two main highly supported distinct subclades are detectable, comprising the vast majority of the genomes. Three additional CC258 genomes are located in the tree as sister groups of the two main clades, and all are representatives of ST11, again a single-nucleotide variation of ST258. The existence of the two main CC258 subclades was reported previously, and a single recombination event was proposed to be the cause of the differentiation between the two (11), while a subsequent work suggested multiple recombination events (37).

Analysis of recombined regions. As described above, the SNP clustering analysis detected high SNP concentrations in two large genomic regions (Fig. 2). The smaller of the two is highly congruent with the ~1.1-Mb recombination found by Chen and colleagues (12), which represents the major evolutive change between the members of ST11 and those in the 2 main subclades of the CC258 clade. Chen and colleagues found this region to be most similar to the corresponding region of isolate Kp13 of ST442 and suggested a recombination event, with the donor strain being a close relative of Kp13. Thus, we investigated whether a recombination event is at the origin of the second, newly detected, highly mutated genomic region, located from positions 4,554,906 to 629,621. We performed a phylogenetic analysis, including all of the 319 *K. pneumoniae* genomes examined in this work, on the core SNPs located in this region and in parallel on the core SNPs located in the ~1.1-Mb region. The phylogenetic analysis of the novel ~1.3-Mb region (see Fig. S4 in the supplemental material) confirms the recombination hypothesis, as the topology of the resulting tree clearly shows that Italian isolate 67BO, of the newly described ST1628, is the sister taxon to the entire

CC258 clade, suggesting that the donor was related to this isolate. The phylogenetic tree obtained from the ~1.1-Mb recombined region (see Fig. S5) confirms the published results, clustering the donor Kp13 as a sister taxon of the CC258 clade, with the exclusion of the outermost ST11 isolates. Thus, we propose an updated scenario in which a first recombination event gave origin to the first CC258 strains (represented by ST11), a second recombination subsequently originated ST258, and a third, smaller recombination initiated the split between the two main ST258 subclades (Fig. 3).

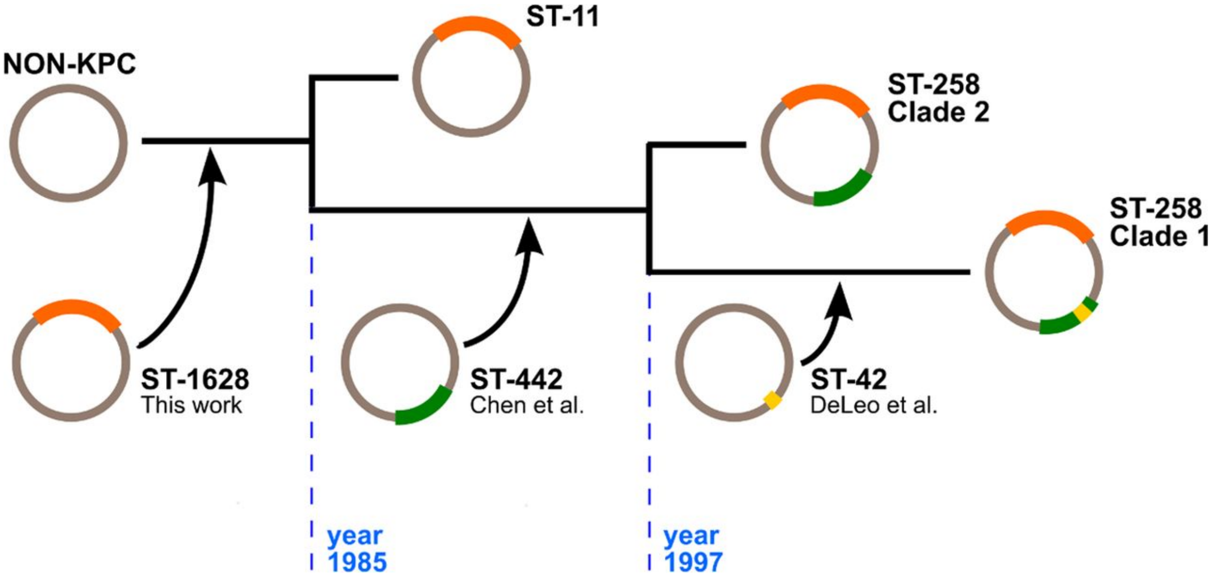


FIG 3. Hypothesis of recombinations occurring in the clonal complex 258 clade. Schematic representation based on the results of the analyses presented. Main nodes of interest are shown, highlighting the hypothesized pattern of three recombination events leading to the current state of clonal complex 258. Dates are inferred based on the molecular clock analysis depicted in Fig. 4.

In order to investigate the potential effect of the newly discovered recombination on the phenotype of the acceptor CC258, the presence of genes possibly related to antibiotic resistance and virulence was investigated in the corresponding region of the genome of strain NJST258_1, using a specifically designed database (see Materials and Methods). Interestingly, 51 genes were detected in the region (see Table S2 in the supplemental material), grouped in three main categories: LPS modification (such as the *waa* operon), bacterial efflux transporters (i.e., efflux pumps and permeases), and regulators (e.g., *ompR-envZ* operon) (see Discussion for an analysis of the detected genes).

Molecular clock. In order to date the origin of the CC258 clade and its subclades, we performed a molecular clock analysis using the software BEAST (33). We produced a reduced data set of 3,615 core SNPs present in a selected subset of taxa (174 CC258 and three closely related non-KPC *K. pneumoniae* genomes used as outgroups), derived from the previously filtered data set, in which the potentially recombined regions of the genome were excluded (Fig. 4). Compared with the dates indicated in published reports, our estimations appear to be fairly accurate. For example, the molecular clock analysis dates the appearance of ST512 to 2007, close to the first report in Israel, i.e., 2006 (20). Additionally, the molecular clock analysis dates the radiation of American and European ST258 isolates to 1997, a time point coherent with the first report of KPC-bearing *K. pneumoniae*, i.e., 1996 (2). Thus, our calibration of the evolutionary rate, superimposed on the phylogenetic tree (Fig. 4), could be used to infer unavailable dates on the global pandemic of CC258 *K. pneumoniae*. See Discussion for further discussion of the estimated dates.

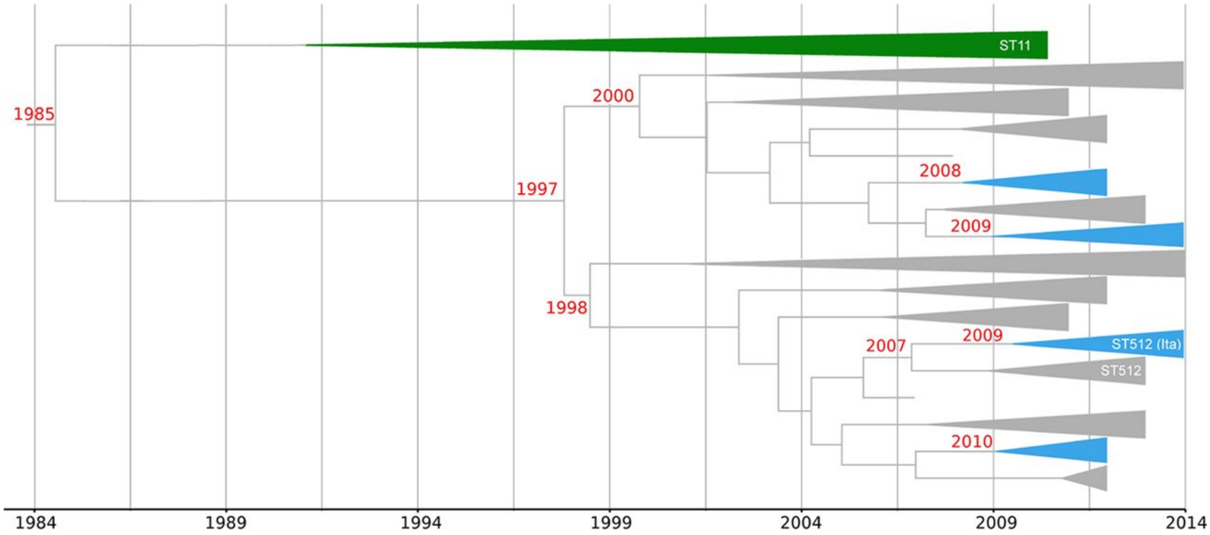


FIG 4 Estimation of divergence times in clonal complex 258. A schematic version of the time-scaled phylogeny was obtained using BEAST software with an uncorrelated log-normal relaxed clock and GTR model with no correction for site rate heterogeneity. The analysis was run for 1,000,000,000 steps, with sampling every 10,000 steps and 25% burn-in. The Italian monophyla are highlighted in blue, while the sequence type 11 (ST11) Asian clade is highlighted in green. All of the phyla with no indication of ST are comprised mainly of isolates of ST258. The dates indicated in the figure, for selected branches and nodes, were inferred from the analysis described above; for a comparison with the dates of isolation of strains, see Discussion.

Italian strains. The structure of the phylogenomic tree allows us to depict the scenario of the epidemiology of *K. pneumoniae* in Italy (Fig. 1 and 4). While susceptible and ESBL Italian strains are homogeneously distributed on the tree and belong to a number of different STs (24 and 15, respectively), all of the KPC strains sequenced in Italy belong to CC258, indicating a strong epidemiological prevalence of this clonal complex in the Italian hospitals. Within CC258, Italian isolates are well clustered in four monophyla, three composed mostly of isolates sequenced in this study and one encompassing two isolates from a previous study (38). Of the four Italian CC258 monophyla, the one including the most isolates is composed solely of ST512 (*n* = 24), confirming the multiple reports that indicate this ST as being of great epidemiological importance, at least in this country. Our phylogenetic analysis clearly indicates that this ST512 monophylum is found within the diversity of ST258.

DISCUSSION

***Klebsiella pneumoniae* in Italy.** We sequenced the genomes of 89 *K. pneumoniae* strains isolated in Italy, among them 31 KPC producers, 29 ESBL producers, and 29 strains susceptible to beta-lactams and carbapenems. Based on our phylogenomic analysis, the 29 genomes from susceptible *K. pneumoniae* strains isolated in Italy are scattered along the tree, showing no evident sign of clusterization. The sequencing of these isolates allowed us to expand the known diversity of the *K. pneumoniae* species, detecting seven novel MLST profiles and contributing to the overall robustness of current and future phylogenetic analyses. The genomes obtained from 29 ESBL isolates also show a considerable diversity, as they are distributed on the phylogenetic tree and belong to 15 different STs, among them a newly found ST.

Regarding KPC isolates, all Italian sequenced strains are found in CC258. Since no *a priori* selection of STs was performed, this result indicates a strong prevalence of CC258 among KPC *K. pneumoniae* isolates in Italy, even though isolates from different STs have been reported previously by nongenomic studies (e.g., reference 19), and a wider genomic sampling surely would allow us to obtain genomes of KPC isolates belonging to other STs. The genomes of KPC-producing *K. pneumoniae* strains isolated in Italy cluster in four monophyletic groups. If we consider that the first reported case of KPC in Italy occurred in 2008, we can use the dates obtained from the molecular clock to conclude that these monophyletic groups represent four different entrances of KPC *K. pneumoniae* in Italy (Fig. 4). This indicates that KPC strains can move effectively among different countries and continents, and that the current Italian scenario of widespread KPC resistance has been caused by multiple overlapping outbreaks. Additional sampling from Italian CC258 isolates could either confirm these results or detect novel monophyla, possibly discovering additional entrance events.

Among the four Italian CC258 monophyla, one is composed entirely of isolates of ST512. This KPC sequence type was first reported in Israel in 2006 (20) but has been spreading since then, mostly in Italy and South America (11, 17). In accordance with these reports, the four available ST512 genomes from South American isolates cluster in our phylogeny as a sister group of the Italian ST512 clade (Fig. 1 and 4). The molecular clock analysis dates the common ancestor of all members of ST512 to 2007, in relative agreement with the first report of this ST, i.e., 2006 (20). Considering that this ST is known to be a single-nucleotide variant of ST258, these results indicate that a mutational event occurred around 2006, giving rise to this sequence type, that then spread to Israel, South America, and Italy. Genome sequencing of isolates of this ST from Israel, currently unavailable, could allow us to perform phylogenetic analyses aimed at better understanding the geographical and temporal origin of the ST512 clade.

Origin of the CC258 clade. Our phylogenomic analysis, coupled with the detection of recombination events and with the molecular clock analysis, allow us to update the hypothesis regarding the origin and evolution of CC258, the most widespread bearer of KPC resistance worldwide (Fig. 3). We postulate a first recombination event that occurred before 1985 between a donor similar to ST1628 and a receiver, an ancestor of ST11. This event, which transferred a region of ~1.3 Mb to the current ST11, gave rise to the basal lineage of CC258. Since only three genomes of ST11 currently are available, all isolated from Asian patients, the current phylogeny suggests that this first recombination event occurred on the Asian continent. However, additional genome sequences of ST11 from different geographic locations are necessary to support or falsify this hypothesis. Our molecular clock analysis also can be useful to date the two subsequent, previously reported (11, 12) recombination events. The second recombination event, confirmed by our phylogenies, gave rise to ST258, having as a recipient ST11 and a donor similar to ST442 (12). Our molecular clock analysis dates this event to between 1985 and 1997. Considering that all of the known genomic CC258 diversity from the American and European continents is included within the subclade that originated in 1997 (Fig. 4), this second event could have been pivotal in the subsequent pandemic of KPC-bearing CC258. Finally, we can date the third smaller recombination event, the one that gave origin to the differentiation between the two main CC258 subclades (11), to between 1999 and 2001. Thus, we can hypothesize that these three events have produced a genomic background apt to bear and diffuse KPC plasmids, contributing to the success of the KPC pandemic.

The proposed scenario suggests that the genomic diversity of the whole *K. pneumoniae* species constitutes a reservoir of genetic variability capable of recombination events of large portions of the genome, with subsequent generation of novel variants. In this scenario, we hypothesize that large genomic recombinations are at the basis of important phenotypic/functional changes that, together with the acquisition and diffusion of plasmids bearing antibiotic resistance genes, have led to the current global epidemic. This hypothesis is supported by the multiple detected recombination events, as well as by the limited number of SNPs identified outside the recombined regions (a total of 1,086 core SNPs in the 206 analyzed CC258 genomes), and finally by the current impossibility to phenotypically differentiate the isolates of subclade ST512 from those of ST258. An alternative hypothesis is that the main reason for the diffusion of CC258 is simply the acquisition of the resistance to carbapenemic antibiotics, and that the genomic variations, whether they are recombinations or point mutations, do not provide a specific fitness benefit but are merely an example of genetic hitchhiking.

In order to investigate the importance of the recombination event described in this work, the gene content of the ~1.3-Mb region was analyzed. Fifty-one genes in this genomic context were found to be potentially related to virulence or antibiotic resistance (see Table S2 in the supplemental material). The presence of LPS synthesis genes is worth a mention because of the multiple linkages between the outer membrane and virulence (39). Genes of the operon *waa* (also known as *rfa*) are responsible for the biogenesis of the core LPS, while genes of the family *arn* control the modifications of lipid A. Modifications in membrane composition can lead to changes in surface charge and interfere with the activity of antibiotics that act on LPS, such as polymyxins and novobiocin (40). Moreover, the presence of *mla* genes in the recombined region is worth being highlighted. These genes are presumed to maintain lipid asymmetry in the Gram-negative outer membrane, as they transport phospholipids to the inner side of the membrane. *mla* genes were reported as virulence factors in *Escherichia coli* and in other Gram-negative bacteria, as mutations in these genes can lead to a change in the permeability of the outer membrane and to a subsequent variation in virulence (41). The presence of fumarate reductase genes of the family *fmr* in the recombined region suggests a link with the variation of virulence of CC258. In fact, fumarate reductase is a virulence determinant in *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Actinobacillus pleuropneumoniae*, and *Salmonella enterica*, as mutants of these genes show variations in virulence (32). Finally, the *ompR*-*envZ* operon, present in the recombined region, is a two-component system that acts as a transcription regulator, affecting the expression of the genes *ompF* and *ompC* (42). Mutations in the *ompR* and *envZ* genes have been shown to reduce the expression of outer membrane porins OmpF and OmpC (43). This in turn can have drastic effects on both the virulence and antibiotic resistance of mutant strains. It has been reported in particular that *OmpR* mutations can lead to reduced susceptibility to carbapenemic antibiotics in *Enterobacteriaceae* (44).

Further functional investigations aimed at unveiling the reasons for the success of the CC258 clade, possibly focusing on the detected recombinant regions, would greatly improve our understanding of the *K. pneumoniae* pandemic and would provide important tools in the fight against KPC-producing strains. Finally, our conclusions should lead to additional studies focused on the recombination potential of other STs of *K. pneumoniae*. If this capacity were found to be widespread, we should be aware that future recombination events could lead to the diffusion of novel epidemic clones.

REFERENCES

1. Knothe H, Shah P, Krcmery V, Antal M, Mitsuhashi S. 1983. Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of Klebsiella pneumoniae and Serratia marcescens. Infection 11:315–317. 10.1007/BF01641355.

2. Yigit HA, Queenan M, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, Alberti S, Bush K, Tenover FC. 2001. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. Antimicrob Agents Chemother 45:1151–1161. 10.1128/AAC.45.4.1151-1161.2001.

3. Woodford N, Tierno PM, Young K, Tysall L, Palepou MF, Ward E, Painter RE, Suber DF, Shungu D, Silver LL, Inglima K, Kornblum J, Livermore DM. 2004. Outbreak of Klebsiella pneumoniae producing a new carbapenem-hydrolyzing class A beta-lactamase, KPC-3, in a New York medical center. Antimicrob Agents Chemother 48:4793–4799. 10.1128/AAC.48.12.4793-4799.2004.

4. Bratu S, Landman D, Haag R, Recco R, Eramo A, Alam M, Quale J. 2005. Rapid spread of carbapenem-resistant Klebsiella pneumoniae in New York City: a new threat to our antibiotic armamentarium. Arch Intern Med 165:1430–1435. 10.1001/archinte.165.12.1430.

5. Nordmann P, Cuzon G, Naas T. 2009. The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. Lancet Infect Dis 9:228–236. 10.1016/S1473-3099(09)70054-4.

6. Cantón R, Akóva M, Carmeli Y, Giske CG, Glupczynski Y, Gniadkowski M, Livermore DM, Miriagou V, Naas T, Rossolini GM, Samuelsen Ø, Seifert H, Woodford N, Nordmann P. 2012. Rapid evolution and spread of carbapenemases among Enterobacteriaceae in Europe. Clin Microbiol Infect 18:413–431. 10.1111/j.1469-0691.2012.03821.x.

7. Richter SN, Frasson I, Bergo C, Parisi S, Cavallaro A, Palù G. 2011. Transfer of KPC-2 carbapenemase from Klebsiella Pneumoniae to Escherichia Coli in a patient: first case in Europe. J Clin Microbiol 49:2040–2042. 10.1128/JCM.00133-11.

8. Luo Y, Yang J, Ye L, Guo L, Zhao Q, Chen R, Chen Y, Han X, Zhao J, Tian S, Han L. 2014. Characterization of KPC-2-producing Escherichia coli, Citrobacter freundii, Enterobacter cloacae, Enterobacter aerogenes, and Klebsiella oxytoca isolates from a Chinese hospital. Microb Drug Resist 4:264–269. 10.1089/mdr.2013.0150.

9. Chen L, Chavda KD, Melano RG, Hong T, Rojtman AD, Jacobs MR, Bonomo RA, Kreiswirth BN. 2014. A molecular survey of the dissemination of two blaKPC-harboring IncFIA plasmids in New Jersey and New York hospitals. Antimicrob Agents Chemother 58:2289–2294. 10.1128/AAC.02749-13.

10. Andrade LN, Curiao T, Ferreira JC, Longo JM, Clímaco EC, Martinez R, Bellissimo-Rodrigues F, Basile-Filho A, Evaristo MA, Del Peloso PF, Ribeiro VB, Barth AL, Paula MC, Baquero F, Cantón R,

Darini AL, Coque TM. 2011. Dissemination of blaKPC-2 by the spread of Klebsiella pneumoniae clonal complex 258 clones (ST258, ST11, ST437) and plasmids (IncFII, IncN, IncL/M) among Enterobacteriaceae species in Brazil. Antimicrob Agents Chemother 55:3579–3583. 10.1128/AAC.01783-10.

11. Deleo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, Chavda KD, Jacobs MR, Mathema B, Olsen RJ, Bonomo RA, Musser JM, Kreiswirth BN. 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae. Proc Natl Acad Sci U S A 111:4988–4993. 10.1073/pnas.1321364111.

12. Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. 2014. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. mBio 5:e01355-14. 10.1128/mBio.01355-14.

13. Pagani L, Ronza P, Giacobone E, Romero E. 1994. Extended-spectrum beta-lactamases from Klebsiella pneumoniae strains isolated at an Italian hospital. Eur J Epidemiol 10:533–540. 10.1007/BF01719569.

14. Perilli M, Dell'Amico E, Segatore B, de Massis MR, Bianchi C, Luzzaro F, Rossolini GM, Toniolo A, Nicoletti G, Amicosante G. 2002. Molecular characterization of extended-spectrum beta-lactamases produced by nosocomial isolates of Enterobacteriaceae from an Italian nationwide survey. J Clin Microbiol 40:611–614. 10.1128/JCM.40.2.611-614.2002.

15. D'Andrea MM, Arena F, Pallecchi L, Rossolini GM. 2013. CTX-M-type β-lactamases: a successful story of antibiotic resistance. Int J Med Microbiol 303:305–317. 10.1016/j.ijmm.2013.02.008.

16. Giani T, D'Andrea MM, Pecile P, Borgianni L, Nicoletti P, Tonelli F, Bartoloni A, Rossolini GM. 2009. Emergence in Italy of Klebsiella pneumoniae sequence type 258 producing KPC-3 carbapenemase. J Clin Microbiol 47:3793–3794. 10.1128/JCM.01773-09.

17. Gaibani P, Ambretti S, Berlingeri A, Gelsomino F, Bielli A, Landini MP, Sambri V. 2011. Rapid increase of carbapenemase-producing Klebsiella pneumoniae strains in a large Italian hospital: surveillance period 1 March–30 September 2010. Euro Surveill 16:19800.

18. Comandatore F, Gaibani P, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Sassera D. 2013. Draft genome of Klebsiella pneumoniae sequence type 512, a multidrug-resistant strain isolated during a recent KPC outbreak in Italy. Genome Announc 1:e00035-12. 10.1128/genomeA.00035-12.

19. Giani T, Pini B, Arena F, Conte V, Bracco S, Migliavacca R, Pantosti A, Pagani L, Luzzaro F, Rossolini GM. 2013. Epidemic diffusion of KPC carbapenemase-producing Klebsiella pneumoniae in Italy: results of the first countrywide survey, 15 May to 30 June 2011. Euro Surveill 18:20489.

20. Warburg G, Hidalgo-Grass C, Partridge SR, Tolmasky ME, Temper V, Moses AE, Block C, Strahilevitz J. 2012. A carbapenem-resistant Klebsiella pneumoniae epidemic clone in Jerusalem:

sequence type 512 carrying a plasmid encoding aac(6′)-Ib. J Antimicrob Chemother 67:898–901. 10.1093/jac/dkr552.

21. Doyle D, Peirano G, Lascols C, Lloyd T, Church DL, Pitout JD. 2012. Laboratory detection of Enterobacteriaceae that produce carbapenemases. J Clin Microbiol 50:3877–3880. 10.1128/JCM.02117-12.

22. Poirel L, Walsh TR, Cuvillier V, Nordmann P. 2011. Multiplex PCR for detection of acquired carbapenemase genes. Diagn Microbiol Infect Dis 70:119–123. 10.1016/j.diagmicrobio.2010.12.002.

23. Clinical and Laboratory Standards Institute. 2011. Performance standards for antimicrobial susceptibility testing; 21st informational supplement. CLSI M100-S121, vol 31. Clinical and Laboratory Standards Institute, Wayne, PA.

24. European Committee on Antimicrobial Susceptibility Testing. 2014. Breakpoint tables for interpretation of MICs and zone diameters. Version 4.0. http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/Breakpoint_table_v_4.0.pdf.

25. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. Proceedings of the 1999 German Conference on Bioinformatics.

26. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. 10.1101/gr.074492.107.

27. Darling AE, Mau B, Perna NT. 2010. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. 10.1371/journal.pone.0011147.

28. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. 10.1093/bioinformatics/btu033.

29. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res 40:e6. 10.1093/nar/gkr928.

30. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 57:3348–3357. 10.1128/AAC.00419-13.

31. Liu B, Pop M. 2009. ARDB-Antibiotic Resistance Genes Database. Nucleic Acids Res 37:D443–D447. 10.1093/nar/gkn656.

32. Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, Barbe V, Bengoechea JA, Sansonetti P, Brisse S, Tournebize R. 2014. Comparative analysis of Klebsiella pneumoniae genomes identifies a phospholipase D family protein as a novel virulence factor. BMC Biol 12:41. 10.1186/1741-7007-12-41.

33. Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. 10.1186/1471-2148-7-214.

34. Castillo-Ramirez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ. 2011. The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7:e1002129. 10.1371/journal.ppat.1002129.

35. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. 2005. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. J Clin Microbiol 43:4178–4182. 10.1128/JCM.43.8.4178-4182.2005.

36. Villa L, Feudi C, Fortini D, García-Fernández A, Carattoli A. 2014. Genomics of KPC-producing Klebsiella pneumoniae sequence type 512 clone highlights the role of RamR and ribosomal S10 protein mutations in conferring tigecycline resistance. Antimicrob Agents Chemother 58:1707–1712. 10.1128/AAC.01803-13.

37. Wright MS, Perez F, Brinkac L, Jacobs MR, Kaye K, Cober E, van Duin D, Marshall SH, Hujer AM, Rudin SD, Hujer KM, Bonomo RA, Adams MD. 2014. Population structure of KPC-producing Klebsiella pneumoniae isolates from midwestern U.S. hospitals. Antimicrob Agents Chemother 58:4961–4965. 10.1128/AAC.00125-14.

38. Comandatore F, Sassera D, Ambretti S, Landini MP, Daffonchio D, Marone P, Sambri V, Bandi C, Gaibani P. 2013. Draft genome sequences of two multidrug resistant Klebsiella pneumoniae ST258 isolates resistant to colistin. Genome Announc 1:e00113–12. 10.1128/genomeA.00113-12.

39. Heinrichs DE, Yethon JA, Whitfield C. 1998. Molecular basis for structural diversity in the core regions of the lipopolysaccharides of Escherichia coli and Salmonella enterica. Mol Microbiol 30:221–232. 10.1046/j.1365-2958.1998.01063.x.

40. Goldberg JB. 1999. Genetics of bacterial polysaccharides. CRC Press, London, United Kingdom.

41. Malinverni JC, Silhavy TJ. 2009. An ABC transport system that maintains lipid asymmetry in the gram-negative outer membrane. Proc Natl Acad Sci U S A 12:8009–8014. 10.1073/pnas.0903229106.

42. Buckler DR, Anand GS, Stock AM. 2000. Response-regulator phosphorylation and activation: a two-way street? Trends Microbiol 8:153–156. 10.1016/S0966-842X(00)01707-8.

43. Yuan J, Wei B, Shi M, Gao H. 2011. Functional assessment of EnvZ/OmpR two-component system in Shewanella oneidensis. PLoS One 6:e23701. 10.1371/journal.pone.0023701.

44. Tängdén T, Adler M, Cars O, Sandegren L, Löwdin E. 2013. Frequent emergence of porin-deficient subpopulations with reduced carbapenem susceptibility in ESBL-producing Escherichia coli during exposure to ertapenem in an in vitro pharmacokinetic model. J Antimicrob Chemother 68:1319–1326. 10.1093/jac/dkt044.

ARTICLE 2

Insertion sequences proliferation and limited genomic plasticity in *Acinetobacter baumannii* sequence type 78, a persistent clone in Italian hospitals

Submitted to: Genome Biology and Evolution

Insertion sequences proliferation and limited genomic plasticity in *Acinetobacter baumannii* sequence type 78, a persistent clone in Italian hospitals

Stefano Gaiarsa[a,b], Ibrahim Bitar[c], Francesco Comandatore[d], Marta Corbella[a], Aurora Piazza[c,d], Erika Scaltriti[e], Laura Villa[f], Piero Marone[a], Laura Pagani[c], Stefano Pongolini[e], Roberta Migliavacca[c], and Davide Sassera[g]

Author Affiliations

[a] S.C. Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

[b] Dipartimento di Bioscienze, Università degli Studi di Milano, Milan, Italy

[c] Dipartimento di Scienze Clinico chirurgiche, Diagnostiche e Pediatriche, Università degli Studi di Pavia, Pavia, Italy.

[d] Pediatric Clinical Research Center, Università degli Studi di Milano, Milan, Italy

[e] Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Sezione di Parma, Parma, Italy

[f] Dipartimento di Malattie Infettive, Parassitarie ed Immunomediate, Istituto Superiore di Sanitá , Rome, Italy

[g] Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy

Address correspondence to Davide Sassera, davide.sassera@unipv.it.

Data deposition: this project has been deposited at EMBL EBI under the study accession number PRJEB19248, while plasmid sequences are deposited NCBI genbank as accession numbers KY202456, KY202457, KY202458 and KY652669

Supplemental material is available at the end of the thesis

ACKNOWLEDGMENTS. The authors would like to thank Raffaele Zarrilli for providing the 3909 sample and Jason Sahl for providing metadata on the strains TG22142, TG22146 and TG22150. SG, FC and DS would also like to thank Rossella Matteri and Davide Boselli for help and inspiration.

Keywords: ***Acinetobacter baumannii*, genomic plasticity, ST78, insertion sequences, SMAL**

ABSTRACT

Acinetobacter baumannii is a known opportunistic pathogen. Its genome has been described as characterized by a very high plasticity, with high frequency of homologous recombinations and proliferation of insertion sequences. The SMAL pulsotype is an *A. baumannii* strain putatively isolated only in Italy, characterized by a low incidence and a high persistence over the years. In the present work, we have conducted a comparative genomic analysis on this clone. All genomes presented the Sequence Type 78 (ST78) and were analysed in comparison with 11 other assemblies of the same ST. The phylogeny highlighted the presence of two different clades, one of which (ST78A) encompasses all the SMAL genomes and three others. ST78A resulted to have a low rate of homologous recombination and low gene content variability. Surprisingly, genomes inside the clade present a high number of Insertion Sequences (IS), mostly absent in the other genomes of the ST. Among these IS, one IS66 was found to interrupt the gene *comEC/rec2*, involved in the acquisition of exogenous DNA. This leads to the depiction of an evolutionary scenario in which the proliferation of IS is slowing the acquisition of exogenous DNA, thus limiting genome plasticity. Such genomic architecture can explain the epidemiological behaviour of high persistence and low incidence of the clone, and provides an interesting framework to compare ST78 with the highly epidemic international clones, characterized by high genomic plasticity.

INTRODUCTION

The bacterium *Acinetobacter baumannii* is an opportunistic pathogen diffused worldwide, part of the so called ESKAPE group of microbial threats of our age (1). It has emerged in recent decades as a clinically relevant pathogen causing a wide range of both nosocomial and community-acquired infections, including injured soldiers (2), also thanks to its ability to colonize skin, plastic intravascular devices and mucous membranes and to survive in the hospital environment (3). Indeed *A. baumannii* was also known as "Iraqibacter", when it became a major threat for troops on a mission in the Middle East (4).

The genome of hundreds of *A. baumannii* isolates have been sequenced, showing a strong level of genome plasticity, in the form of a tendency to undergo frequent and substantial rearrangements, including recombinations (5, 6) and movement of insertion sequences (IS) (7).

Generally, proliferation of IS elements in bacteria brings genomic variability, which can lead to adaptation in a new niche. Thus, bacterial genomes with higher numbers of IS copies have been observed to lead to virulent clones in multiple species (8,9) which often have a successful worldwide spread (10,11). Moreover, IS elements act as anchors for homologous recombination process, leading to internal genome rearrangements but also very often to the inclusion of exogenous DNA in the chromosome. Such homologous recombination events were found to be crucial for the evolution and adaptivity of *A. baumannii* and other pathogens; therefore, this genomic feature granted the bacterium a place among the so-called "bacterial hopeful monsters", a set of microorganisms able to rapidly modify their genotype through recombination events, and thus capable of quickly adapting to novel environmental conditions (12). The role of ISs in *A. baumannii* has been investigated mainly for what concerns specific classes. For example, the presence of the Insertion Sequence IS*Aba1* upstream the gene encoding for the beta-lactamase OXA-51 grants *A. baumannii* resistance to carbapenemic antibiotics. The gene *bla*$_{\text{OXA-51}}$ is always found in clinical isolates of *A. baumannii*, but only through the presence of the insertion sequence it can confer the bacterium the resistant phenotype (13).

The most commonly isolated strains of *A. baumannii* in Europe belong to the two main European Clones, i.e. ECI and ECII (also known as International or Global Clones, IC or GC). In the early 2000s these two clones were the first discovered to carry the plasmid-encoded gene *bla*$_{\text{OXA-58}}$, which confers resistance to carbapenems. More recently, strains of both European Clones have been reported to carry another determinant of resistance to carbapenems, the gene *bla*$_{\text{OXA-23}}$, which can be either plasmid or chromosome borne (14). While report of *bla*$_{\text{OXA-23}}$ strains are increasing, the gene *bla*$_{\text{OXA-58}}$ is being found less commonly in clinical isolates. This has been hypothesized to be due to the fitness cost that the gene carries (15, 16). Strong evidence of recombination events and movement of IS has been found in the European Clones, not rarely related to a change in antimicrobial resistance pattern (7, 17, 18).

In the last 15 years, a different clone of *A. baumannii,* not evolutionary related to the two main European Clones, has been isolated multiple times in Italian hospitals. The clone was identified by Pulse Field Gel Electrophoresis (PFGE) and named SMAL (based on the hospitals from which it was first isolated, San Matteo and Salvatore Maugeri Acute care and Long term care facilities, respectively) (19). Concurrently it was characterized by MultiLocus Sequence Typing (MLST) as ST78, by multiplex-PCR as Sequence Group 6 (20, 21), by repetitive-sequence-based PCR (rep-PCR) as type 3 and by Amplification Fragment Length Polymorfism (AFLP^{TM}) analysis as type 21; finally termed "Italian Clone" (22). Within a National survey on the spread of carbapenem-resistant A. baumannii strains, the majority of the isolates (n=52) belonged to ICII/ST2; however, 3/55 genotyped strains showed a SMAL pulsotype (15). Despite the low number of isolates reported, the SMAL Clone represents an endemic reality in Italy since its first detection, in 2002 (Migliavacca, personal communication).

SMAL strains often show resistance to carbapenems in *in vitro* testing. The mechanisms underlying such phenotype are heterogeneous: mainly the overexpression of the chromosomic *bla*$_{OXA-51}$-like gene or the presence of acquired resistance determinants (e.g. *bla*$_{OXA-58}$ and *bla*$_{OXA-23}$). The firstly collected SMAL isolates were characterized by the presence of a *bla*$_{OXA-58}$ acquired determinant; such resistance gene was then almost completely replaced in 2009 by a *bla*$_{OXA-23}$ gene (19, 22). Therefore, the SMAL Clone seems appears to have followed the same evolutionary response (in terms of resistance genes acquisition) of the European Clones I and II (23).

SMAL isolates showing Minimum Inhibitory Concentration (MIC) values for carbapenems under clinical susceptibility breakpoint could be very difficult to treat in case of localized and device associated infections. Therapeutic failures can be at least partially due to the SMAL biofilm forming ability. This feature has been well demonstrated in previous studies, which showed how the SMAL biofilm production was considerably higher in comparison with the ATCC19606 reference strain (19, 22).

The purpose of the present work was to characterize, from the genomic point of view, 15 isolates belonging to the SMAL pulsotype, investigating the possible links between the phenotypic and genetic features of this Clone.

RESULTS

Genome sequencing and assembly. The total DNA of 15 isolates of *A. baumannii* previously recognized as SMAL was sequenced and assembled resulting in draft genomes with an average of 193 contigs above 1000bp, an average total length of 4,293,267 bp and an average N50 of 48,860 bp. Draft genomes are published on EBI-EMBL under the study accession number PRJEB19248. One plasmid was detected in each of four genomes (i.e. 14336, 2RED09, 20C15 and 5MO, respective accession numbers KY202456, KY202457, KY202458 and KY652669).

Global phylogeny of the species. All available genomes of the species *A. baumannii* were retrieved from the NCBI database and joined with those of the isolates presented in this work, to obtain a database of 1,412 genomes. Genomes were aligned, SNPs were called in order to perform a Maximum Likelihood phylogeny. The 15 SMAL isolates were *in silico* assigned to the ST78 and resulted to be clustered on the global tree in a single monophylum, together with 11 database genomes of the same ST (Supplementary Figure 1).

Phylogeny of the ST78. The ST78 monophylum was selected to be further investigated. Core SNPs were called and used to create a 2,760 core SNP alignment (945 informative sites) to perform a phylogeny of the ST78 (Figure 1). The ST78 monophylum was split in two well supported monophyletic groups (called ST78A and ST78B) plus a single evolutionary distant genome (strain ABBL025). The 15 novel genomes were clustered in a smaller single monophylum, together with strain 3909, within ST78A. Analysis of available metadata (see Table 1) revealed that strain 3909 was isolated in Italy in 2007 (24). The isolate was obtained from the hospital of origin and analyzed with PFGE. The resulting pattern of migration showed that isolate 3909 belongs to the pulsotype SMAL as well. For this reason, from now on, the 16 genomes in this monophylum will be addressed as the SMAL cluster. Within ST78A, three other genomes isolated in Arizona (USA) in 2011 (J. Sahl, personal communication) form a highly supported clade that is positioned as a closely evolutionarily related sister group of the SMAL cluster. SNP distribution along the genome was plotted in order to spot possible recombinations (See Figure 2). One ~34 Kbp region with higher SNPs density was detected, and identified as common to the genomes of both main clades (i.e. ST78A and ST78B), indicating that it could represent a recombination of the common ancestor of the two clades, or of the sole genome ABBL025.

Strain	Year of isolation	Country of Isolation	City/State/Institute of isolation	Genome length (bp)	N of predicted ORFs
3909	2007	Italy	Napoli, Ospedale Monaldi	3948828	3719
14336	2010	Italy	Firenze, Ospedale Careggi	3961089	3682
831240	NA	USA	NA	3970899	3700
855125	NA	USA	NA	4401277	4109
1096934	NA	USA	NA	4330037	4010
103SM	2012	Italy	Pavia, Policlinico San Matteo	3984084	3705
20C15	2011	Italy	Napoli, Ospedale Cardarelli	4014203	3749
25C30	2011	Italy	Catania, Policlinico di Catania	4002419	3722
2MG	2012	Italy	Pavia, Fondazione Salvatore Maugeri	4028799	3760
2RED09	2009	Italy	Milano, Istituto Geriatrico “P. Redaelli”	3983268	3710
5MO	2009	Italy	Monza, Ospedale San Gerardo	4026195	3760
61SM01	2006	Italy	Pavia, Policlinico San Matteo	4020527	3733
65SM01	2006	Italy	Pavia, Policlinico San Matteo	4002746	3717
68SM01	2007	Italy	Pavia, Policlinico San Matteo	4015623	3738
72SM01	2007	Italy	Pavia, Policlinico San Matteo	4001559	3726
74SM01	2007	Italy	Pavia, Policlinico San Matteo	3969850	3681
96SM	2012	Italy	Pavia, Policlinico San Matteo	4010167	3739
ABBL025	2006	USA	Chicago (IL)	4067524	3765
ABBL026	2006	USA	Chicago (IL)	3934606	3660
MGTN	2004	Italy	Pavia, Fondazione Salvatore Maugeri	4002517	3722
MONUR	2004	Italy	Pavia, Fondazione Salvatore Maugeri	4001892	3721
TG22142	2011	USA	Arizona	4229994	3937
TG22146	2011	USA	Arizona	4240225	3978
TG22150	2011	USA	Arizona	3978141	3708
UH1752	2007	USA	Cleveland (OH)	4007789	3740
UH5207	2007	USA	Cleveland (OH)	4009458	3739

Table 1. Microbiologic and genomic information on the strains of the Sequence Type 78. From left to right are reported: name of the isolate, year of isolation, country, city and institution (where available) of isolation, size of the complete genome (in base pairs), and number of predicted ORFs.

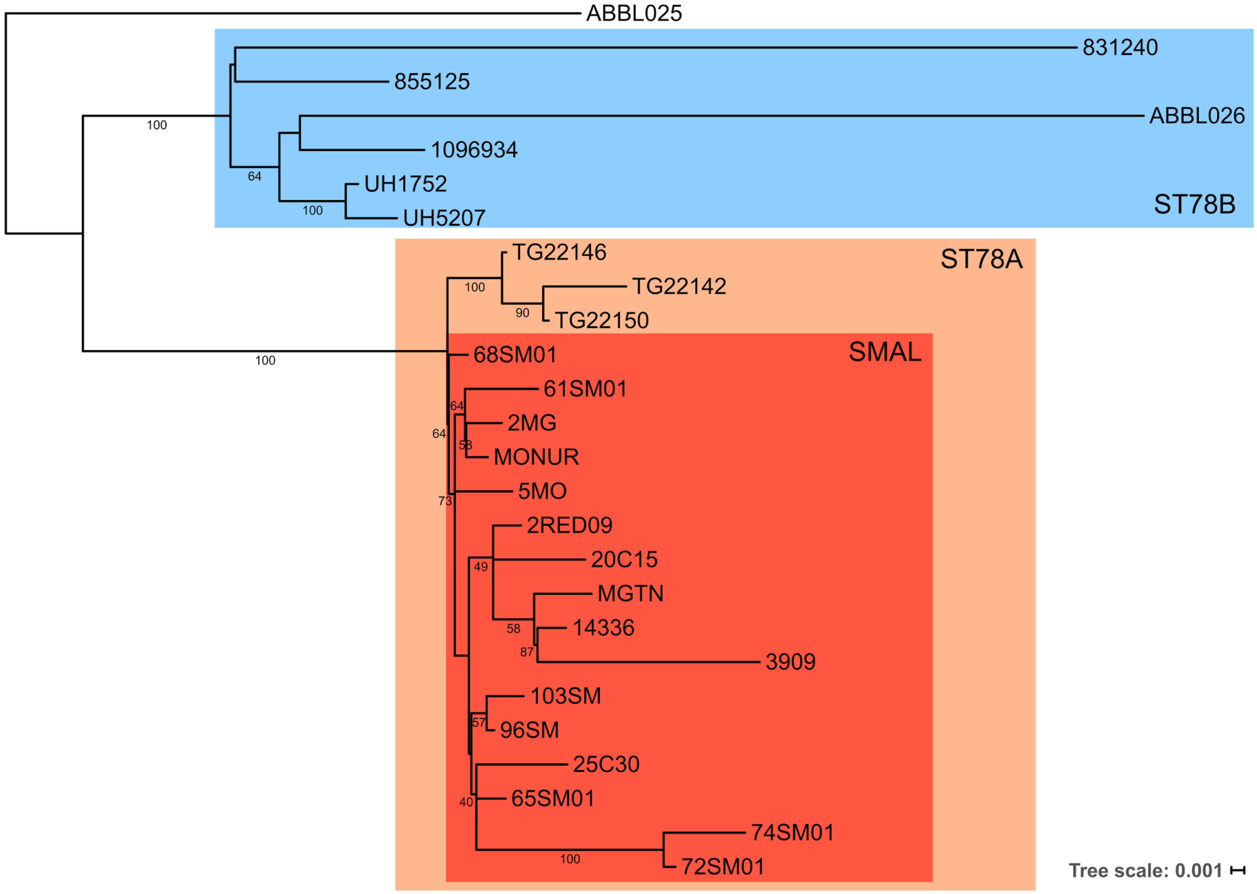


Figure 1. Phylogeny of the Sequence Type 78 obtained with the software RAxML on a dataset of core Single Nucleotide Polymorphisms. Bootstrap values over 40 are reported next to the relative node. The main clades are highlighted with colored boxes.

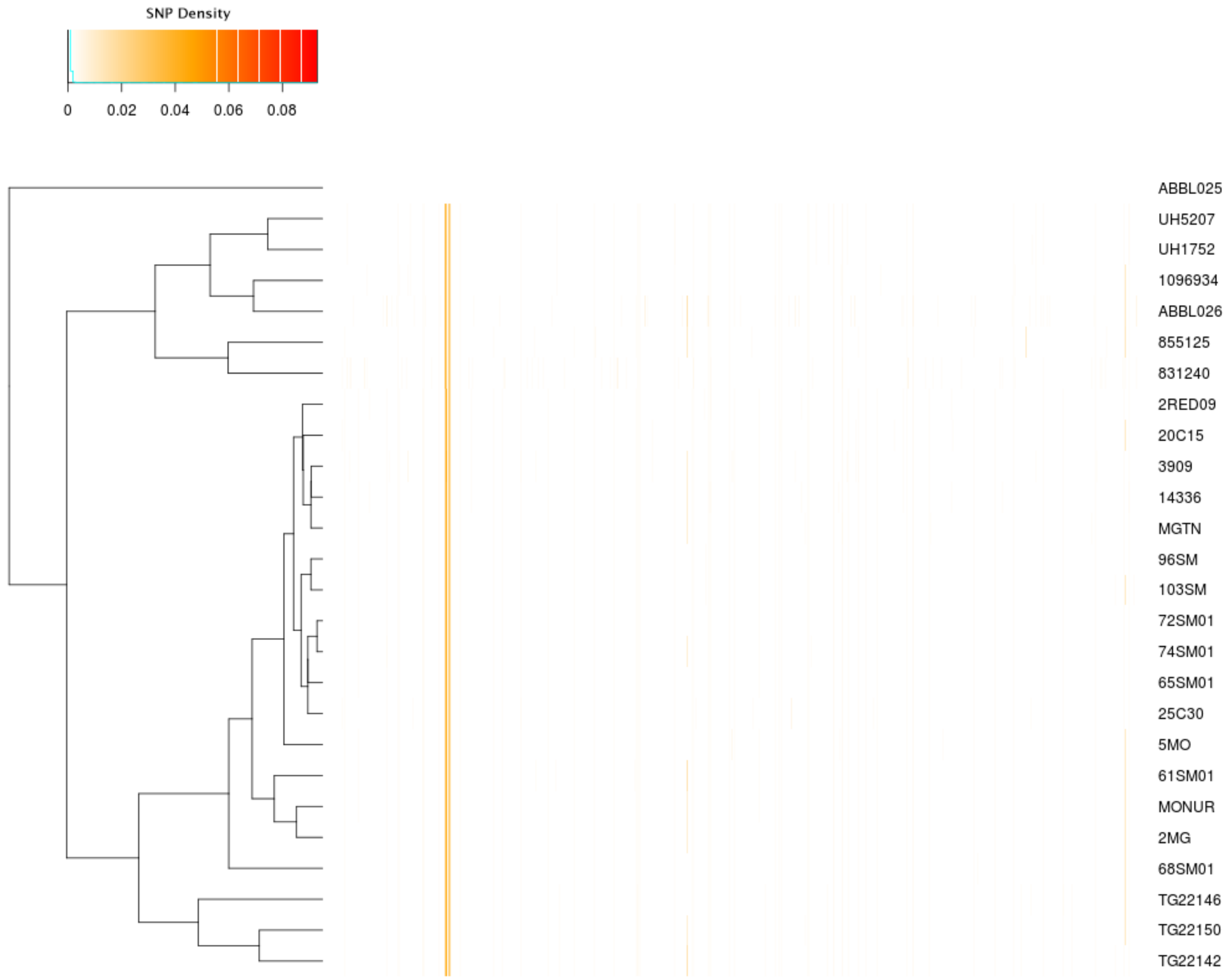


Figure 2. Concentration of core Single Nucleotide Polymorphisms, calculated in windows of 1000bp along the whole genome (using the genome of strain ABBL025 as a reference). Genomes are ordered following the phylogenetic tree of Sequence Type 78, which is reported on the left.

Phenotypic characterization. SMAL clade susceptibility profiles were determined using routine diagnostic automated systems; MIC values were assessed by E-test. Six out of 16 strains (3909, 2RED09, 14336, 5MO, 20C15 and 25C30) showed an intermediate/resistant phenotype to both Meropenem (MER) and Imipenem (IPM), with MIC values ranging from 6 to >32 mg/L (MIC50= >32 mg/L). Susceptibility was retained in 7/16 and 9/16 strains for MER and IPM, respectively. Full results of resistance profiles are reported in Supplementary Table 1.

Biofilm formation was tested for all SMAL cluster strains and no significant difference was assessed; all strains showed a strong biofilm formation ability, higher than that comparatively observed for the reference strain ATCC19606 (19). Averages of the three replicate values ranged from 0.4 to 0.7 OD_{600}. Full results are reported in Supplementary Table 1.

Analysis of gene content. Coding sequences were called and ortholog proteins were predicted in the 26 genomes of ST78. Results indicate that the gene content of the SMAL isolates is highly conserved (gene dispersion of 37.38 genes/taxon), while the dispersion of the whole ST78 is 73.58 genes/taxon. Moreover, the three Arizonan isolates, besides having a very low dispersion themselves (30.67 genes/taxon), when combined with the SMAL genomes (thus the ST78A clade), show a dispersion of 35.84 genes/taxon. Distribution of the accessory genes was plotted on the tree (Figure 3A). The high internal and reciprocal similarity of the two genome clusters is clearly shown, and it is strongly highlighted in the second plot, where a distance analysis of the accessory gene content was run, using gene presences as characters (Figure 3B). We thus wondered if the limited gene content variation of ST78A was indeed indicative of low genomic plasticity, or if it was due to a small evolutionary distance. To address this question, the phylogenetic distance between each pair of genomes of the two main clusters (ST78A and ST78B) was calculated and plotted against the binary distance of gene content (Figure 3C). The plot shows that the lower gene presence distance of ST78A does not correspond to a lower phylogenetic distance, supporting the hypothesis of a lower genomic plasticity of ST78A.

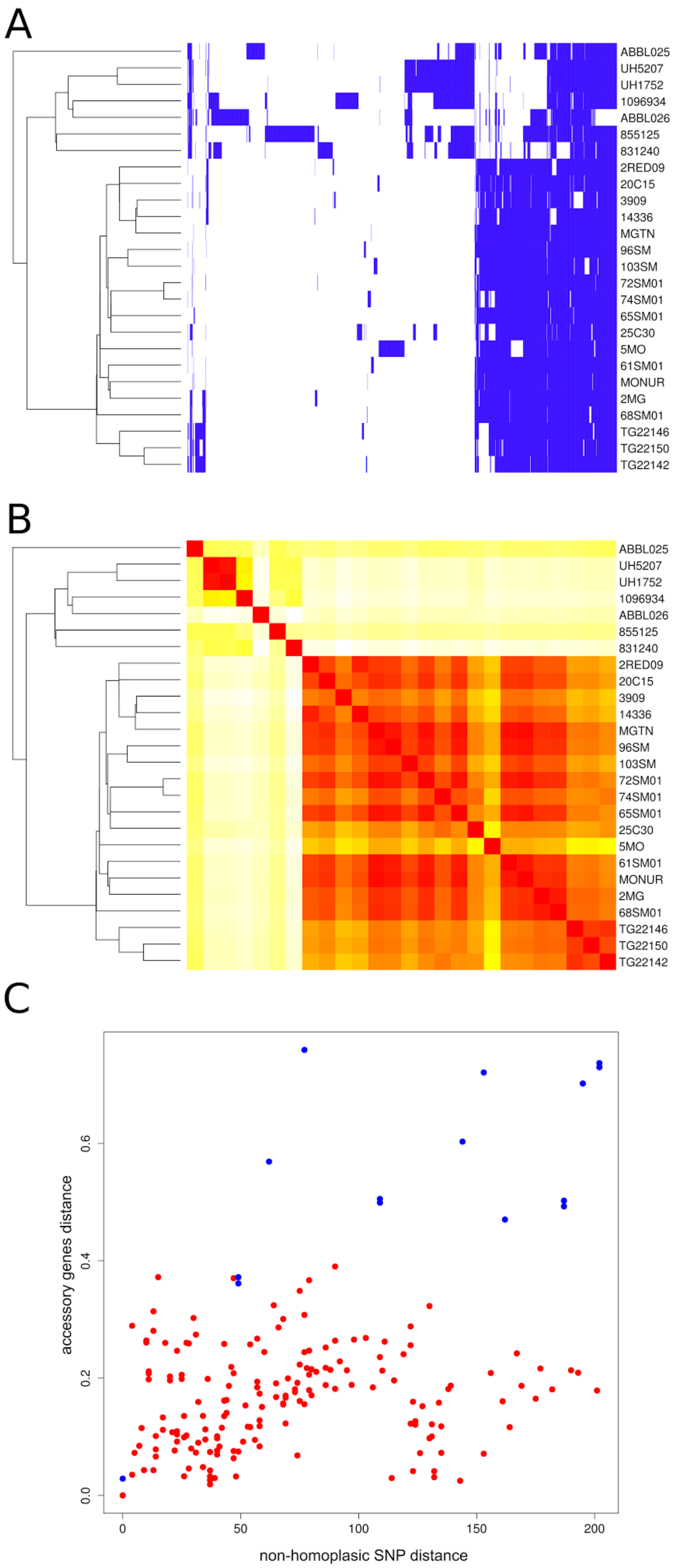


Figure 3. Representation of presence of accessory genes in the Sequence Type 78. A) From left to right: the phylogeny of the sequence type (with tips aligned) and the matrix of gene presence (blue squares are detected genes); B) From left to right: the phylogeny of the sequence type (with tips aligned) and the distance matrix of the genomes, obtained using gene presence as characters; C) Plot of the intergenomic distances. Each genome pair is represented on the X axis by their phylogenetic distance (expressed as number of non-homoplasic core SNP between the two genomes) and on the Y axis by the gene content distance (expressed by the binary distance between the two genomes, calculated on a matrix of accessory gene presence). Genome pairs inside the ST78A clade are reported in red; pairs inside the ST78B clade are reported in blue

Resistome. Resistance gene content of the 26 genomes of ST78 was analyzed by manually curated Blast alignments against custom and public databases. Full results are reported in Supplementary Table 2. Genes encoding the intrinsic OXA-51-like beta-lactamase were detected in all the ST78 genomes. One of the novel genomes presents a previously unknown variant of this gene, that was assigned the code *bla*$_{\text{OXA-545}}$. This result prompted to test the presence and genomic position of the known resistance-enhancer insertion sequence IS*Aba1*. Such IS is present in all the genomes of the ST78A cluster, while absent in all the other ST78 genomes, except UH1752. The sequence was found to be upstream the resistance gene *bla*$_{\text{OXA-90}}$ in five genomes (25C30, 103SM, TG22142, TG22146, TG22150). The carbapenem-resistance gene *bla*$_{\text{OXA-58}}$ was found to be encoded in four genomes of the dataset, all belonging to the SMAL cluster (2RED09, 20C15, 14336, and 3909). These genomes are clustered with a fifth one (MGTN) *bla*$_{\text{OXA-58}}$ negative, in a low supported monophyletic clade. The gene *bla*$_{\text{OXA-23}}$, instead, was detected in the genome of strains 5MO and 20C15. This pattern of gene presence is in accordance with the results of MIC tests. Indeed, five of the six intermediate/resistant strains carry genes encoding for carbapenemases and one (25C30) has the IS*Aba1* enhancer upstream the *bla*$_{\text{OXA-90}}$ gene. Surprisingly, the isolate 103SM, which also presents the resistance enhancing genotype, resulted to have a susceptible phenotype.

The gene *bla*$_{\text{ADC-52}}$ was detected in all the 26 genomes in analysis. In the ST78A cluster, the gene presents an unpublished single-SNP mutation. Furthermore, the *carO* gene and the *adeS*-*adeR* regulation system are present in all genomes. The sequence of *adeS* was found to be interrupted by the end of the contig in the assemblies of five strains (3909, 855125 and the three Arizonan ones) suggesting a possible presence of an insertion sequence and a consequent loss of function.

Detection of other genes of interest. Gene content of the ST78 genomes was further investigated for the presence of virulence factors, competence and biofilm formation capability. Full results are reported in Supplementary Table 2. Virulence genes present a very conserved distribution. In fact, 26 genes, including the whole set of pilum-related *pil* genes and the transporter-encoding *ptk* are present in all the ST78 genomes. Notable exceptions are the genes *cap8J* (present only in the genomes of the ST78A clade) and *epsA* (present in the seven genomes of the ST78B clade). Twelve Genes coding for biofilm formation were also searched, detecting multiple point mutations and two frame-shifting insertions, thus showing no particular pattern relationship to the phylogeny. Based on the results of the *in vitro* biofilm formation assay (i.e. all isolates are strong producers) we can conclude that none of the detected mutation inhibit biofilm formation.

Lastly, all genomes in the ST78A cluster present the genes of the O-antigen group A. The remaining seven genomes present matches with some genes of the group C but the evidence was not strong enough to assign the classification.

Analysis of mobilome. All 26 genomes were analyzed for presence of insertion sequences (IS) obtaining the results depicted in Figure 4 (full results reported in Supplementary Table 3). Interestingly, the ST78A contains strikingly more IS than the others. Furthermore, the 19 genomes in this cluster possess a set of exclusive IS classes which is absent in the other seven organisms.

This result prompted us to investigate whether the genes of interest that were found to be truncated, were interrupted by IS. The interruption in the competence gene *comEC/rec2* was found to be caused by an insertion sequence of the class IS66 in all the genomes of the SMAL clade. The gene *adeS,* interrupted in five strains, was found to have an insertion of IS66 in the three Arizonan genomes. For the other two genomes, the assembly did not allow to retrieve the sequence that interrupts the genes *adeS* and *ompF* because the contigs end with the interrupted gene.

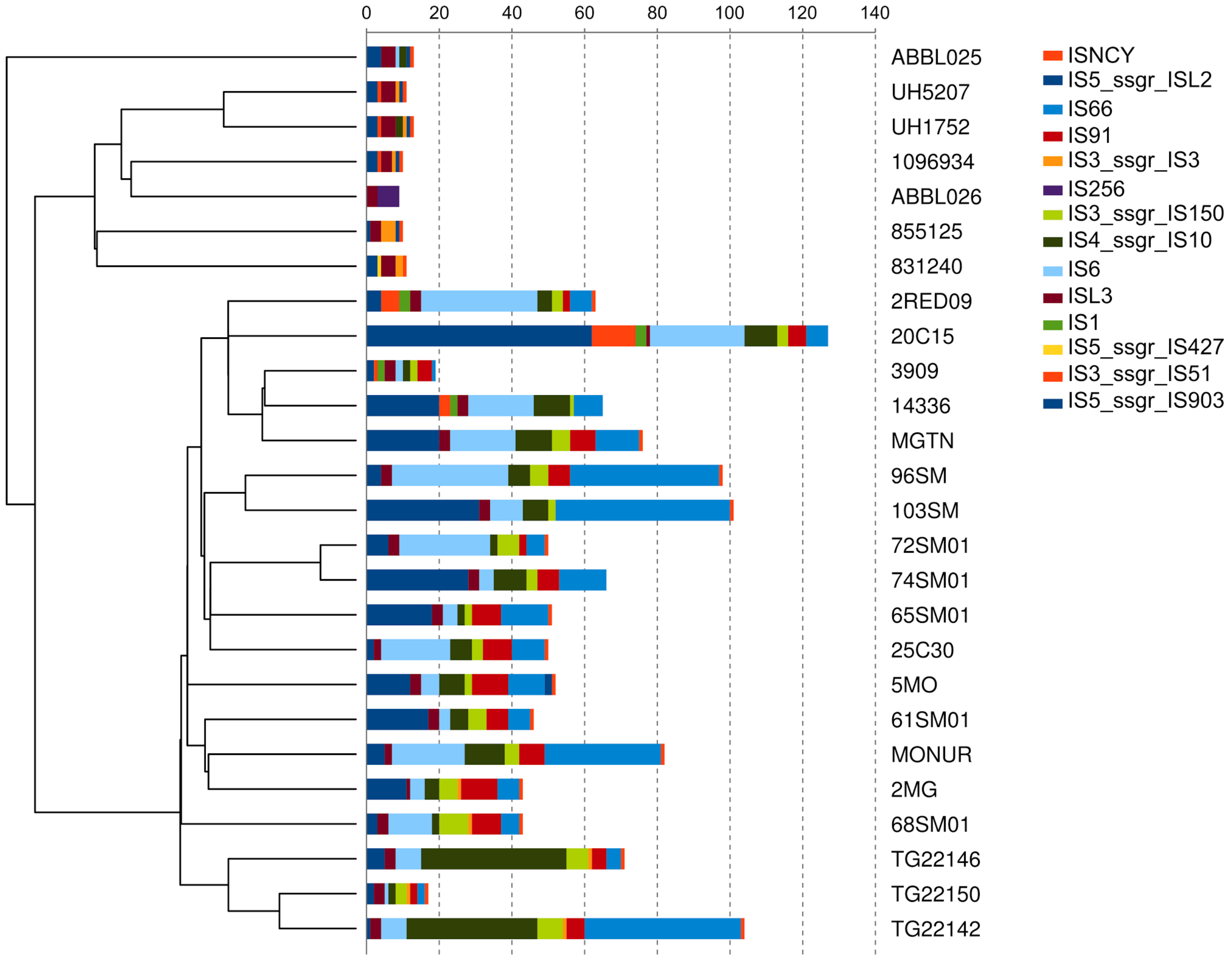


Figure 4. Histogram of the insertion sequences detected by ISSaga on the genomes of the Sequence Type 78. On the left, the phylogeny tree of the sequence type is reported

Plasmid analysis. Three of the four detected plasmids (pIBAC_oxa58_2RED of 25311 bp in strain 2RED09, pIBAC_oxa58_1433 of 26496 bp in strain 14336 and pIBAC_oxa58_20C15 of 26781 bp in strain 20C15) contained one single copy of gene *bla*$_{\text{OXA-58}}$. This result is congruent with the already published sequence of the plasmid of strain 3909 (24). Plasmid sequences (including the one of strain 3909) showed to have a good reciprocal synteny when analyzed with the software Mauve (25) (see Supplementary Figure 2), with some small internal reorganization. Conversely, the gene *bla*$_{\text{OXA-23}}$, detected in the strains 20C15 and 5MO, was found to be chromosomally encoded in both cases. The *bla*$_{\text{OXA-23}}$ site was recognised in both cases to have a 100% sequence similarity with the previously described transposon Tn2006 which was found to transfer the resistant determinant from *Acinetobacter radioresistens* to *A. baumannii* (26). Plasmid annotation analysis showed strong backbone similarity among the three sequences, and with the one found in strain 3909 as well. All four sequences presented two replication initiation sites *repAci1* and *repAci2* and two genes encoding for conjugal transfer proteins, *trbL* and *traA* (except for 3909 harboring only *traA),* followed by two IS*Aba25* insertion sequences. Some differences, however, were found around the *bla*$_{\text{OXA-58}}$ locus. Indeed, the plasmid of isolate 14336 (*pAB14336*) had opposite orientation of the IS*Aba2*/IS*Aba3*-*bla*$_{\text{OXA-58}}$-IS*Aba3* cluster when compared to the previously described plasmid *p183Eco* (27). Moreover, one IS26 sequence is missing. On the plasmid of strain 2RED09, instead, *bla*$_{\text{OXA-58}}$ is surrounded by two IS*Aba3* with opposite orientation (see Supplemental Figure 3). On the plasmid of strain 20C15, *bla*$_{\text{OXA-58}}$ is surrounded by IS*Aba3* and IS*Aba2*. Lastly, on the plasmid of 3909, *bla*$_{\text{OXA-58}}$ is surrounded by IS*Aba2* and IS*Aba3* with different orientation for *bla*$_{\text{OXA-58}}$ when compared to p14336.

Two plasmid replication sites were found in the assembly of strain 5MO. One was the replication site of plasmid pAB5MO (published in the present work). The second was found on a contig possibly integrated in the chromosome. Indeed, the read coverage on the contig carrying the repAci (*aci6*) was similar to the average of the chromosomic contigs. The presence of *aci6,* even though it does not directly correspond to the resistant gene, suggests that it could be responsible for the transfer of plasmid carrying the carbapenemase gene, as previously described by Bertini and coworkers (28). Indeed, the strain 5MO was found to encode a copy of gene *bla*$_{\text{OXA-23}}$ integrated in its chromosome.

DISCUSSION

The genomes of 15 strains of *A. baumannii* with SMAL pulsotype were sequenced and compared with the genomic variability of the species as a whole, and of the ST they were found to belong to, ST78. A phylogenomic approach focused on 26 ST78 genomes led to the identification of an evolutionary monophyletic group of 16 Italian genomes (previously assigned to the SMAL pulsotype. Three other genomes isolated in Arizona (USA) formed a closely related clade, while the remaining seven strains resulted more divergent. These data lead to conclude that the SMAL clone was imported in Italy in one single event.

Four of the 15 sequenced genomes were found to carry plasmids. The complete sequence of the three plasmids carrying the gene *bla*$_{\text{OXA-58}}$ was compared with the existing ones. Limited differences in the global structures of the plasmids harboured by the Italian ST78 strains were detected. The site including the resistance gene, on the other hand, showed high variability, especially concerning the IS. These variations surrounding the *bla*$_{\text{OXA-58}}$ locus suggest a lack of stability of this site. Such instability, together with the high energetic burden of maintaining a plasmid, could lead to a possible loss of carbapenem resistance. These observations could indicate an ongoing switch, in epidemiological terms, from the plasmid encoded *bla*$_{\text{OXA-58}}$ to the more stable and chromosomally mediated *bla*$_{\text{OXA-23}}$. Other cases of replacement between the two determinants have been reported for other Clones both in Italy and elsewhere (29, 30).

The 16 Italian strains present three interesting genome features: i) limited or absent recombination signal (Figure 2), ii) highly conserved gene content (Figure 3), and iii) a strong proliferation of multiple classes of IS elements, including class 66 (Figure 4). These three characteristics seems to be in contrast, as the first two features suggest genome stability, while IS proliferation is considered a trademark of genomic plasticity. Here we propose a genome evolution scenario that starts with the proliferation of ISs, including IS66 elements. One IS66 then inactivated the gene *comEC/Rec2*, an event clearly shown by our genome data to have occurred just once, at the basis of the Italian clade. ComEC is the inner membrane protein responsible for the intake of DNA in the cytoplasm (31, 32). In our scenario, the interruption of this gene reduced the capability of DNA exchange of the Italian strains. In parallel, IS elements proliferation played other roles in the evolution of the clade, such as affecting the stability of the blaOXA-58 gene and causing the loss of function of other genes, which could have in turn contributed to the current low genomic plasticity of the entire ST78A clade, or to branches of it.

These key elements depict a scenario in which the force driving the evolution of the SMAL clone is selfish DNA, in the form of IS elements. IS are known to be, usually, anchors for homologous recombination processes and are thus considered carriers of genomic plasticity and responsible for the evolution of virulent clones in multiple bacterial species. In the case described in the present work, the SMAL clone possesses a high number of IS elements but surprisingly there is no evidence of homologous recombination. On the contrary, the interruption of the *comEC/rec2* gene by IS66 may contribute to the reduction of import of exogenous DNA. Thus, in this case, recombination events are not coupled with IS proliferation but seem to be in a competitive relationship.

MATERIALS AND METHODS

Pulsed Field Gel Electrophoresis. PFGE of *A. baumannii* was performed after *Apa*I digestion using a method described previously (33). Genomic DNA was prepared in agarose plugs, and DNA restriction was carried out at 30°C for 16 h. PFGE was performed in a CHEF DRII system (Bio-Rad, Hercules, CA, USA), with pulses ranging from 0.5 to 15 s at a voltage of 6 V/cm at 14°C for 20 h. Lambda 48.5-kb concatemers (New England BioLabs, Beverly, MA, USA) were used as molecular size markers. Isolates showing three or fewer band differences were regarded as a single PFGE type, according to the criteria described previously by Tenover *et al.* (34).

Biofilm formation capability assay. One millilitre of fresh medium in borosilicate (15×125 mm), polystyrene (12×75 mm) or polypropylene (12×75 mm) sterile tubes was inoculated with 0.01 ml of an overnight culture. Triplicate cultures for each sample were incubated for 8 h shaking (at 200 r.p.m. in an orbital shaker) at 37°C. The supernatant of the tube was aspirated and rinsed thoroughly with distilled water. The cells attached to the tube walls were visualized and quantified by staining with crystal violet and solubilization with ethanol–acetone as described by Thomas and coworkers (35). The OD_{600} was detected using a spectrophotometer and compared to that of 2MG and 65SM01 (i.e. known biofilm forming SMAL strains, already included in the work by Nucleo and coworkers (19)).

Identification and Antibiotic resistance profiling. Identification and susceptibility profiles were initially established using MicroScan4 (Beckman Coulter) NBC46 panels. MICs of imipenem (IPM) and meropenem (MER) (carbapenem resistance) were obtained by Etest strips (bioMérieux). Results were interpreted according to the latest recommendations EUCAST guidelines (http://www.eucast.org/clinical%20breakpoints/).

DNA extraction, sequencing and assembly. Bacterial strains were cultivated in MacConkey medium in petri dishes. One single colony per strain was used for the downstream genomic analyses treated in this work. DNA was extracted using NucleoSpin Tissue (Macherey-Nagel) kit, library were prepared using Nextera XT kits and sequenced with the Illumina MiSeq technology with 2x250 paired-end runs. Reads were assembled with the Mira 4.0 assembler (36) using the default settings for Illumina reads and excluding the control for high coverage.

Global database of *Acinetobacter baumannii* genomes. All available *A. baumannii* genomes (April 2016) were downloaded from the NCBI ftp site. All genomes were merged to those sequenced in this work to form the global database of sequenced strains of this species. The Multilocus Sequence Type of all genomes was determined using an in-house script and the Pasteur profiling scheme (37). Genomic sequences were aligned to each other using an in-house Perl script and the Mauve software (25). Small Nucleotide Polymorphisms (SNPs) were extracted from regions where all genomes aligned to the others. SNPs were used to investigate the evolution of the species. The software fasttree 2.1.7 SS3 (38) was used to build a maximum likelihood phylogeny using the alignment of single nucleotide variants as input.

Fine phylogeny of the SMAL and closely related strains. 26 genomes were aligned to the evolutionary closest available complete genome (i.e. AB031, according to the global phylogeny). Each global genomic alignment was performed using the software Mauve and a set of in-house Perl and Python scripts for output formatting. A global alignment of the 26 genomes of interest was obtained and used to extract SNPs, which are used for phylogeny. The evolutionary analysis was performed using the software RAxML (39) with the ASC_GTRGAMMA evolution model and 100 bootstrap replicates, using the ascertainment bias correction of Lewis.

Recombination analysis. The presence of recombination in the dataset was tested by plotting the concentration of core SNPs along the genomes. Assemblies were aligned to the complete genomic sequence of strain AB031 as for the phylogeny of ST78 (see paragraph for details). Concentration of core SNP was calculated in windows of 1000bp along the whole genome (using the genome of strain ABBL025 as a reference) and plotted using the R software.

Pan/core-genome analysis. All 26 genomes of ST78 were annotated with the automatic pipeline Prokka (40) using the default settings for bacteria and avoiding to call rRNA sequences. Gene presence and absence was calculated using PanOCT (41), using the Prokka annotation and a reciprocal blastp analysis as input and adopting the same parameters as the work by Chan and colleagues (42). Reciprocal blast was performed using blast+ program instead of blastall in order to minimize the time of calculations). Pan-genome and core-genome of the whole dataset and of single genomic clusters were calculated using PanOCT and R. An analysis of binary distance using gene presence as characters was carried out using R. Dispersion in gene content for each clade of interest was calculated as follows: (pan-genome of the clade – core-genome of the clade) / number of organisms in the clade. The evolutionary distance between two genomes was expressed as number of non-homoplasic core SNPs between the two genomes. Non-homoplasic sites were obtained using Noisy (43) on the core SNP alignment previously used for the phylogeny of ST78. The evolutionary distance was plotted against the binary distance of gene content for each couple of genomes inside the clusters ST78A and ST78B.

Gene content analyses. The presence of genes coding for antimicrobial resistance and competence factors was tested using blast with ad hoc prepared set of genes (see Supplementary Table 2). Virulence gene database was obtained (44) and tested with blast. More research was performed on the whole ResFinder and VirulenceFinder databases (https://cge.cbs.dtu.dk/services/data.php), using a permissive blast search and checking positive results manually. O-antigen genes were extracted from the reference genomes of the strains ABNIH1 (Biosample SAMN00855421), ABNIH2 (Biosample SAMN00857848), and ABNIH3 (Biosample SAMN00857859).

***In-silico* Plasmid extraction and characterization.** Assembled genomes and contigs were blasted against in-house generated database of plasmid replication sites of *A. baumannii*, while resistant genes were determined by uploading the contigs to ResFinder database (www.cge.cbs.dtu.dk/services/ResFinder, (45)). Genomes positive to the plasmid replication site search, were reassembled using SPAdes (46). The contigs containing plasmid sequences were detected, analyzed and closed using the software Bandage (47). ORFs and their relative amino acids were predicted using Artemis (48). The annotation was performed manually using the online blast tool on the nr database. Genbank files were formatted and uploaded using the Sequin tool.

BIBLIOGRAPHY

1. Rice LB. 2008. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. J Infect Dis 197:1079–1081.

2. Peleg AY, Seifert H, Paterson DL. 2008. Acinetobacter baumannii: Emergence of a Successful Pathogen. Clin Microbiol Rev 21:538–582.

3. Imperi F, Antunes LCS, Blom J, Villa L, Iacono M, Visca P, Carattoli A. 2011. The genomics of Acinetobacter baumannii: insights into genome plasticity, antimicrobial resistance and pathogenicity. IUBMB Life 63:1068–1074.

4. Howard A, O'Donoghue M, Feeney A, Sleator RD. 2012. Acinetobacter baumannii: an emerging opportunistic pathogen. Virulence 3:243–250.

5. Snitkin ES, Zelazny AM, Montero CI, Stock F, Mijares L, NISC Comparative Sequence Program, Murray PR, Segre JA. 2011. Genome-wide recombination drives diversification of epidemic strains of Acinetobacter baumannii. Proc Natl Acad Sci U S A 108:13758–13763.

6. Feng Y, Ruan Z, Shu J, Chen C-L, Chiu C-H. 2016. A glimpse into evolution and dissemination of multidrug-resistant Acinetobacter baumannii isolates in East Asia: a comparative genomics study. Sci Rep 6:24342.

7. Li H, Liu F, Zhang Y, Wang X, Zhao C, Chen H, Zhang F, Zhu B, Hu Y, Wang H. 2015. Evolution of carbapenem-resistant Acinetobacter baumannii revealed through whole-genome sequencing and comparative genomic analysis. Antimicrob Agents Chemother 59:1168–1176.

8. Beare PA, Unsworth N, Andoh M, Voth DE, Omsland A, Gilk SD, Williams KP, Sobral BW, Kupko JJ 3rd, Porcella SF, Samuel JE, Heinzen RA. 2009. Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus Coxiella. Infect Immun 77:642–656.

9. Rohmer L, Fong C, Abmayr S, Wasnick M, Larson Freeman TJ, Radey M, Guina T, Svensson K, Hayden HS, Jacobs M, Gallagher LA, Manoil C, Ernst RK, Drees B, Buckley D, Haugen E, Bovee D, Zhou Y, Chang J, Levy R, Lim R, Gillett W, Guenthener D, Kang A, Shaffer SA, Taylor G, Chen J, Gallis B, D'Argenio DA,

Forsman M, Olson MV, Goodlett DR, Kaul R, Miller SI, Brittnacher MJ. 2007. Comparison of Francisella tularensis genomes reveals evolutionary events associated with the emergence of human pathogenic strains. Genome Biol 8:R102.

10. Bouchami O, de Lencastre H, Miragaia M. 2016. Impact of Insertion Sequences and Recombination on the Population Structure of Staphylococcus haemolyticus. PLoS One 11:e0156653.

11. Leavis HL, Willems RJL, van Wamel WJB, Schuren FH, Caspers MPM, Bonten MJM. 2007. Insertion Sequence–Driven Diversification Creates a Globally Dispersed Emerging Multiresistant Subspecies of E. faecium. PLoS Pathog 3:e7.

12. Croucher NJ, Klugman KP. 2014. The emergence of bacterial “hopeful monsters.” MBio 5:e01550–14.

13. Turton JF, Ward ME, Woodford N, Kaufmann ME, Pike R, Livermore DM, Pitt TL. 2006. The role of ISAba1 in expression of OXA carbapenemase genes in Acinetobacter baumannii. FEMS Microbiol Lett 258:72–77.

14. Turton JF, Kaufmann ME, Glover J, Coelho JM, Warner M, Pike R, Pitt TL. 2005. Detection and typing of integrons in epidemic strains of Acinetobacter baumannii found in the United Kingdom. J Clin Microbiol 43:3074–3082.

15. Principe L, Piazza A, Giani T, Bracco S, Caltagirone MS, Arena F, Nucleo E, Tammaro F, Rossolini GM, Pagani L, Luzzaro F, AMCLI-CRAb Survey Participants. 2014. Epidemic diffusion of OXA-23-producing Acinetobacter baumannii isolates in Italy: results of the first cross-sectional countrywide survey. J Clin Microbiol 52:3004–3010.

16. Migliavacca R, Espinal P, Principe L, Drago M, Fugazza G, Roca I, Nucleo E, Bracco S, Vila J, Pagani L, Luzzaro F. 2013. Characterization of resistance mechanisms and genetic relatedness of carbapenem-resistant Acinetobacter baumannii isolated from blood, Italy. Diagn Microbiol Infect Dis 75:180–186.

17. Wright MS, Haft DH, Harkins DM, Perez F, Hujer KM, Bajaksouzian S, Benard MF, Jacobs MR, Bonomo RA, Adams MD. 2014. New insights into dissemination and variation of the health care-associated pathogen Acinetobacter baumannii from genomic analysis. MBio 5:e00963–13.

18. Holt K, Dougan G, Kenyon JJ, Schultz MB, Hamidian M, Pickard DJ, Hall R. 2016. Five decades of genome evolution in the globally distributed, extensively antibiotic-resistant Acinetobacter baumannii global clone 1. Microbial Genomics 2.

19. Nucleo E, Steffanoni L, Fugazza G, Migliavacca R, Giacobone E, Navarra A, Pagani L, Landini P. 2009. Growth in glucose-based medium and exposure to subinhibitory concentrations of imipenem induce biofilm formation in a multidrug-resistant clinical isolate of Acinetobacter baumannii. BMC Microbiol 9:270.

20. Giannouli M, Tomasone F, Agodi A, Vahaboglu H, Daoud Z, Triassi M, Tsakris A, Zarrilli R. 2009. Molecular epidemiology of carbapenem-resistant Acinetobacter baumannii strains in intensive care units of multiple Mediterranean hospitals. J Antimicrob Chemother 63:828–830.

21. Giannouli M, Cuccurullo S, Crivaro V, Di Popolo A, Bernardo M, Tomasone F, Amato G, Brisse S,

Triassi M, Utili R, Zarrilli R. 2010. Molecular Epidemiology of Multidrug-Resistant Acinetobacter baumannii in a Tertiary Care Hospital in Naples, Italy, Shows the Emergence of a Novel Epidemic Clone. J Clin Microbiol 48:1223–1230.

22. Carretto E, Barbarini D, Dijkshoorn L, van der Reijden TJK, Brisse S, Passet V, Farina C, APSI Acinetobacter Study Group. 2011. Widespread carbapenem resistant Acinetobacter baumannii clones in Italian hospitals revealed by a multicenter study. Infect Genet Evol 11:1319–1326.

23. Cherkaoui A, Emonet S, Renzi G, Schrenzel J. 2015. Characteristics of multidrug-resistant Acinetobacter baumannii strains isolated in Geneva during colonization or infection. Ann Clin Microbiol Antimicrob 14:42.

24. Zarrilli R, Giannouli M, Rocco F, Loman NJ, Haines AS, Constantinidou C, Pallen MJ, Triassi M, Di Nocera PP. 2011. Genome sequences of three Acinetobacter baumannii strains assigned to the multilocus sequence typing genotypes ST2, ST25, and ST78. J Bacteriol 193:2359–2360.

25. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403.

26. Poirel L, Figueiredo S, Cattoir V, Carattoli A, Nordmann P. 2008. Acinetobacter radioresistens as a silent source of carbapenem resistance for Acinetobacter spp. Antimicrob Agents Chemother 52:1252–1256.

27. Bertini A, Poirel L, Bernabeu S, Fortini D, Villa L, Nordmann P, Carattoli A. 2007. Multicopy blaOXA-58 gene as a source of high-level resistance to carbapenems in Acinetobacter baumannii. Antimicrob Agents Chemother 51:2324–2328.

28. Bertini A, Poirel L, Mugnier PD, Villa L, Nordmann P, Carattoli A. 2010. Characterization and PCR-based replicon typing of resistance plasmids in Acinetobacter baumannii. Antimicrob Agents Chemother 54:4168–4177.

29. Wu W, He Y, Lu J, Lu Y, Wu J, Liu Y. 2015. Transition of blaOXA-58-like to blaOXA-23-like in Acinetobacter baumannii Clinical Isolates in Southern China: An 8-Year Study. PLoS One 10:e0137174.

30. D'Arezzo S, Principe L, Capone A, Petrosillo N, Petrucca A, Visca P. 2010. Changing carbapenemase gene pattern in an epidemic multidrug-resistant Acinetobacter baumannii lineage causing multiple outbreaks in central Italy. J Antimicrob Chemother 66:54–61.

31. Krüger N-J, Stingl K. 2011. Two steps away from novelty--principles of bacterial DNA uptake. Mol Microbiol 80:860–867.

32. Wilharm G, Piesker J, Laue M, Skiebe E. 2013. DNA uptake by the nosocomial pathogen Acinetobacter baumannii occurs during movement along wet surfaces. J Bacteriol 195:4146–4153.

33. Bou G, Cerveró G, Domínguez MA, Quereda C, Martínez-Beltrán J. 2000. PCR-based DNA fingerprinting (REP-PCR, AP-PCR) and pulsed-field gel electrophoresis characterization of a nosocomial outbreak caused by imipenem- and meropenem-resistant Acinetobacter baumannii. Clin Microbiol Infect 6:635–643.

34. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239.

35. Tomaras AP, Dorsey CW, Edelmann RE, Actis LA. 2003. Attachment to and biofilm formation on abiotic surfaces by Acinetobacter baumannii: involvement of a novel chaperone-usher pili assembly system. Microbiology 149:3473–3484.

36. Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information, p. 45–56. *In* German Conference on Bioinformatics.

37. Diancourt L, Passet V, Nemec A, Dijkshoorn L, Brisse S. 2010. The population structure of Acinetobacter baumannii: expanding multiresistant clones from an ancestral susceptible genetic pool. PLoS One 5:e10034.

38. Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650.

39. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

40. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

41. Fouts DE, Brinkac L, Beck E, Inman J, Sutton G. 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40:e172.

42. Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y, Huang X-Z, Beck E, Harkins DM, Kim M, Lesho EP, Nikolich MP, Fouts DE. 2015. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of Acinetobacter baumannii. Genome Biol 16:143.

43. Dress AWM, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF. 2008. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol 3:7.

44. Sahl JW, Del Franco M, Pournaras S, Colman RE, Karah N, Dijkshoorn L, Zarrilli R. 2015. Phylogenetic and genomic diversity in isolates from the globally distributed Acinetobacter baumannii ST25 lineage. Sci Rep 5:15188.

45. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644.

46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

47. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352.

48. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

ARTICLE 3

Tracking nosocomial *Klebsiella pneumoniae* infections and outbreaks by whole-genome analysis: small-scale Italian scenario within a single hospital

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

2015. 53:2861–2868.
doi:10.1128/JCM.00545-15.

Tracking Nosocomial *Klebsiella pneumoniae* Infections and Outbreaks by Whole-Genome Analysis: Small-Scale Italian Scenario within a Single Hospital

Raffaella Onori[a], Stefano Gaiarsa[b,c], Francesco Comandatore[c], Stefano Pongolini[d], Sylvain Brisse[e], Alberto Colombo[a], Gianluca Cassani[a], Piero Marone[b], Paolo Grossi[a], Giulio Minoja[a], Claudio Bandi[c], Davide Sassera[f] and Antonio Toniolo[a]

Author Affiliations

[a] University of Insubria and Ospedale di Circolo e Fondazione Macchi, Varese, Italy
[b] Fondazione IRCCS Policlinico S. Matteo, Pavia, Italy
[c] Università degli Studi di Milano, Milan, Italy
[d] Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Parma, Italy
[e] Institut Pasteur and CNRS, UMR 3525, Paris, France
[f] Università degli Studi di Pavia, Pavia, Italy

D. J. Diekema, Editor | **R.O. and S.G. contributed equally to this article.**
Received 26 February 2015. Returned for modification 25 March 2015. Accepted 29 May 2015. Accepted manuscript posted online 1 July 2015.

Address correspondence to Davide Sassera, **davide.sassera@unipv.it**.

Supplemental material is available at the end of the thesis

This article can be found at **http://dx.doi.org/10.1128/JCM.00545-15**.
Please scan this QR code to access the website from the printed version

ACKNOWLEDGMENTS. We thank Alexis Criscuolo (Institut Pasteur) for support in cgMLST analysis. We acknowledge the generous support of Regione Lombardia (Milan, Italy) (Nosocomial Infections grant to A.T.). We declare that we have no conflicts of interest to report.

ABSTRACT

Multidrug-resistant (MDR) *Klebsiella pneumoniae* is one of the most important causes of nosocomial infections worldwide. After the spread of strains resistant to beta-lactams at the end of the previous century, the diffusion of isolates resistant to carbapenems and colistin is now reducing treatment options and the containment of infections. Carbapenem-resistant *K. pneumoniae* strains have spread rapidly among Italian hospitals, with four subclades of pandemic clonal group 258 (CG258). Here we show that a single Italian hospital has been invaded by three of these subclades within 27 months, thus replicating on a small scale the "Italian scenario." We identified a single clone responsible for an epidemic outbreak involving seven patients, and we reconstructed its star-like pattern of diffusion within the intensive care unit. This epidemiological picture was obtained through phylogenomic analysis of 16 carbapenem-resistant *K. pneumoniae* isolates collected in the hospital during a 27-month period, which were added to a database of 319 genomes representing the available global diversity of *K. pneumoniae* strains. Phenotypic and molecular assays did not reveal virulence or resistance determinants specific for the outbreak isolates. Other factors, rather than selective advantages, might have caused the outbreak. Finally, analyses allowed us to identify a major subclade of CG258 composed of strains bearing the yersiniabactin virulence factor. Our work demonstrates how the use of combined phenotypic, molecular, and whole-genome sequencing techniques can help to identify quickly and to characterize accurately the spread of MDR pathogens.

INTRODUCTION

Klebsiella pneumoniae is a major nosocomial pathogen that is rapidly spreading in hospitals worldwide, mainly due to the common occurrence of multidrug-resistant (MDR) strains (1). Infections caused by this pathogen are difficult to eradicate, since *K. pneumoniae* carries genes for resistance to the majority of antimicrobial drugs, including carbapenems (2, 3). The first strain of carbapenem-resistant *K. pneumoniae* was isolated in 1996; the plasmid-encoded determinant was named *K. pneumoniae* carbapenemase (KPC) and was indicated as the bla_{KPC} gene (4). Since then, KPC-producing *K. pneumoniae* strains have been spreading worldwide. Additional carbapenemases (bla_{NDM}, $bla_{\mathrm{OXA\text{-}48}}$, bla_{VIM}, and $bla_{\mathrm{IMP\text{-}1}}$) have now been reported for MDR *Enterobacteriaceae*, including *K. pneumoniae* (5–8). A last-resort treatment for infections caused by MDR Gram-negative bacteria is represented by membrane-acting polymyxins such as colistin, but resistance to this antibiotic in *K. pneumoniae* is also emerging (9, 10).

In addition to the study of genes providing resistance to antibiotics, genetic factors involved in the variable levels of virulence of different isolates of *K. pneumoniae* are currently highly investigated but only partially understood. Among the most important virulence factors are fimbrial genes (*mrk* and *fim* operons), which mediate adherence to surfaces and host tissues (11, 12). Another important aspect involved in the colonization of the host is the presence of genes for iron uptake systems such as aerobactin (13), enterobactin (*ent* operon) (14), and yersiniabactin (*irp* and *ybt* genes) (15). Capsular types, particularly K1 and K2, and hypermucoviscosity, favored by the positive regulator genes *rmpA* and *rmpA*2, are also important for *K. pneumoniae* virulence. Capsule production increases resistance to phagocytosis and other immune response components (16, 17). For detailed descriptions of these and other potential virulence factors of *K. pneumoniae*, see references 18 and 19.

Most of the KPC-producing *K. pneumoniae* strains isolated worldwide have been attributed to clonal group 258 (CG258) (19, 20). Recent phylogenomic analyses showed that four different subclades of pandemic CG258 are present in Italy, indicating entrance into the country on at least four different occasions during the period of 2008 to 2010 (21). The spread of MDR *K. pneumoniae* in hospitals and nursing homes in Italy is known to have occurred very rapidly, with a diffusion pattern that has been described as the “Italian scenario” (22). The worldwide spread of *K. pneumoniae* is due, in part, to failures in the early identification of MDR strains, as well as high rates of recombination and horizontal gene transfer (21, 23, 24).

Whole-genome sequencing is now offering the possibility of in-depth characterization of bacterial isolates, and it holds the potential to reconstruct the origin and diffusion of nosocomial infections and outbreaks (19, 25). Here we present a phylogenomic study of 16 isolates from a single hospital in northwestern Italy that were collected between 2011 and 2013, including an epidemic outbreak in 2013 that involved seven patients. Genomes from these isolates were compared with 319 publicly available genomes, representing the available global genomic diversity of *K. pneumoniae*. Phylogenomic analysis, together with phenotyping assays and molecular characterization of drug resistance determinants and virulence genes, allowed us to trace the origins of sporadic infections and the outbreak, to describe the monophyletic origin of a yersiniabactin-positive subclade of CG258, and to detect a common genetic trait in colistin-resistant strains.

MATERIALS AND METHODS

Nosocomial infections with *K. pneumoniae* and hospital outbreak. Between January 2011 and March 2013, 16 cases of infection due to carbapenem-resistant *K. pneumoniae* occurred at the Ospedale di Circolo e Fondazione Macchi (Varese, Italy). Seven cases that occurred in the intensive care unit (ICU) during a short period were part of a single epidemic event that started in February 2013 (Fig. 1). Evidence indicated that a 69-year-old man was patient zero (indicated as KpVA-8 in Fig. 1). He had been transferred to the ICU from a nearby hospital, with an already diagnosed infection due to KPC-producing *K. pneumoniae*. During his stay in the ICU, infection spread to six other patients.

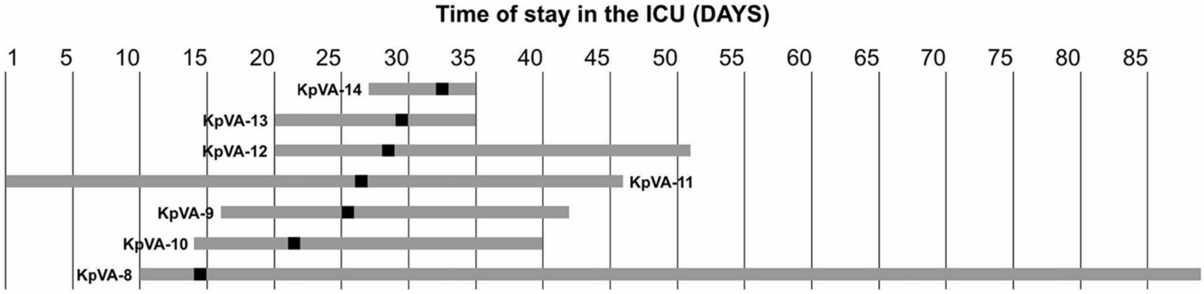


FIG 1. Time frames of stays in the ICU for the seven patients involved in the *K. pneumoniae* outbreak. Horizontal bars, length of stay for each patient. Black squares, day of the first isolation of *K. pneumoniae* for each patient.

Bacterial isolates. A total of 16 non-duplicated isolates of *K. pneumoniae* were investigated, specifically, the first isolate obtained from each patient. Multiple *K. pneumoniae* isolates were obtained subsequently from each patient, for clinical reasons (e.g., spread of infection to novel body sites) or in the course of surveillance studies. Clinical specimens included urine, blood, bronchoalveolar lavage fluid, sputum, tracheal aspirate, and wound specimens. During the outbreak period, ICU patients were screened every 3 days for surveillance, using nasal, armpit, inguinal, and rectal swabs. Species identification and antibiotic susceptibility tests were performed with the FDA-approved Phoenix automated microbiology system (Becton, Dickinson, Sparks, MD). Additional quantitative assays were performed using Etest strips (bioMérieux, Marcy l'Etoile, France) on Mueller-Hinton agar II plates (Becton, Dickinson), according to clinical breakpoints from the European Committee on Antimicrobial Susceptibility Testing (EUCAST). Short descriptions of the investigated isolates are presented in Table 1, while Etest MICs are reported in Table 2.

Patient no.	Clinical isolate status	Date of isolation (mo/day/yr)	Source*	Sequence type
KpVA-4	Sporadic	1/11/11	B	ST258
KpVA-5	Sporadic	1/28/11	B	ST258
KpVA-6	Sporadic	3/14/11	B	ST258
KpVA-7	Sporadic	5/3/11	B	ST258
KpVA-1	Sporadic	10/31/12	SP	ST512
KpVA-2	Sporadic	1/30/13	BAL	ST258
KpVA-8	Epidemic	2/1/13	BAL	ST512
KpA-10	Epidemic	2/8/13	BAL	ST512
KpVA-9	Epidemic	2/12/13	BAL	ST512
KpVA-11	Epidemic	2/13/13	B	ST512
KpVA-12	Epidemic	2/15/13	TA	ST512
KpVA-13	Epidemic	2/16/13	WS	ST512
KpVA-14	Epidemic	2/19/13	BAL	ST512
KpVA-3	Sporadic	3/14/13	B	ST258
KpVA-15	Sporadic	3/24/13	U	ST512
KpVA-16	Sporadic	3/24/13	U	ST512

TABLE 1. Clinical isolates and main properties. * B, blood; WS, wound sample; SP, sputum; BAL, bronchoalveolar lavage fluid; TA, tracheal aspirate; U, urine.

Characteristic	KPVA-4	KPVA-5	KPVA-6	KPVA-7	KPVA-1	KPVA-2	KPVA-8	KPVA-10	KPVA-9	KPVA-11	KPVA-12	KPVA-13	KPVA-14	KPVA-3	KPVA-15	KPVA-16
Date of isolation (mo/day/yr)	1/11/11	1/28/11	3/14/11	5/3/11	10/31/12	1/30/13	2/1/13	2/8/13	2/12/13	2/13/13	2/15/13	2/16/13	2/19/13	3/14/13	3/24/13	3/24/13
								Sensitivity and MIC (mg/liter)*								
Ampicillin	R, >16	R, >16	R, >16	R, >16	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8
Amoxicillin-clavulanate	R, >16/8	R, >16/8	R, >16/8	R, >16/8	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2	R, >8/2
Ceftazidime	R, >16	R, >16	R, >16	R, >16	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8	R, >8
Cefotaxime	R, >32	R, >32	R, >32	R, >32	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4	R, >4
Aztreonam	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16	R, >16
Ertapenem	R, >1	R, >32	R, >1	R, >1	R, >1	R, >1	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, >1	R, >1	R, >1
Imipenem	R, >32	R, 8	R, 32	R, 8	R, >8	R, >8	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, 8	R, >8	R, >8
Meropenem	R, >32	R, 32	R, >32	I, 6	R, >8	R, >8	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, >32	R, 12	R, >8	R, >8
Ciprofloxacin	R, >2	R, >2	R, >2	R, >2	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1	R, >1
Levofloxacin	R, >4	R, >4	R, >4	R, >4	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2	R, >2
Amikacin	R, 48	R, 64	R, 48	R, 64	R, >16	R, >16	S, 1	S, 1	S, 1	S, 1	S, 1	S, 1	S, 1	R, 48	R, >16	R, >16
Gentamicin	S, 0.5	S, 1.5	S, 1	S, 2	S, 2	S, 2	S, 0.25	S, 0.25	S, 0.25	S, 0.25	S, 0.25	S, 0.25	S, 0.25	S, 1.5	S, 4	S, 2
Tobramycin	S, 0.5	R, 16	R, 16	R, 24	R, >4	R, >4	S, 0.38	S, 0.38	S, 0.38	S, 0.38	S, 0.38	S, 0.38	S, 0.38	R, 12	R, >4	R, >4
Colistin	S, 0.12	S, 0.19	S, 0.38	S, 0.19	R, >4	S <1	R, 8	R, 8	R, 8	R, 8	R, 8	R, 8	R, 8	S, 0.19	R, >4	R, >4

TABLE 2. Antimicrobial susceptibility profiles of 16 investigated *K. pneumoniae* isolates. * R, resistant; I, intermediate; S, susceptible.

Direct sequencing of 16S rRNA, drug resistance genes, and virulence factors. Starting from pure cultures on Mueller-Hinton agar, bacterial DNA was obtained by lysozyme pretreatment (Sigma-Aldrich, Milan, Italy) followed by extraction with a QIAmp DNA Blood minikit (Qiagen, Milan, Italy). Confirmatory identification was performed via direct sequencing of the 16S rRNA gene. PCR was performed using AmpliTaq Gold with buffer I (Applied Biosystems, Life Technologies, Monza, Italy) in 50-μl mixtures, according to the manufacturer's instructions. PCR primers were synthesized by Sigma-Genosys (Haverhill, United Kingdom). Published primers and thermal protocols were used (26). DNA fragments were analyzed by electrophoresis on a 1.5% agarose gel in TBE buffer (89 mM Tris-borate and 2 mM EDTA [pH 8.3]) containing GelRed (10,000× in water; Biotium, DBA Italy, Segrate, Italy). PCR products were purified and sequenced on an ABI Prism 310 sequencer (Life Technologies). Sequences were compared with those in GenBank.

PCR assays for detecting antimicrobial resistance genes (27) and virulence factors were performed according to published protocols. Genes coding for adhesion fimbriae, enterobactin, and yersiniabactin siderophores were searched for, as follows: *fimH* gene, coding for type 1 fimbriae (28); *mrkA* gene, coding for the major subunit protein, and *mrkD* gene, coding for the adhesin, for type 3 fimbriae (29, 30); e*ntE* gene, coding for synthase subunit E, and e*ntB* gene, coding for isochorismatase, for enterobactin siderophore synthesis (14); y*btS* gene, coding for salicylate synthase, for yersiniabactin siderophore synthesis; *irp-1* and *irp-2* genes, related to yersiniabactin siderophore (15). Direct sequencing was performed as reported above.

Whole-genome sequencing and assembly. Whole-genome DNA was sequenced using an Illumina MiSeq platform (Illumina Inc., San Diego, CA), with a paired-end run of 2 by 250 bp, after Nextera XT paired-end library preparation. Sequencing reads were assembled using MIRA 4.0 software (31) with accurate *de novo* settings.

***In silico* MLST and gene mining.** Multilocus sequence typing (MLST) profiles were obtained *in silico* by analyzing appropriate gene variants (http://bigsdb.web.pasteur.fr/perl/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef_public&page=downloadAlleles) for each genome, using an in-house Python script. The presence of selected genes coding for antibiotic resistance and virulence factors was determined by using BLAST with a specifically designed database, BIGSdb-Kp (http://bigsdb.web.pasteur.fr/perl/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef_public&page=sequenceQuery) (19). All hits were manually checked, and genes requiring specificity for a particular variant (e.g., *bla*$_{\text{KPC}}$ versus *bla*$_{\text{OXA-48}}$) were requested to have 100% identity with the database sequences. BLAST searches and filters were also used to test for the presence of yersiniabactin genes in all genomes used for the global phylogenetic analysis (see Results for details). Analysis of the presence of insertion sequences within the *mgrB* gene (a putative determinant of colistin resistance) (32) was performed with a manually corrected BLAST search.

Core SNP detection and phylogeny. Whole-genome sequences of the 16 isolates were added to a previously described database of 319 genomes of *K. pneumoniae* strains isolated throughout the world (21). Single-nucleotide polymorphisms (SNPs) were detected using an in-house pipeline based on Mauve software (33), using the published NJST258_1 complete genome as a reference (21). Briefly, each genome was individually aligned with the reference and alignments were merged with Perl scripts to obtain a global alignment. Core SNPs, defined as single-nucleotide variations flanked by at least one identical nucleotide on both sides in all genomes analyzed (34), were detected. Maximum likelihood phylogenetic analysis was performed using core SNPs merged in a multialignment file. RAxML software was used (35) with the generalized time-reversible (GTR) model and 100 bootstraps.

Core genome MLST. Core genome MLST (cgMLST) analysis was performed using the BIGSdb software and database (19, 36). cgMLST profiles made of allelic variants at 694 loci were obtained for 219 genomes of CG258, representing the 16 genomes presented in this work. cgMLST profiles were used to produce a tree of all 219 genomes, using the unweighted pair group method with arithmetic mean (UPGMA) approach.

Outbreak reconstruction. The spreading routes of outbreak strains were reconstructed by combining core SNPs and the dates of sample collection, applying the SeqTrack method implemented in the R package Adegenet (37). The outbreak chain of transmission was then obtained using the R package Outbreaker (38).

Nucleotide sequence accession number. Genome assemblies were deposited in the EMBL database under accession number PRJEB7661.

RESULTS

Species identification and antimicrobial susceptibility. In this work, the first isolate obtained from each patient was investigated. Species identification was performed with biochemical and molecular assays. The seven isolates collected in February 2013 were suspected to belong to a single outbreak; these strains are referred to as epidemic. The remaining isolates are termed sporadic. Five of 16 isolates were obtained from blood cultures (3 of 7 for patients involved in the ICU outbreak).

Phenotypic assays detected resistance to imipenem, meropenem, and ertapenem in all isolates, regardless of the isolation date. All were thus classified as carbapenem resistant. The seven epidemic isolates were resistant to colistin but susceptible to aminoglycosides (gentamicin, amikacin, and tobramycin). Three of 9 sporadic isolates were also resistant to colistin.

Drug resistance determinants and virulence factors. Isolates were subjected to a set of PCR assays to detect drug resistance genes and virulence factors. The bla_{KPC} gene was detected in all isolates, while other carbapenem resistance genes (bla_{NDM}, $bla_{\mathrm{IMP\text{-}1}}$, $bla_{\mathrm{OXA\text{-}48}}$, and bla_{VIM}) were not detected. Genes coding for type 1 and type 3 fimbriae (*fim* and *mrk* operons, respectively) were detected in all isolates, as was the enterobactin siderophore located in the *ent* operon (Table 3). Three sporadic isolates carried genes for yersiniabactin, i.e., *ybtS* and the iron-repressible genes *irp1* and *irp2* (14, 39, 40). In *K. pneumoniae*, the yersiniabactin siderophore is expressed together with or instead of enterobactin. Finally, genomes were scanned for the presence of any beta-lactamase gene with the web tool BIGSdb (19). Genes coding for $\mathrm{Bla}_{\mathrm{SHV}}$ were detected in all isolates, while genes coding for $\mathrm{Bla}_{\mathrm{TEM}}$ were detected in 13 of the 16 genomes, being absent only in KpVa-2, KpVA-3, and KpVA-4. None of the analyzed genomes were found to carry genes of any of the other 17 beta-lactamase families.

Characteristic	KPVA-4	KPVA-5	KPVA-6	KPVA-7	KPVA-1	KPVA-2	KPVA-8	KPVA-10	KPVA-9	KPVA-11	KPVA-12	KPVA-13	KPVA-14	KPVA-3	KPVA-15	KPVA-16
Date of isolation (mo/day/yr)	1/11/11	1/28/11	3/14/11	5/3/11	10/31/12	1/30/13	2/1/13	2/8/13	2/12/13	2/13/13	2/15/13	2/16/13	2/19/13	3/14/13	3/24/13	3/24/13
								Presence and type of antibiotic resistance determinants								
blaKPC	2	3	3	3	3	2	3	3	3	3	3	3	3	2	3	3
blaVIM	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
blaNDM1	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
blaIMP	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
blaOXA	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
blaSHV	12	11	11	11	11	12	11	11	11	11	11	11	11	11	11	11
blaTEM	No	1	1	1	1	No	1	1	1	1	1	1	1	No	1	1
mgrB insertion	No	No	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
								Presence of virulence determinants								
fimACDEFH	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
mrkABCDF	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
rpmA	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
magA	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
entABCDEF	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ybtA	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	No	No
ybtS	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	No	No
irp1	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	No	No
irp2	Yes	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes	No	No

TABLE 3. Antimicrobial resistance and virulence determinants in 16 investigated *K. pneumoniae* isolates

Whole-genome sequencing and characterization. Whole-genome sequences were obtained for all 16 *K. pneumoniae* isolates. The assembled genomes were characterized *in silico* for MLST and were searched for genes coding for drug resistance determinants and virulence factors (Tables 1 and 3). MLST analysis enabled identification of two groups of isolates. Six isolates (KpVA-2, KpVA-3, KpVA-4, KpVA-5, KpVA-6, and KpVA-7) were of sequence type 258 (ST258); 10 isolates, including the epidemic ones, belonged to ST512. ST512 differs from ST258 at a single nucleotide, thus belonging to the same clonal group, CG258 (19, 20). Confirming the results obtained by molecular analysis, all strains were found to carry the *bla*$_{\mathrm{KPC}}$ gene. The *bla*$_{\mathrm{KPC2}}$ variant (KPC2) was found in three of six isolates belonging to ST258, while the remaining 13 isolates carried the *bla*$_{\mathrm{KPC3}}$ variant (KPC3).

The three isolates belonging to ST258 and possessing the *bla*$_{\mathrm{KPC2}}$ variant also presented unique profiles of virulence and drug resistance factors. Strains KpVA-2, KpVA-3, and KpVA-4 harbored the *irp1*, *irp2*, y*btA*, and *ybtS* genes, which were not detected in other strains. These four genes encode yersiniabactin, a virulence factor expressed by *Yersinia* and other enterobacteria, including *K. pneumoniae* (19, 41). All 16 strains analyzed here possessed the *mrk* and *fim* operons, coding for fimbrial genes (11, 12), and the *ent* operon, coding for enterobactin (14), consistent with previous results showing that these genes are highly conserved in *K. pneumoniae* (19, 30). None of the isolates demonstrated *rmpA* and *wzy*-K1 (*magA*) genes, which are hypermucoviscosity-associated genes (16, 17, 19).

Genes related to colistin resistance were also investigated. The entire set of *pmr* genes was highly conserved among the 16 strains, including the *pmrB* locus, which has been indicated as a colistin resistance determinant (42). All colistin-resistant strains harbored a variant of the *mgrB* gene interrupted by an IS*5*-like transposon (Table 3). Insertion of a transposon in this gene has been reported as a determinant of colistin resistance (32).

Global core SNP phylogeny. A global genome phylogeny of *K. pneumoniae*, including the 16 isolates investigated in this study, was obtained by adding the novel genomes to a previously constructed database of 319 isolates (21). Phylogeny was obtained in order to contextualize our strains among the previously sequenced *K. pneumoniae* isolates. The 16 novel genomes clustered in 5 monophyletic groups on the global tree (Fig. 2). As expected, they fit within CG258. Interestingly, these 16 isolates were assigned to three of the four previously identified groups of Italian isolates of CG258 (21). Inclusion of the investigated genomes in the global phylogeny allowed us to define the relationships among the isolates at the investigated hospital. KpVA-2, KpVA-3, and KpVA-4, the three isolates that were yersiniabactin positive, clustered together in a clade containing nine additional Italian strains and two U.S. strains, all demonstrating the yersiniabactin genes (which could also contribute to copper toxicity) (38). Thirty-nine additional isolates, belonging to different sequence types and scattered on the global *K. pneumoniae* phylogeny, also demonstrated yersiniabactin. The genomes of the seven isolates collected in February 2013 and hypothesized to represent a single epidemic event clustered together in a single, well-supported, phylogenetic clade (Fig. 2). This result confirmed the original hypothesis of a single clone being responsible for the seven infections that occurred in the ICU.

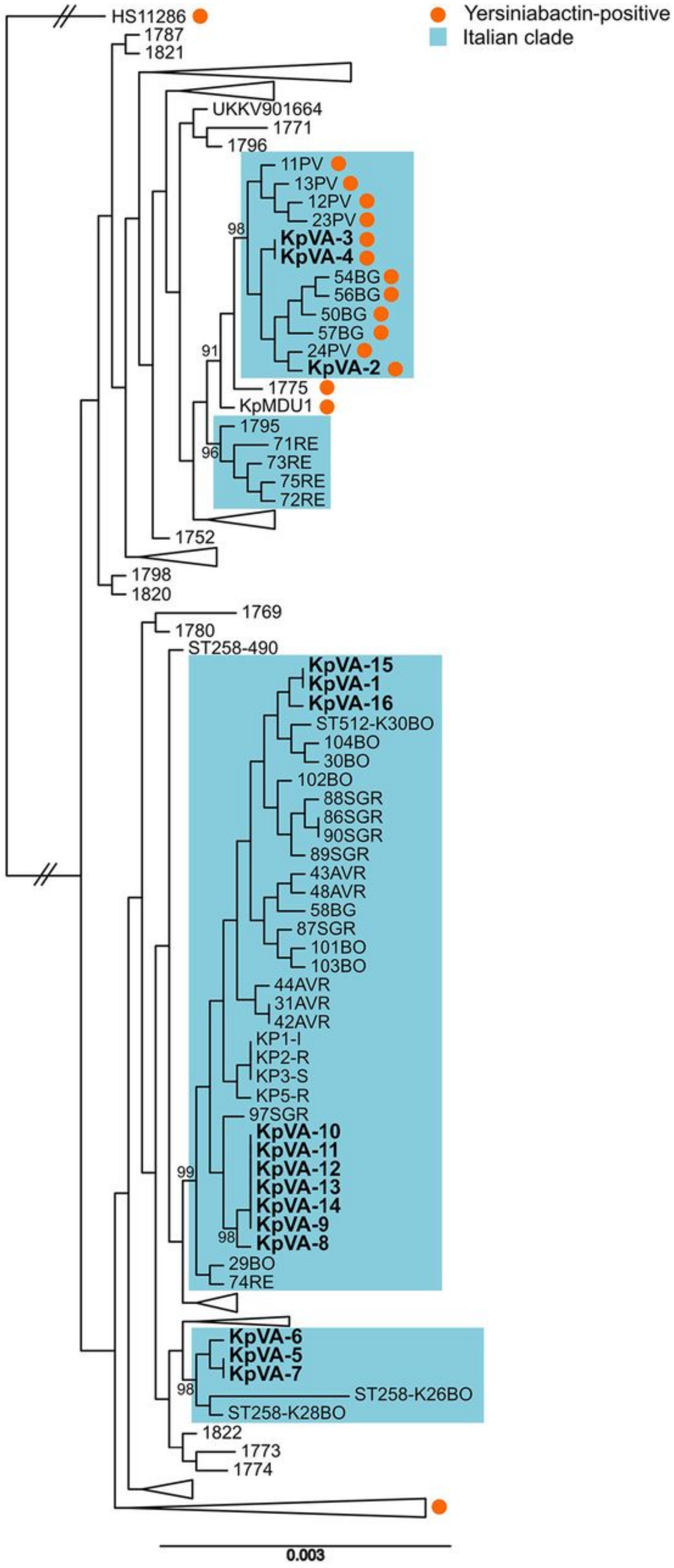


FIG 2. Representation of the phylogenetic relationships between isolates of clonal group 258 of *Klebsiella pneumoniae*, reconstructed using RaxML software with 100 bootstrap replicates and the generalized time-reversible model. The 16 novel isolates investigated in this study are highlighted in bold type. Highlighted in blue boxes are the four clades encompassing Italian isolates (both newly sequenced and taken from databases). Triangles represent coherent monophyletic clades of isolates from other countries, and orange dots indicate the presence of yersiniabactin genes. Bootstrap values are indicated only on nodes of interest, for the sake of image clarity. For a complete phylogeny of 335 *K. pneumoniae* isolates worldwide, see Fig. S1 in the supplemental material.

In analyses of the numbers of SNPs differentiating the isolates, the seven strains belonging to the investigated outbreak presented an average of 20 SNPs per genome in comparisons among them. Interestingly, a similar average number (27 SNPs per genome) differentiated strains KpVA-1, KpVA-15, and KpVA-16, which are grouped in a single clade but have been sampled over a longer time (about 5 months). This could indicate a difference in the measured paces of the molecular clock between the two clusters. Multiple hypotheses could explain the observed situation, such as the presence of different environmental conditions or a conservative pressure from purifying selection.

Core genome MLST. cgMLST analysis was performed with the same genomic data set used for the SNP phylogeny presented in Fig. 2 (219 *K. pneumoniae* genomes belonging to CG258). The resulting UPGMA tree (see Fig. S2 in the supplemental material) is largely consistent with the tree resulting from the SNP-based phylogenomic analysis. Specifically, in both analyses, the main subdivisions of CG258 (23) were clearly detectable, while the 16 genomes presented in this work were clustered in four monophyletic groups, one of which corresponded to the seven outbreak strains.

Outbreak reconstruction. To define the investigated ICU outbreak in greater detail, a genomic network was built using the core SNPs identified among the epidemic isolates, which were ordered according to the isolation date. The resulting structure showed a star-like topology (Fig. 3) centered around the isolate obtained from patient zero (KpVA-8). This structure suggested a nonlinear spread of infection. Thus, multiple events of contagion probably took place, all starting from patient zero and infecting six ICU patients (Fig. 3). It is important to note that patient zero (KpVA-8) stayed in the ICU for >2 months and the stays of the other infected patients coincided with his presence (Fig. 1). In addition, the location of his bed in the ward was not related to the date of infection (i.e., patients in beds closer to the bed of patient zero were not infected before patients in beds more distant from the bed of patient zero). Interestingly, the phylogenetic analysis of SNPs (Fig. 2) showed KpVA-8 (i.e., the isolate from patient zero) as the sister taxon of a single clade including the other six epidemic strains (Fig. 2). The result confirms that isolate KpVA-8 was at the origin of the outbreak.

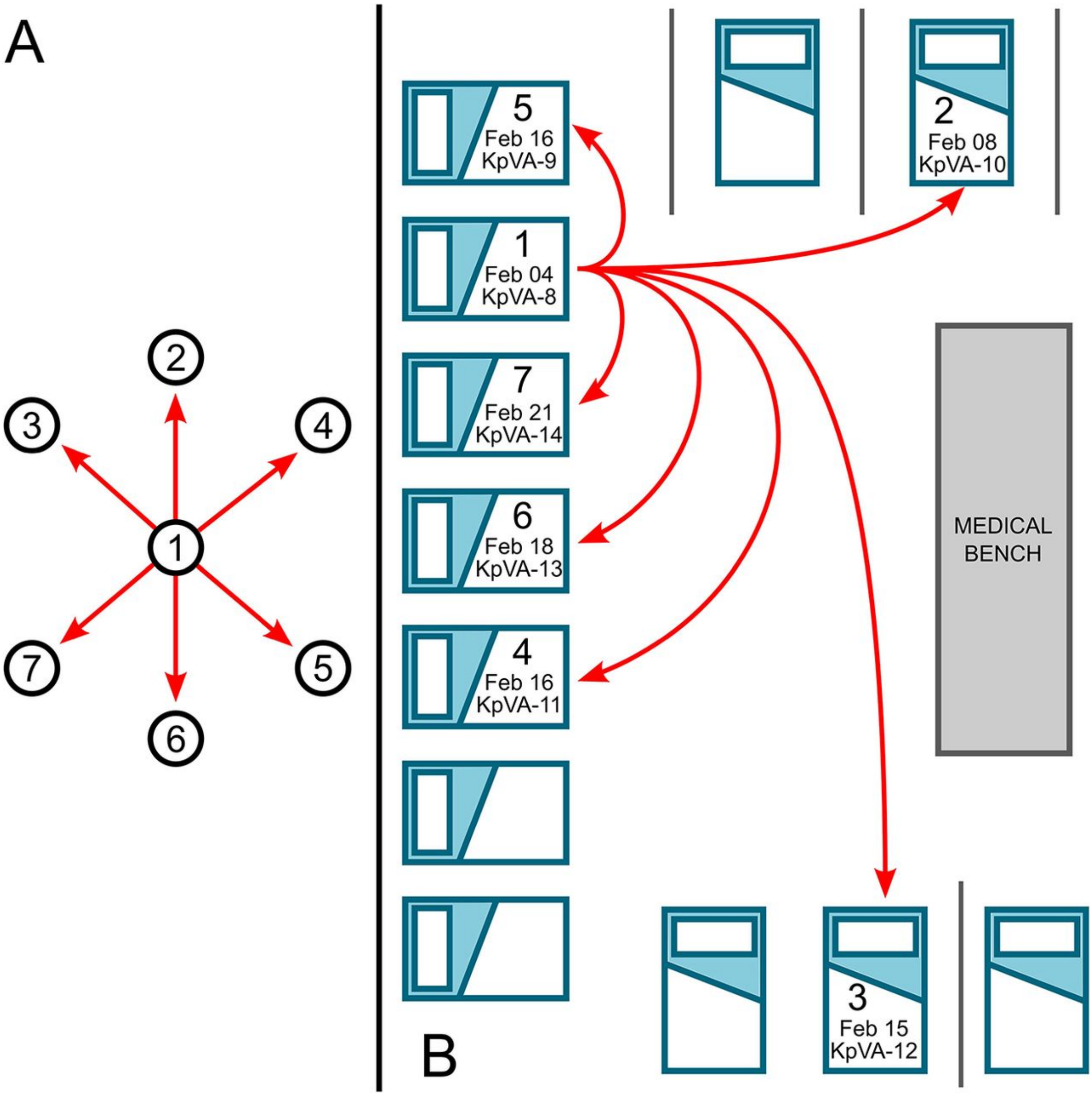


FIG 3. (A) Reconstruction of the star-like diffusion pattern, starting from isolate KpVA-8, among the seven *K. pneumoniae* isolates belonging to a single outbreak event that occurred in the ICU of Ospedale di Circolo e Fondazione Macchi in February 2013. The star-like topology was obtained using the R package Outbreaker. Numbers in bold indicate the temporal order of contagion. (B) Graphic representation of the bed-to-bed spread of infection on a map of the ICU.

DISCUSSION

This work was aimed at characterizing 16 carbapenem-resistant *K. pneumoniae* isolates collected from a single hospital during a 27-month period, including an epidemic that occurred in February 2013. This allowed us to identify the genomic characteristics of the isolates, to elucidate the epidemiological relationships among them, and to place them in the context of the global phylogeny of *K. pneumoniae*. In particular, whole-genome analyses allowed us to characterize the diversity of this bacterial species within a single hospital, to contextualize the local features within the global genomic spectrum of the species, and to reconstruct the spreading route of seven isolates within the ICU.

The 16 carbapenem-resistant isolates were shown to belong to CG258, the most prevalent KPC-producing *K. pneumoniae* lineage. Indeed, all 16 genomes demonstrated the gene *bla*$_{\text{KPC}}$, and none of them had other known carbapenem resistance genes. This is not surprising, considering previous reports that showed the worldwide diffusion and high prevalence of KPC isolates of this clonal group among carbapenem-resistant *K. pneumoniae* strains (21, 23).

When the drug resistance profiles of the investigated isolates were compared, the main difference was colistin resistance. In 10 of the 16 isolates, the colistin MIC was 4 mg/liter or higher (Table 2), thus above the EUCAST MIC breakpoint. Genomic analysis aimed at detecting the determinants for this resistance trait found that the 10 resistant isolates demonstrated insertion of an IS*5*-like transposon in the *mgrB*, a gene that regulates a pathway of lipopolysaccharide biosynthesis (43). Insertion of an IS*5*-like sequence in the *mgrB* gene has indeed been proposed as a determinant of colistin resistance (32). Our results appear to support this causative link, considering that none of the six colistin-sensitive strains presented the aforementioned insertion.

The global phylogeny of *K. pneumoniae* (Fig. 2) reveals that the 16 isolates are assigned to three of the four previously characterized Italian clades of CG258. These four clades have been proposed to represent four different dissemination events for KPC isolates in Italy, from 2008 to 2010 (21). Therefore, within a time span of 27 months, three of the four main Italian lineages of CG258 KPC-producing *K. pneumoniae* were detected in a single hospital in northwestern Italy. This result clearly indicates that at least three of the four lineages are currently circulating, creating a scenario of multiple, contemporary, overlapping epidemics.

With regard to virulence genes, all 16 isolates possessed the operons *fim*, *ent*, and *mrk* but lacked the genes *rmpA* and *wzy*-K1 (*magA*). The only detected difference among the 16 isolates was the presence in three isolates of four genes responsible for yersiniabactin synthesis (*ybtA*, *ybtS*, *irp1*, and *irp2*). Yersiniabactin has been reported to provide advantages in metabolism and multiplication, particularly in mixed infections and under iron-deprived conditions and especially in pulmonary infections, according to recent studies (14, 39, 40). When investigating the presence of yersiniabactin genes in the global CG258 *K. pneumoniae* phylogeny (Fig. 1), we detected a monophyletic group encompassing the three novel isolates as well as all other Italian isolates belonging to the same subclade, in addition to two U.S. isolates. This result clearly indicates that these genes were acquired before the diversification of this specific subclade and have been maintained since, which allows the characterization of this group of isolates as a yersiniabactin-positive monophyletic lineage of strains within CG258.

Seven of the 16 isolates, which were collected from ICU patients over a period of 17 days, were hypothesized to represent a single epidemic event. The monophyletic relationships among the seven epidemic isolates (Fig. 2) confirmed the hypothesis. A specific analysis for outbreak reconstruction showed that the epidemic isolates were connected in a star-like diagram (Fig. 3) originating from the isolate from patient zero. This is congruent with the rapid spread among the seven ICU patients.

Epidemic isolates could not be differentiated from sporadic isolates based on a specific pattern of the presence/absence of genes coding for virulence factors. The drug resistance profiles of these isolates also were identical to those of some sporadic isolates. This indicates that the spread of the outbreak was not related to genes conferring a specific advantage to the epidemic clone. Rather, it suggests that external factors might have caused the spread of the clone among ICU patients.

In conclusion, this study shows how whole-genome analysis can facilitate accurate reconstruction of the spread of bacterial pathogens. The wealth of data from genome sequencing allows reconstruction of the relationships of isolates from single hospitals and outbreaks and placement of the isolates in overall global phylogenies, and comparisons of genomes permit determination of whether strains involved in a specific outbreak share common characteristics that might confer specific selective advantages. The introduction of bacterial genomics into clinical settings will thus allow reconstruction of the routes and causes of nosocomial infections, with estimation of the relative roles of human- and microbe-related factors.

REFERENCES

1. Podschun R, Ullmann U. 1998. Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. Clin Microbiol Rev 11:589–603.

2. Elemam A, Rahimian J, Mandell W. 2009. Infection with pan resistant Klebsiella pneumoniae: a report of 2 cases and a brief review of the literature. Clin Infect Dis 49:271–274. 10.1086/600042.

3. Endimiani A, Hujer AM, Perez F, Bethel CR, Hujer KM, Kroeger J, Oethinger M, Paterson DL, Adams MD, Jacobs MR, Diekema DJ, Hall GS, Jenkins SG, Rice LB, Tenover FC, Bonomo RA. 2009. Characterization of blaKPC-containing Klebsiella pneumoniae isolates detected in different institutions in the eastern USA. J Antimicrob Chemother 63:427–437. 10.1093/jac/dkn547.

4. Yigit HA, Queenan M, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, Alberti S, Bush K, Tenover FC. 2001. Novel carbapenem-hydrolyzing β-lactamase, KPC-1, from a carbapenem-resistant strain of Klebsiella pneumoniae. Antimicrob Agents Chemother 45:1151–1161. 10.1128/AAC.45.4.1151-1161.2001.

5. Kumarasamy KK, Toleman MA, Walsh TR, Bagaria J, Butt F, Balakrishnan R, Chaudhary U, Doumith M, Giske CG, Irfan S, Krishnan P, Kumar AV, Maharjan S, Mushtaq S, Noorie T, Paterson DL, Pearson A, Perry C, Pike R, Rao B, Ray U, Sarma JB, Sharma M, Sheridan E, Thirunarayan MA, Turton J, Upadhyay S, Warner M, Welfare W, Livermore DM, Woodford N. 2010. Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study. Lancet Infect Dis 10:597–602. 10.1016/S1473-3099(10)70143-2.

6. Wesselink JJ, López-Camacho E, de la Peña S, Ramos-Ruiz R, Ruiz-Carrascoso G, Lusa-Bernal S, Fernández-Soria VM, Gómez-Gil R, Gomez-Puertas P, Mingorance J. 2012. Genome sequence of OXA-48 carbapenemase-producing Klebsiella pneumoniae KpO3210. J Bacteriol 194:6981. 10.1128/JB.01897-12.

7. Miriagou V, Tzelepi E, Gianneli D, Tzouvelekis LS. 2003. Escherichia coli with a self-transferable, multiresistant plasmid coding for metallo-β-lactamase VIM-1. Antimicrob Agents Chemother 47:395–397. 10.1128/AAC.47.1.395-397.2003.

8. Fukigai S, Alba J, Kimura S, Iida T, Nishikura N, Ishii Y, Yamaguchi K. 2007. Nosocomial outbreak of genetically related IMP-1 β-lactamase-producing Klebsiella pneumoniae in a general hospital in Japan. Int J Antimicrob Agents 29:306–310. 10.1016/j.ijantimicag.2006.10.011.

9. Bogdanovich T, Adams-Haduch JM, Tian GB, Nguyen MH, Kwak EJ, Muto CA, Doi Y. 2011. Colistin-resistant Klebsiella pneumoniae carbapenemase (KPC)-producing Klebsiella pneumoniae belonging to the international epidemic clone ST258. Clin Infect Dis 53:373–376. 10.1093/cid/cir401.

10. Mammina C, Bonura C, Di Bernardo F, Aleo A, Fasciana T, Sodano C, Saporito MA, Verde MS, Tetamo R, Palma DM. 2012. Ongoing spread of colistin-resistant Klebsiella pneumoniae in different

wards of an acute general hospital, Italy, June to December 2011. Euro Surveill 17(33):pii=20248. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20248.

11. Struve C, Bojer M, Krogfelt KA. 2009. Identification of a conserved chromosomal region encoding Klebsiella pneumoniae type 1 and type 3 fimbriae and assessment of the role of fimbriae in pathogenicity. Infect Immun 77:5016–5024. 10.1128/IAI.00585-09.

12. Oteo J, Saez D, Bautista V, Fernández-Romero S, Hernández-Molina JM, Pérez-Vázquez M, Aracil B, Campos J, Spanish Collaborating Group for the Antibiotic Resistance Surveillance Program. 2013. Carbapenemase-producing Enterobacteriaceae in Spain in 2012. Antimicrob Agents Chemother 57:6344–6347. 10.1128/AAC.01513-13.

13. Nassif X, Sansonetti PJ. 1986. Correlation of the virulence of Klebsiella pneumoniae K1 and K2 with the presence of a plasmid encoding aerobactin. Infect Immun 54:603–608.

14. Bachman MA, Oyler JE, Burns SH, Caza M, Lépine F, Dozois CM, Weiser JN. 2011. Klebsiella pneumoniae yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. Infect Immun 79:3309–3316. 10.1128/IAI.05114-11.

15. Lawlor MS, O'Connor C, Miller VL. 2007. Yersiniabactin is a virulence factor for Klebsiella pneumoniae during pulmonary infection. Infect Immun 75:1463–1472. 10.1128/IAI.00372-06.

16. Lai YC, Peng HL, Chang HY. 2003. RmpA2, an activator of capsule biosynthesis in Klebsiella pneumoniae CG43, regulates K2 cps gene expression at the transcriptional level. J Bacteriol 185:788–800. 10.1128/JB.185.3.788-800.2003.

17. Wu MF, Yang CY, Lin TL, Wang JT, Yang FL, Wu SH, Hu BS, Chou TY, Tsai MD, Lin CH, Hsieh SL. 2009. Humoral immunity against capsule polysaccharide protects the host from magA+ Klebsiella pneumoniae-induced lethal disease by evading Toll-like receptor 4 signaling. Infect Immun 77:615–621. 10.1128/IAI.00931-08.

18. Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, Barbe V, Bengoechea JA, Sansonetti P, Brisse S, Tournebize R. 2014. Comparative analysis of Klebsiella pneumoniae genomes identifies a phospholipase D family protein as a novel virulence factor. BMC Biol 12:41. 10.1186/1741-7007-12-41.

19. Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, Garin B, Le Hello S, Arlet G, Nicolas-Chanoine MH, Decré D, Brisse S. 2014. Genomic definition of hypervirulent and multidrug-resistant Klebsiella pneumoniae clonal groups. Emerg Infect Dis 20:1812–1820. 10.3201/eid2011.140206.

20. Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia M, Landini MP, Pongolini S, Sambri V, Bandi C, Marone P, Sassera D. 2015. Genomic epidemiology of Klebsiella pneumoniae: the Italian scenario, and novel insights into the

origin and global evolution of resistance to carbapenem antibiotics. Antimicrob Agents Chemother 59:389–396. 10.1128/AAC.04224-14.

21. Nordmann P. 2014. Carbapenemase-producing Enterobacteriaceae: overview of a major public health challenge. Med Mal Infect 44:51–56. 10.1016/j.medmal.2013.11.007.

22. Deleo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, Chavda KD, Jacobs MR, Mathema B, Olsen RJ, Bonomo RA, Musser JM, Kreiswirth BN. 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae. Proc Natl Acad Sci U S A 111:4988–4993. 10.1073/pnas.1321364111.

23. Chen L, Mathema B, Pitout JD, DeLeo FR, Kreiswirth BN. 2014. Epidemic Klebsiella pneumoniae ST258 is a hybrid strain. mBio 5(3):e01355–14.

24. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, NISC Comparative Sequencing Program Group, Henderson DK, Palmore TN, Segre JA. 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 4:148ra116.

25. Mancini F, Carniato A, Ciervo A. 2009. Pneumonia caused by Shigella sonnei in man returned from India. Emerg Infect Dis 15:1874–1875. 10.3201/eid1511.090126.

26. Andrade LN, Curiao T, Ferreira JC, Longo JM, Clímaco EC, Martinez R, Bellissimo-Rodrigues F, Basile-Filho A, Evaristo MA, Del Peloso PF, Ribeiro VB, Barth AL, Paula MC, Baquero F, Cantón R, Darini AL, Coque TM. 2011. Dissemination of blaKPC-2 by the spread of Klebsiella pneumoniae clonal complex 258 clones (ST258, ST11, ST437) and plasmids (IncFII, IncN, IncL/M) among Enterobacteriaceae species in Brazil. Antimicrob Agents Chemother 55:3579–3583. 10.1128/AAC.01783-10.

27. Poirel L, Walsh TR, Cuvillier V, Nordmann P. 2011. Multiplex PCR for detection of acquired carbapenemase genes. Diagn Microbiol Infect Dis 70:119–123. 10.1016/j.diagmicrobio.2010.12.002.

28. Stahlhut SG, Chattopadhyay S, Struve C, Weissman SJ, Aprikian P, Libby SJ, Fang FC, Krogfelt KA, Sokurenko EV. 2009. Population variability of the FimH type 1 fimbrial adhesin in Klebsiella pneumoniae. J Bacteriol 191:1941–1950. 10.1128/JB.00601-08.

29. Wu CC, Lin CT, Cheng WY, Huang CJ, Wang ZC, Peng HL. 2012. Fur-dependent MrkHI regulation of type 3 fimbriae in Klebsiella pneumoniae CG43. Microbiology 158:1045–1056. 10.1099/mic.0.053801-0.

30. Brisse S, Fevre C, Passet V, Issenhuth-Jeanjean S, Tournebize R, Diancourt L, Grimont P. 2009. Virulent clones of Klebsiella pneumoniae: identification and evolutionary scenario based on genomic and phenotypic characterization. PLoS One 4:e4982. 10.1371/journal.pone.0004982.

31. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Comput Sci Biol Proc Ger Conf Bioinformatics 99:45–56.

32. Cannatelli A, D'Andrea MM, Giani T, Di Pilato V, Arena F, Ambretti S, Gaibani P, Rossolini GM. 2013. In vivo emergence of colistin resistance in Klebsiella pneumoniae producing KPC-type carbapenemases mediated by insertional inactivation of the PhoQ/PhoP mgrB regulator. Antimicrob Agents Chemother 57:5521–5526. 10.1128/AAC.01480-13.

33. Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. 10.1371/journal.pone.0011147.

34. Sassera D, Comandatore F, Gaibani P, D'Auria G, Mariconti M, Landini MP, Sambri V, Marone P. 2014. Comparative genomics of closely related strains of Klebsiella pneumoniae reveals genes possibly involved in colistin resistance. Ann Microbiol 64:887–890. 10.1007/s13213-013-0727-5.

35. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. 10.1093/bioinformatics/btu033.

36. Jolley KA, Maiden MC. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. 10.1186/1471-2105-11-595.

37. Jombart T, Eggo RM, Dodd PJ, Balloux F. 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106:383–390. 10.1038/hdy.2010.78.

38. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol 10:e1003457. 10.1371/journal.pcbi.1003457.

39. Koczura R, Mokracka J, Kaznowski A. 2012. The Yersinia high-pathogenicity island in Escherichia coli and Klebsiella pneumoniae isolated from polymicrobial infections. Pol J Microbiol 61:71–73.

40. Chaturvedi KS, Hung CS, Crowley JR, Stapleton AE, Henderson JP. 2012. The siderophore yersiniabactin binds copper to protect pathogens during infection. Nat Chem Biol 8:731–736. 10.1038/nchembio.1020.

41. Bach S, de Almeida A, Carniel E. 2000. The Yersinia high-pathogenicity island is present in different members of the family Enterobacteriaceae. FEMS Microbiol Lett 183:289–294. 10.1111/j.1574-6968.2000.tb08973.x.

42. Cannatelli A, Di Pilato V, Giani T, Arena F, Ambretti S, Gaibani P, D'Andrea MM, Rossolini GM. 2014. In vivo evolution to colistin resistance by PmrB sensor kinase mutation in KPC-producing Klebsiella pneumoniae is associated with low-dosage colistin treatment. Antimicrob Agents Chemother 58:4399–4403. 10.1128/AAC.02555-14.

43. Lippa AM, Goulian M. 2009. Feedback inhibition in the PhoQ/PhoP signaling system by a membrane peptide. PLoS Genet 5:e1000788. 10.1371/journal.pgen.1000788.

ARTICLE 4

Differential single nucleotide polymorphism-based analysis of an outbreak caused by *Salmonella enterica* serovar Manhattan reveals epidemiological details missed by standard pulsed-field gel electrophoresis

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

2015. 53:1227–1238.
doi:10.1128/JCM.02930-14.

Differential Single Nucleotide Polymorphism-Based Analysis of an Outbreak Caused by *Salmonella enterica* Serovar Manhattan Reveals Epidemiological Details Missed by Standard Pulsed-Field Gel Electrophoresis

Erika Scaltriti[a], Davide Sassera[b], Francesco Comandatore[b,c], Marina Morganti[a], Carmen Mandalari[a], Stefano Gaiarsa[c,d], Claudio Bandi[c], Gianguglielmo Zehender[e], Luca Bolzoni[f], Gabriele Casadei[a] and Stefano Pongolini[a,f]

Author Affiliations

[a] Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Sezione di Parma, Parma, Italy
[b] Dipartimento di Biologia e Biotecnologie, Università di Pavia, Pavia, Italy
[c] Dipartimento di Scienze Veterinarie e Sanità Pubblica (DIVET), Università degli Studi di Milano, Milan, Italy
[d] Fondazione IRCCS Policlinico San Matteo, Pavia, Italy
[e] Dipartimento di Scienze Cliniche L. Sacco, Università degli Studi di Milano, Milan, Italy
[f] Direzione Sanitaria – Servizio di Analisi del Rischio, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Parma, Italy

D. J. Diekema, Editor
Received 13 October 2014. Returned for modification 13 November 2014. Accepted 25 January 2015.
Accepted manuscript posted online 4 February 2015.

Address correspondence to Stefano Pongolini, **stefano.pongolini@izsler.it**.

Supplemental material is available at the end of the thesis

This article can be found at **http://dx.doi.org/10.1128/JCM.02930-14**.
Please scan this QR code to access the website from the printed version

ACKNOWLEDGMENTS. We acknowledge Elena Carra for providing some of the isolates included in the study. We thank Roberto Alfieri for technical assistance in remote analysis at the University of Parma Department of Physics and Earth Sciences. This study was supported by Regione Lombardia grant delibera regionale 001051/22122000 and by the Italian Ministry of Health grant IZSLER-PRC2012/006.

ABSTRACT

We retrospectively analyzed a rare *Salmonella enterica* serovar Manhattan outbreak that occurred in Italy in 2009 to evaluate the potential of new genomic tools based on differential single nucleotide polymorphism (SNP) analysis in comparison with the gold standard genotyping method, pulsed-field gel electrophoresis. A total of 39 isolates were analyzed from patients (n = 15) and food, feed, animal, and environmental sources (n = 24), resulting in five different pulsed-field gel electrophoresis (PFGE) profiles. Isolates epidemiologically related to the outbreak clustered within the same pulsotype, SXB_BS.0003, without any further differentiation. Thirty-three isolates were considered for genomic analysis based on different sets of SNPs, core, synonymous, nonsynonymous, as well as SNPs in different codon positions, by Bayesian and maximum likelihood algorithms. Trees generated from core and nonsynonymous SNPs, as well as SNPs at the second and first plus second codon positions detailed four distinct groups of isolates within the outbreak pulsotype, discriminating outbreak-related isolates of human and food origins. Conversely, the trees derived from synonymous and third-codon-position SNPs clustered food and human isolates together, indicating that all outbreak-related isolates constituted a single clone, which was in line with the epidemiological evidence. Further experiments are in place to extend this approach within our regional enteropathogen surveillance system.

INTRODUCTION

Salmonellosis is a major food-borne disease worldwide, with an estimated 93.8 million cases occurring each year, resulting in 155,000 deaths (1). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks (2) indicated that nontyphoid salmonellosis was the second most reported food-borne zoonosis in Europe in 2012, trailing only behind *Campylobacter jejuni* infection. The 2012 overall notification rate for human salmonellosis in the European Union (EU) was 22.2 episodes per 100,000 population, for a total of 91,034 confirmed cases, with hospitalization and mortality rates of 45.1% and 0.14%, respectively. The highest proportions of *Salmonella*-positive foodstuff samples were reported for fresh turkey, poultry, and pork at 4.4%, 4.1%, and 0.7%, respectively (2). In order to manage this food-borne infection and to limit its health and economic burdens, surveillance programs have developed and implemented DNA-based subtyping methods to identify outbreaks in a timely manner and to trace infections back to their food sources. Over the past decades, the two most intensively used protocols for *Salmonella* subtyping have been pulsed-field gel electrophoresis (PFGE) and multilocus variable-number tandem-repeat analysis (MLVA) (3). Unfortunately, these methods rely on just few features of the entire bacterial genome (rare restriction sites for PFGE or few polymorphic loci for MLVA) to assess the relatedness of different isolates. During epidemiological investigations of food-borne outbreaks, this limitation might lead to difficulties in distinguishing outbreak-related from outbreak-unrelated *Salmonella enterica* subsp. *enterica* isolates due to the high genetic homogeneity of this subspecies (4). Multilocus sequence typing (MLST) is another molecular tool for bacterial typing based on allelic differences in the loci of specified housekeeping genes (5). While proposed as an alternative to classical serotyping (6), MLST does not seem to be discriminatory enough when all isolates being tested belong to the same serotype (7). With the aim of improving resolution in molecular epidemiology, the technological advancements of whole-genome sequencing (WGS) may provide an unprecedented opportunity to access the entire genome information at a reasonable cost, as well as to set a new series of high-resolution standards in molecular epidemiology. As PFGE and MLVA are able to resolve more genotypes within a single serovar, WGS has already proved its resolution power to detect variations within otherwise undistinguishable bacterial clones (by PFGE or MLVA), as shown by recent examples in the literature (8, 9). Large studies based on WGS within *S. enterica* subspecies (10) and within serovars in *S. enterica* subsp. *enterica* (11, 12) contributed to the elucidation of *Salmonella* phylogenetic diversity and also accomplished important steps forward in the area of bacterial disease tracking. Moreover, serovar-specific studies on *S. enterica* subsp. *enterica* have highlighted microevolutionary differences among clinical, environmental, and food isolates in *S. enterica* serovars Montevideo (13, 14), Enteritidis (4), Newport (15), Typhimurium (16–18), and Heidelberg (12), which would have been missed by more traditional approaches.

While outbreaks of more common serovars, such as *Salmonella* Typhimurium and *Salmonella* Enteritidis, have been reported and investigated, only a few human outbreaks due to *S. enterica* serovar Manhattan have been reported (19, 20) worldwide in the past 60 years, and none have been characterized at the genomic level. Here, we present a WGS-based retrospective analysis of the only *Salmonella* Manhattan outbreak ever documented in Italy, which occurred from June to July 2009 in a relatively small geographic area in the province of Modena.

The outbreak investigation at the time of the event was carried out by international standard epidemiological techniques (21) and by PFGE on the isolates from patients and food, feed, animal, and environmental sources.

The aim of this study was 2-fold: (i) to evaluate the effectiveness of WGS to accurately identify the relationships among all the outbreak-related isolates with enough resolution to clarify the ambiguities that PFGE was not able to unravel, and (ii) to explore and test new genomic tools for bacterial molecular epidemiology based on synonymous and nonsynonymous single-nucleotide polymorphisms (SNPs) and SNPs in different codon positions.

We selected this specific *Salmonella* Manhattan outbreak to test our WGS pipeline because of three main features that made this outbreak a particularly suitable case study. First, *Salmonella* Manhattan is considered a rare serotype, as confirmed by the regional surveillance system for *Salmonella* of Emilia-Romagna, which over the past 3 years recorded a yearly average of only 5.6 sporadic cases over a total of 924 isolates per year, from a regional population of about 5,000,000 (M. Morganti, E. Scaltriti, L. Bolzoni, G. Casadei, and S. Pongolini; Enter-net Italia, unpublished data). This low prevalence of *Salmonella* Manhattan infection provides a reasonable confidence that virtually all isolates collected in the outbreak area at the time of the episode belonged to the outbreak, therefore preventing the noise effect due to unrelated isolates wrongly assigned to the epidemic. Second, the investigation conducted at the time of the outbreak was successful in tracing the infection back to a food point source using internationally coded epidemiological methods (21); bacterial isolates were also recovered not only from food (pork sausage) at the retail level but also along the food chain up to the raw meat used to prepare the implicated food (at the production establishment). Third, the regional surveillance system for *Salmonella* of

Emilia-Romagna, hosted at the Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), holds a full collection of *Salmonella* Manhattan strains covering the years 2001 to present. This set of isolates was pivotal in the conduct of a successful epidemiological investigation and for testing our WGS-based analyses of this rare serovar.

CASE REPORT

The diagnostic unit of Parma of IZSLER is the Regional Reference Center for Surveillance of Enteropathogens (Enter-net) of clinical, environmental, animal, and food origins. Within this activity, a cluster of 15 human infections caused by *Salmonella* Manhattan was detected in the province of Modena from June to July 2009. All 15 isolates showed the same PFGE profile, SXB_BS.0003, strengthening the hypothesis that the unusually high incidence of this rare serovar was due to an epidemic outbreak. Consequently, an epidemiological investigation was undertaken and, considering the rarity of the serovar involved, all 21 isolates of *Salmonella* Manhattan available from the surveillance collection of IZSLER were genotyped by PFGE to get possible clues about the source of the outbreak. Thirteen isolates from the collection had the same PFGE profile as that of the outbreak strain, but only three of them had been isolated just before the onset of the outbreak (May/June 2009). Two had been isolated from pork sausage at the establishment of an industrial producer that distributed in the outbreak area, while one had been recovered from swine intestine at an establishment near the outbreak area that processed guts for the salami industry. According to the epidemiological investigation, the gut processing establishment had no correlation with the outbreak. However, as its isolate presented the same PFGE pulsotype as that of the outbreak-related isolates, health authorities were left with a certain degree of uncertainty about its possible role. Following the results of the epidemiological and molecular analyses, food samples were collected at retail sources in the outbreak area and at the establishment producing the sausage in order to confirm the source and clonality of the outbreak strain. Two samples from retail-collected sausages, along with a sample from fresh pork supplies of the sausage producer, scored positive for the outbreak pulsotype. Based on these results, the sausage from the implicated producer was recalled, leading to the outbreak extinction.

MATERIALS AND METHODS

Bacterial isolates. A total of 39 *Salmonella* Manhattan isolates were included in the study. Fifteen isolates were involved in the epidemic episode, another three isolates were collected within the epidemiological investigation, and 21 were collected between 2001 and 2009 during the surveillance activity of IZSLER (Table 1). The isolates were isolated and streak purified with standard microbiological techniques and stocked at −80°C. They were cultured on plates with Trypticase soy agar with 5% defibrinated sheep blood (TSA-SB) and incubated overnight at 37°C before being typed by pulsed-field gel electrophoresis, according to the PulseNet protocol (22). The isolates selected for WGS were inoculated into brain heart infusion broth and cultured overnight at 37°C with agitation (200 rpm).

Lab no.	Isolate no. (this study)	Date of isolation (DD/MM/YYYY)	Isolation place (province)	Matrix	PFGE pulsotype
160969_3	SM1[b]	06/30/2009	Modena	Human	SXB_BS.0003
160969_5	SM2[b]	06/30/2009	Modena	Human	SXB_BS.0003
160969_6	SM3[b]	06/30/2009	Modena	Human	SXB_BS.0003
165051_2	SM4[b]	07/03/2009	Modena	Human	SXB_BS.0003
165051_3	SM5[b]	07/03/2009	Modena	Human	SXB_BS.0003
165051_5	SM6[b]	07/03/2009	Modena	Human	SXB_BS.0003
165051_7	SM7[b]	07/30/2009	Modena	Human	SXB_BS.0003
111113	SM8[b]	07/03/2009	Modena	Human	SXB_BS.0003
165051_11	SM9[b]	07/03/2009	Modena	Human	SXB_BS.0003
165051_12	SM10[b]	07/03/2009	Modena	Human	SXB_BS.0003
180073_1	SM11[b]	07/22/2009	Modena	Human	SXB_BS.0003
180073_2	SM12[b]	07/22/2009	Modena	Human	SXB_BS.0003
180073_3	SM13[b]	07/22/2009	Modena	Human	SXB_BS.0003
180073_4	SM14[b]	07/22/2009	Modena	Human	SXB_BS.0003
180073_6	SM15[b]	07/22/2009	Modena	Human	SXB_BS.0003
250920	SM42[b]	08/31/2009	Milano	Pork	SXB_BS.0003
227021	SM32[b]	05/06/2009	Milano	Pork sausage	SXB_BS.0003
188801	SM52[b]	05/06/2009	Milano	Pork sausage	SXB_BS.0003
216630_1	SM53[b]	09/03/2009	Modena	Pork sausage	SXB_BS.0003
216630_2	SM54[b]	09/03/2009	Modena	Pork sausage	SXB_BS.0003
226957	SM16	03/07/2006	Mantova	Swine	SXB_PR.0753
226963	SM17[b]	03/20/2006	Mantova	Swine	SXB_PR.0753
226972	SM19[b]	03/20/2006	Sondrio	Pork salami	SXB_PR.0753
226979_1	SM21[b]	07/31/2006	Cremona	Swine gut	SXB_BS.0003
226985	SM23[b]	08/03/2006	Milano	Pork sausage	SXB_BS.0003
226987	SM24[b]	08/03/2006	Milano	Pork sausage	SXB_BS.0003
226993	SM26	01/22/2007	Ravenna	Hamburger	SXB_BS.0003
226998	SM27[b]	06/29/2007	Milano	Pork	SXB_BS.0003

227002	SM28	09/18/2002	Pavia	Surface water	SXB_BS.0003
227009	SM29[b]	09/02/2002	Bologna	Bovine sausage	SXB_PR.0754
227015	SM31	09/11/2001	Pavia	Surface water	SXB_PR.0751
227033	SM35[b]	11/29/2008	Ravenna	Swine stool	SXB_BS.0003
227039	SM36[b]	09/30/2008	Brescia	Swine stool	SXB_PR.0752
227052	SM38[b]	09/24/2008	Milano	Swine stool	SXB_BS.0003
188806	SM48[b]	06/03/2009	Reggio Emilia	Swine intestine	SXB_BS.0003
188790	SM47	10/01/2002	Pavia	Surface water	SXB_BS.0003
188795	SM49[b]	03/09/2009	Brescia	Chicken farm	SXB_PR.0753
188787	SM51	09/17/2002	Pavia	Surface water	SXB_BS.0003
188781	SM50[b]	07/31/2001	Modena	Minced pork	SXB_PR.0751

TABLE 1. Complete data set of *Salmonella* Manhattan isolates analyzed in this study[a]. [a]The isolates above the line break are the outbreak-related isolates (15 human-origin and 5 food-origin isolates), and those below the line break are the 19 *Salmonella* Manhattan collection isolates. SM32 and SM52 were also collection isolates, but they were eventually attributed to the outbreak, following the results of this study. [b]These *Salmonella* Manhattan isolates were selected for whole-genome sequencing.

Pulsed-field gel electrophoresis. All isolates were genotyped by PFGE, according to the PulseNet protocol (22). Genomic DNA underwent XbaI restriction before electrophoresis in a Chef Mapper XA system (Bio-Rad, CA, USA). The PFGE patterns were analyzed using the BioNumerics Software version 6.6 (Applied-Maths, Sint-Martens-Latem, Belgium) and associated with isolate information in our surveillance database. Clustering of the PFGE profiles was generated using the unweighted-pair group method using averages (UPGMA) based on the Dice similarity index (optimization, 1%; band matching tolerance, 1%). Following a comparison of the electrophoretic profiles, a PFGE pattern (pulsotype) was assigned to each isolate within the Regional Surveillance Database of Emilia-Romagna.

Whole-genome sequencing. All outbreak-related isolates and a selection of the IZSLER *Salmonella* Manhattan collection, representative of the different pulsotypes detected, were subjected to WGS (Table 1), for a total of 33 isolates. Genomic DNA was extracted from overnight cultures using the Qiagen DNeasy blood and tissue kit (Qiagen) and quality controlled and quantified using a Synergy H1 hybrid multimode microplate reader (BioTek, Winooski, VT, USA). The sequencing libraries were prepared with the Nextera XT sample preparation kit (Illumina, San Diego, CA, USA), and sequencing was performed on the Illumina MiSeq platform, with a 2 × 250-bp paired-end run.

Read quality check and assembly. All read sets were evaluated for sequence quality and read-pair length using the softwares FastQC and Flash (23). FastQC allowed us to assess the overall quality of the generated sequences, while Flash was used to measure the distance between the sequence read pairs. All the read sets that passed the quality check (visual check for FastQC and average read pair distance >100 nucleotides [nt] for Flash) were assembled with MIRA 4.0 (24) using accurate settings for *de novo* assembly mode.

***In silico* multilocus sequence typing.** *In silico* MLST was performed using the MLST scheme optimized by the University of Warwick (http://mlst.warwick.ac.uk/mlst/dbs/Senterica).

Comparative genomics by local variation calling. In a previous work, we sequenced and published the first improved high-quality draft genome (25) of *Salmonella* Manhattan (strain 111113) (26). The 18 contigs of the *Salmonella* Manhattan 111113 genome, belonging to a human isolate of the outbreak presented here, were concatenated in a pseudochromosome and used as a reference for alignment of each of the other 32 genome assemblies included in this study, using progressiveMauve (27). A previously described bioinformatic pipeline (28) was then used to merge the results of all isolates for comparison and to extract the coordinates of all local variations spanning from SNPs to longer variations (mutations, insertions, and deletions), based on the annotation of the reference genome of strain 111113. Core SNPs were identified as single nondegenerate mutated bases flanked by identical bases and present in all 33 genomes (including that of strain 111113). Genes presenting at least one core SNP were selected and compared against the Virulence Factors Database (VFDB) (29–31), using a BLAST search with a 10^{-5} E value cutoff.

Analysis of variations. Open reading frames (ORFs) were predicted and translated on all assembled genomes (including the previously published *Salmonella* Manhattan strain 111113 genome [26]) using Prodigal (32). Next, every genomic variation (SNPs, mutations, insertions, and deletions) was parsed in order to assign it to one of the following subsets of isolates: (i) all outbreak-related isolates, irrespective of the human, food, or raw meat origin; (ii) outbreak-related human-origin-only isolates; and (iii) outbreak-related food-origin-only isolates (including those from sausage and raw meat).

Phylogenetic analysis. From the core SNP data set, different subsets were generated: (i) nonsynonymous SNPs, (ii) synonymous SNPs, and (iii) SNPs at the first, second, or third codon position. The core and subsets of SNPs were used as inputs for generating SNP-based phylogenies using the maximum likelihood (ML) or the Bayesian methods. Model choice was evaluated in JModelTest (33). Maximum likelihood analyses were run in PhyML (34), with a generalized time-reversible (GTR) substitution model and 100 bootstrap iterations, while Bayesian analyses were run in MrBayes (35, 36), using the same model for 2,000,000 generations, with chains sampled every 1,000 generations. The final parameter values and trees were summarized after discarding 25% of the posterior sample. The ML and Bayesian trees were displayed and edited for publication with FigTree version 1.4.0.

Nucleotide sequence accession numbers. The genome sequences of *Salmonella* Manhattan (strain 111113; study identification [ID], SM8) contigs were previously deposited at EBI under the accession no. CBKW010000001 to CBKW010000021 (project PRJEB1854). The newly 32 sequenced genomes (contigs) of *Salmonella* Manhattan were deposited at EBI under the project number PRJEB5339 and are summarized here in the format isolate lab no./study identification no.: WGS accession number: 160969_3/SM1: CCBJ010000610 to CCBJ010000701, 160969_5/SM2: CCBJ010000175 to CCBJ010000212, 160969_6/SM3: CCBJ010000291 to CCBJ010000308, 165051_2/SM4: CCBJ010001977 to CCBJ010002069, 165051_3/SM5: CCBJ010002070 to CCBJ010000089, 165051_5/SM6: CCBJ010000001 to CCBJ010000100, 165051_7/SM7: CCBJ010004043 to CCBJ010004081, 165051_11/SM9: CCBJ010003194 to CCBJ010003512, 165051_12/SM10: CCBJ010000309 to CCBJ010000327, 180073_1/SM11: CCBJ010002338 to CCBJ010002378, 180073_2/SM12: CCBJ010003726 to CCBJ010003749, 180073_3/SM13: CCBJ010001070 to CCBJ010001515, 180073_4/SM14: CCBJ010001516 to CCBJ010001924, 180073_6/SM15: CCBJ010000702 to CCBJ010000770, 250920/SM42: CCBJ010000328 to CCBJ010000609, 227021/SM32: CCBJ010004870 to CCBJ010004957, 188801/SM52: CCBJ010002097 to CCBJ010002229, 216630_1/SM53: CCBJ010002817 to CCBJ010003193, 216630_2/SM54: CCBJ010000213 to CCBJ010000238, 226963/SM17: CCBJ010002257 to CCBJ010002337, 226972/SM19: CCBJ010002230 to CCBJ010002256, 226979_1/SM21: CCBJ010000101 to CCBJ010000174, 226985/SM23: CCBJ010003750 to CCBJ010004042, 226987/SM24: CCBJ010003702 to CCBJ010003725, 226998/SM27: CCBJ010000771 to CCBJ010001069, 227009/SM29: CCBJ010001925 to CCBJ010001976, 227033/SM35: CCBJ010000239 to CCBJ010000268, 227039/SM36: CCBJ010000269 to CCBJ010000290, 227052/SM38: CCBJ010002379 to CCBJ010002816, 188806/SM48: CCBJ010003540 to CCBJ010003701, 188795/SM49: CCBJ010004082 to CCBJ010004692, and 188781/SM50: CCBJ010004693 to CCBJ010004869.

RESULTS

We present here reanalysis by WGS of an outbreak caused by *Salmonella* Manhattan in the province of Modena (Italy) in 2009. The isolates from the human cases were SM1, -2, -3, -4, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, and -15. Out of the 21 collection isolates available, all were genotyped by PFGE to search for clues on the source of infection, and SM21, -23, -24, -26, -27, -28, -32, -35, -38, -47, -48, -51, and -52 showed the outbreak pulsotype; however, SM36, -16, -17, -19, -29, -31, -49, and -50 belonged to different pulsotypes, and a selection of them were included in this study as outgroup isolates. SM42, -53, and -54 were isolated during the microbiological follow-up of the episode and presented the outbreak pulsotype.

Pulsed-field gel electrophoresis. The 39 *Salmonella* Manhattan isolates of the study showed five different XbaI-PFGE profiles: SXB_BS.0003, SXB_PR.0753, SXB_PR.0754, SXB_PR.0751, and SXB_PR.0752 (Fig. 1). All the human isolates (SM1 to SM15) showed the same PFGE profile (SXB_BS.0003), supporting the hypothesis that the unusually high incidence of this rare serovar was due to a single epidemic clone.

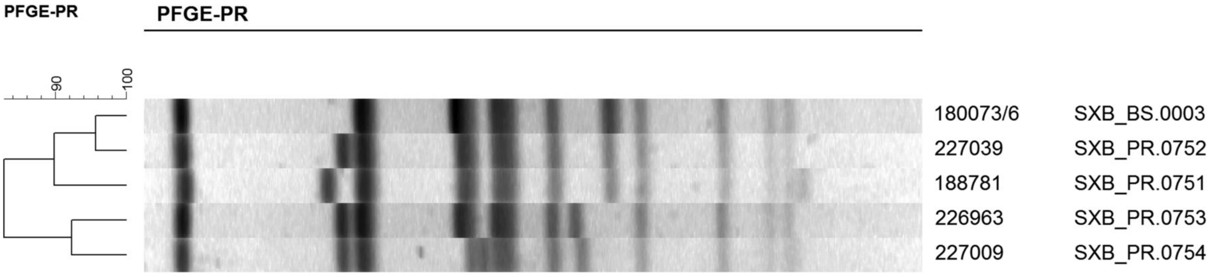


FIG 1. Similarity of *Salmonella* Manhattan isolates, examined in this study, inferred by pulsed-field gel electrophoresis profiles (PFGE-PR). The samples underwent XbaI restriction and pattern analysis according to the standard PulseNet protocol. The UPGMA dendrogram of all the profiles of the study is reported on the left; the ruler indicates the similarity values. The laboratory numbers of the isolates and their pulsotypes are reported on the right.

Another 13 isolates from the IZSLER surveillance collection belonged to genotype SXB_BS.0003. Among these, three (SM32, SM48, and SM52) dated back to just before the outbreak period (May/June 2009) and were pivotal in guiding the epidemiological investigation. SM48 originated from an establishment near the outbreak area that processed swine guts for the salami industry. Due to this microbiological and molecular finding, the establishment was suspected of having a role in the outbreak, although no evident correlation with the human infections was made. More significantly, SM32 and SM52 were isolated just before the onset of the episode from pork sausages produced at an industrial establishment that shipped to retail stores in the outbreak area. Consequently, sausages from this producer, which were on sale in the outbreak area, were sampled along with the pork purchased by the producer. Both the sausages and the pork were positive for *Salmonella* Manhattan with the outbreak pulsotype (SXB_BS.0003) (isolates SM53 and SM54 from the sausages and SM42 from pork). Interestingly, two *Salmonella* Manhattan isolates from our collection, isolated within the own-check hygiene procedures of the producer (SM23 and SM24) 3 years before the outbreak, presented the same genotype. Also, the surveillance collection isolates SM21, -26, -27, -28, -35, -38, -47, and -51 shared the outbreak pulsotype, but they did not seem to be correlated with the outbreak or source of infection.

Among the other non-outbreak PFGE profiles detected, the pulsotype SXB_PR.0752 (isolate SM36) had 95% similarity with the outbreak pulsotype, while the genotypes SXB_PR.0751 (isolates SM31 and SM50), SXB_PR.0753 (isolates SM16, SM17, SM19, and SM49), and SXB_PR.0754 (isolate SM29) were less similar (90%, 84%, and 84%, respectively) (Fig. 1).

Whole-genome sequencing. The genomes of the 33 *Salmonella* Manhattan isolates considered for genomic analysis, including the already deposited genome of strain 111113 (26), were sequenced, quality checked, and assembled to draft status, from an average of 2,593,738 MiSeq paired-end reads per genome. The average sequenced genome characteristics were 4,678,201 nt in length, 150 large (>1,000 nt) contigs, and an N_{50} of 212,360. The genome data for each isolate are listed in Table S1 in the supplemental material. The MLST profile was determined for all draft genomes, which were found to belong to the same sequence type (ST), ST18. All assembled genomes underwent comparative and phylogenetic analyses.

Analysis of variations. A comparative genomic analysis was implemented to detect the differences between the *Salmonella* Manhattan genomes, in terms of nucleotide variations, exclusive to (i.e., present in all the isolates of a group and absent in all the others) the outbreak-related isolates, as divided into the following main groups: (i) all outbreak-related isolates, irrespective of the human, food, or raw meat origin; (ii) outbreak-related human-origin-only isolates; and (iii) outbreak-related food-origin-only isolates (including sausages and raw meat).

Of all the non-degenerate nucleotide variations (total 9,410) discovered by the progressiveMauve algorithm, 14 were outbreak specific, and all were core SNPs (two intergenic, two synonymous, and 10 nonsynonymous), divided as six variations exclusive to all outbreak-related isolates, three variations characteristic of the food-origin-only outbreak-related isolates, and five characteristic of the outbreak-related human-origin-only isolates (Table 2).

Group of isolates	Amino acid change	Codon change	Position CDS[a]	Type of SNP	Gene	Locus tag	Strand	Product name
All outbreak	C→R	TGT→CGT	625	Genic	*cobT*	SMA01→2283	–	Nicotinate-nucleotide–dimethylbenzimidazole phosphoribosyltransferase
	N→N	AAT→AAC	156	Genic	*gntR*	SMA01→3706	–	Gluconate utilization system Gnt-transcriptional repressor
	A→T	GCC→ACC	577	Genic	*ansB*	SMA01→3765	–	L-Asparaginase
	V→A	GTC→GCC	988	Genic	*dcuC*	SMA01→4465	–	Putative cryptic C4-dicarboxylate transporter
				Intergenic				
	K→E	AAA→GAA	70	Genic	*betI*	SMA01→1140	+	Transcriptional regulator, TetR family
Human origin	M→T	ATG→ACG	584	Genic	*dsbI*	SMA01→0572	+	Thiol-disulfide oxidoreductase, DsbB-like
	A→T	GCC→ACC	310	Genic	*sthD*	SMA01→3447	–	β-fimbriae usher protein
	V→V	GTT→GTC	465	Genic	*ispH*	SMA01→3526	+	4-hydroxy-3-methylbut-2-enyl diphosphate reductase
	Q→STOP	CAA→TAA	252	Genic	*rfbD*	SMA01→4557	+	UDP-galactopyranose mutase
				Intergenic				
Food origin	S→I	AGC→ATC	872	Genic	*fliK*	SMA01→2244	+	Flagellar hook-length control protein FliK
	P→L	CCT→CTT	17	Genic		SMA01→0101	+	Hypothetical protein
	A→V	GCC→GTC	1544	Genic	*fdrA*	SMA01→4374	+	Protein FdrA: acyl-CoA synthetase[b]

TABLE 2. Characteristic SNPs of three groups of outbreak-related isolates. [a] CDS, coding sequence. [b] CoA, coenzyme

Phylogenetic analysis. Phylogeny was reconstructed using an SNP-based approach. SNPs were extracted from the assembled genomes using a bioinformatic pipeline (28) based on progressiveMauve (27). Of the 9,410 detected variations, 953 were core SNPs, with 224 being synonymous and 467 being nonsynonymous; the remaining 262 SNPs were marked as intergenic. Among the synonymous SNPs, 6% and 94% were located in the first and third codon positions, respectively, while among the nonsynonymous SNPs, 43% were in the first, 42% in the second (total, 85% for the two positions), and 15% in the third codon position. The number of synonymous and nonsynonymous core SNPs at the first, second, and third positions were 214, 194, and 283, respectively.

The phylogenetic analysis of the study isolates was performed separately based on the different subsets of SNPs considered, namely, core, synonymous, nonsynonymous, and different codon positions using both Bayesian (Fig. 2 to 4) and maximum likelihood algorithms (see Fig. S1 and S2 in the supplemental material). Both algorithms returned the same phylogenetic results on each subset.

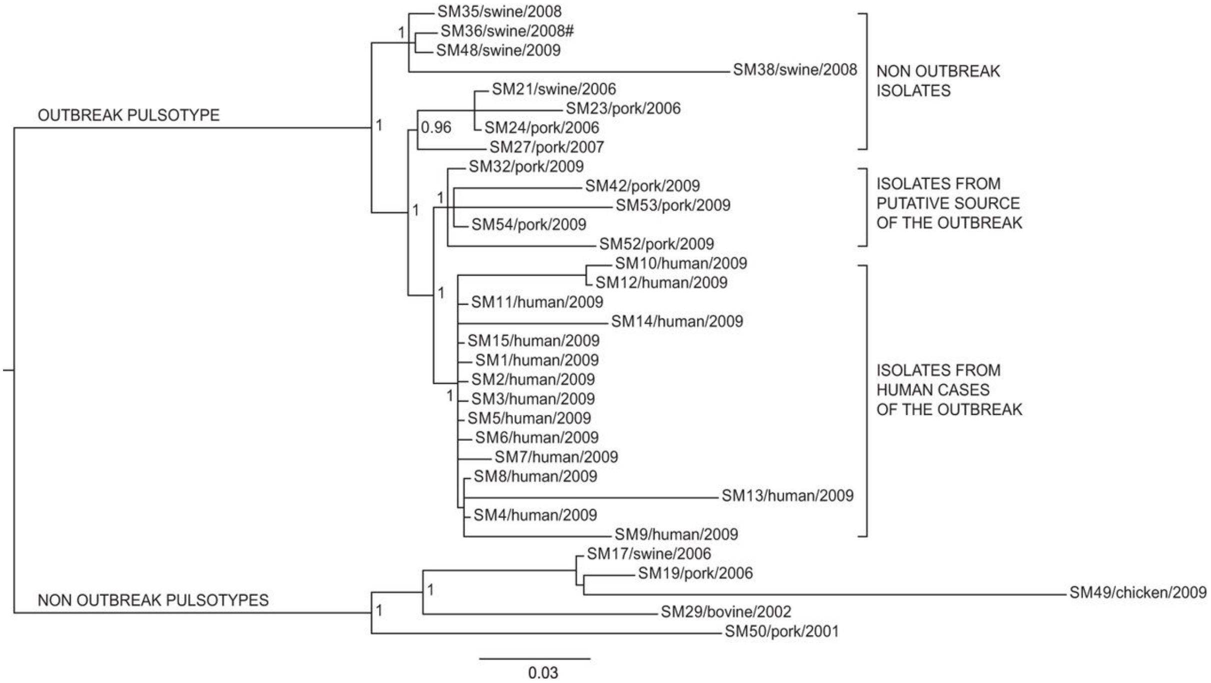


FIG 2. Bayesian phylogeny of the 33 *Salmonella* Manhattan sequenced genomes based on core SNPs. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).

All data sets identified two major clades: one grouping all the isolates belonging to pulsotype SXB_BS.0003 and the highly related SXB_PR.0752 (95% similarity), and the other constituted by isolates with different pulsotypes (SXB_PR.0753, SXB_PR.0754, and SXB_PR.0751). Interestingly, WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of pulsotype SXB_BS0003, meaning they are highly related compared to isolates of the other pulsotypes of the study. Therefore, we considered SXB_PR.0752 together with SXB_BS.0003 for the subsequent analyses of phylogeny and presence of variants.

Phylogeny based on core SNPs revealed four main groups inside the outbreak pulsotype. Isolates that were not epidemiologically related to the outbreak formed two monophyletic clusters, with the outermost one grouping isolates from various locations and previous years but always from swine stool within the own-check procedures of pig farms (isolates SM35, SM36, and SM38) or at food processing plants (isolate SM48). The other group included isolates collected at the sausage-producing establishment within its hygiene monitoring system 3 years before the outbreak (SM23 and SM24), along with an isolate collected on a pig farm in the same period (SM21). Isolate SM27 originated from another food processing plant in the same area of the sausage producer, but that was never linked to the outbreak.

The two innermost clusters included all the outbreak-related isolates. Five strains isolated from sausages prepared by the implicated producer (SM32, SM42, SM52, SM53, and SM54), both at a retail locations in the outbreak area and at the establishment, which were distinct from the cluster of human isolates of the outbreak (from SM1 to SM15). All outbreak-related isolates are monophyletic, confirming their derivation from a common ancestor. In order to better investigate the relationships among those isolates, we performed additional analyses on specific subsets of the core SNPs to take into account the possible effects of selective evolutionary pressure. We separately considered nonsynonymous SNPs, synonymous SNPs, and SNPs at the first, second, and third codon positions as presumptively subjected to decreasing selective pressures (37). The trees corresponding to the different subsets of SNPs are shown in Fig. 3 and 4. The trees generated by nonsynonymous SNPs and SNPs at the first plus second and second codon positions showed the same topology described by the whole data set of core SNPs, with a clear distinction between outbreak-related isolates of human and food origins. The phylogenies generated by SNPs under minor selective pressure (i.e., third position) revealed different scenarios, with the loss of a node inside the outbreak cluster showing isolates of human origin as a subgroup within the food-origin outbreak isolates. Considering synonymous SNPs only, the outbreak isolates of human and food origins are grouped in one cluster, being indicative of a single circulating clone. The phylogenetic inferences made by Bayesian and maximum likelihood algorithms gave identical results (see Fig. S1 and S2 in the supplemental material).

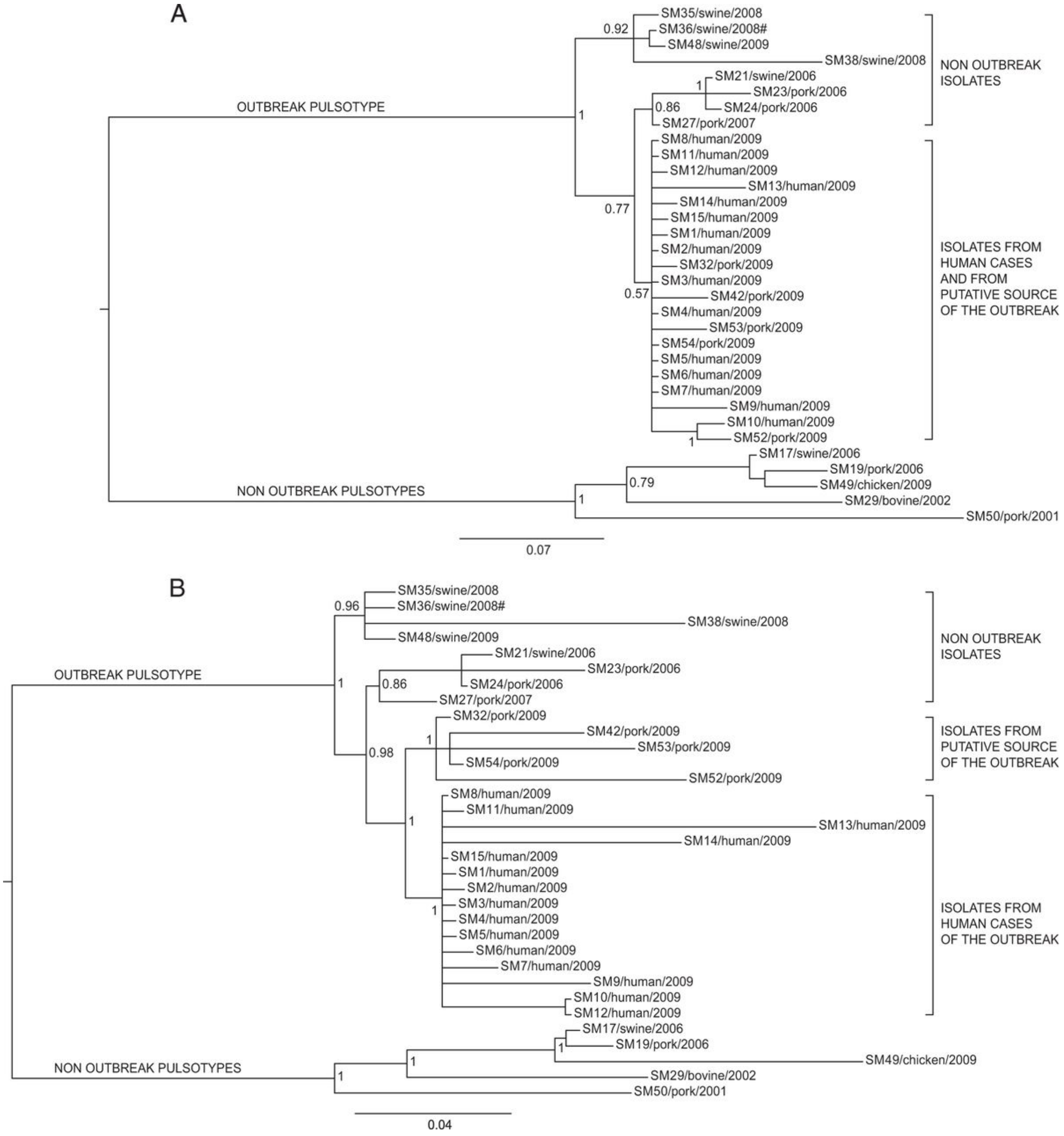


FIG 3. Phylogenetic Bayesian analysis of the 33 *Salmonella* Manhattan sequenced genomes based on synonymous (A) and nonsynonymous (B) SNP data sets. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).

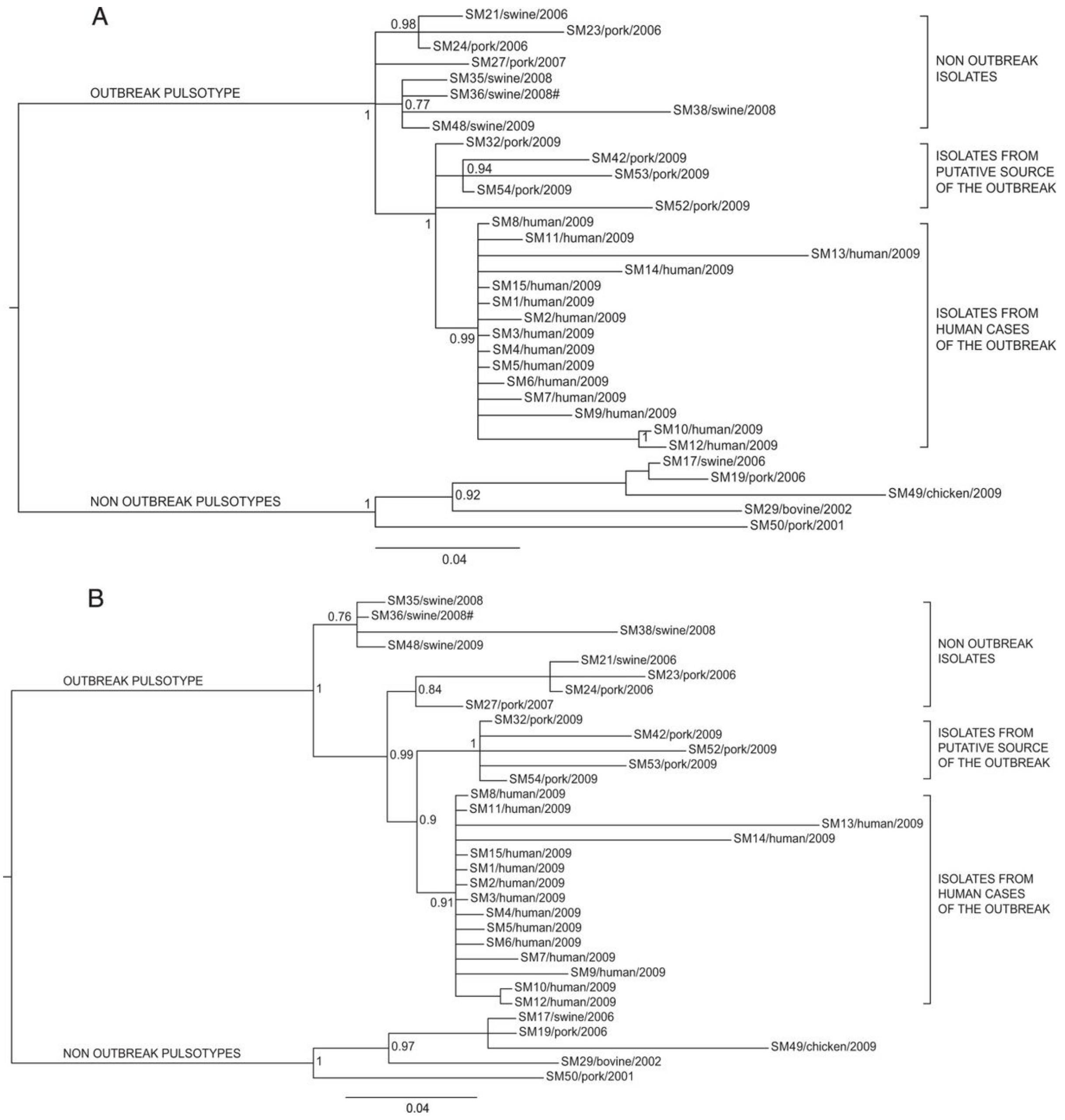


FIG 4 (A-B). Phylogenetic Bayesian analysis of the 33 *Salmonella* Manhattan sequenced genomes based on SNPs in first (A), second (B), third (C), and first plus second codon position (D) data sets. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).

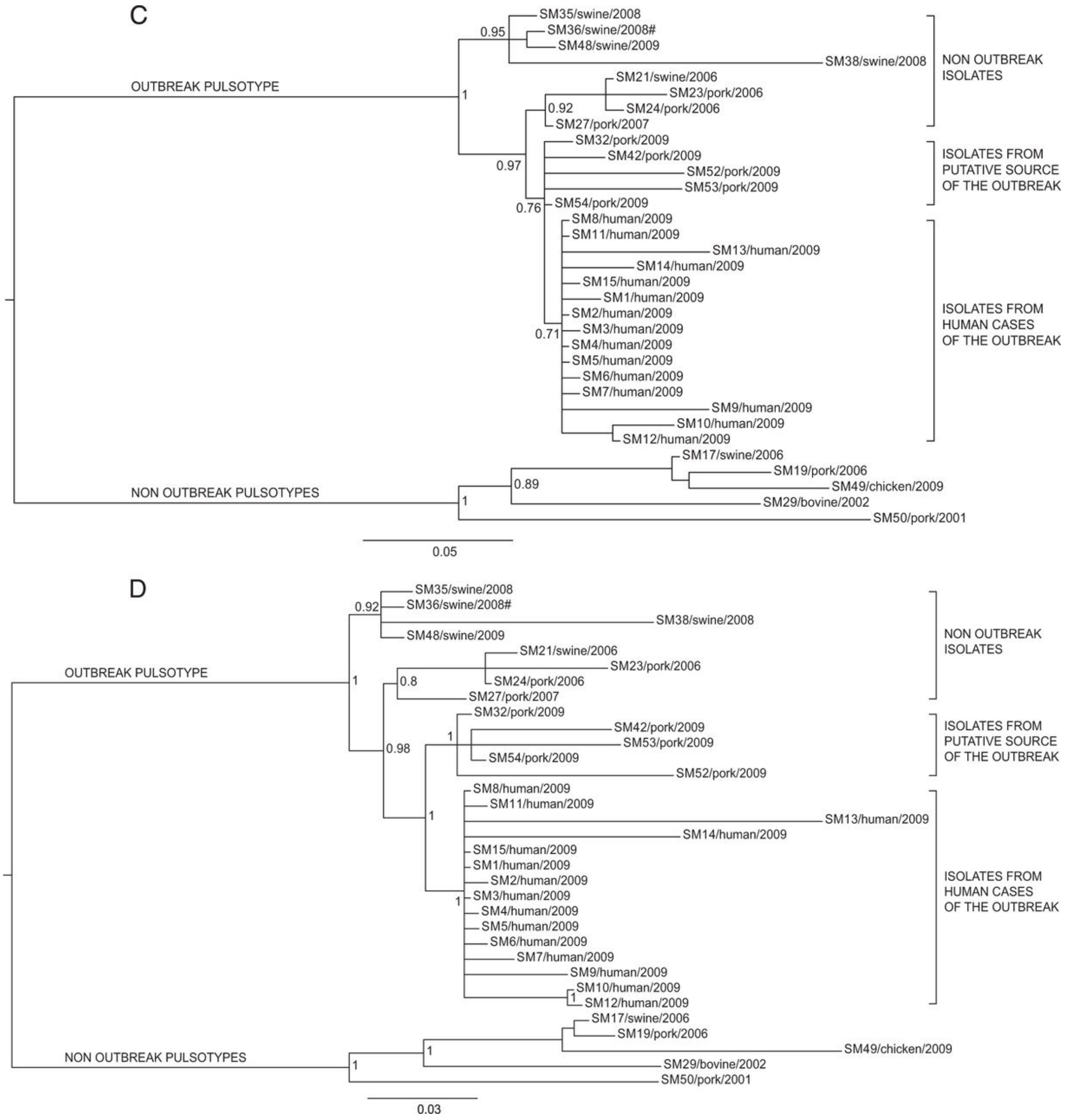


FIG 4 (C-D). Phylogenetic Bayesian analysis of the 33 *Salmonella* Manhattan sequenced genomes based on SNPs in first (A), second (B), third (C), and first plus second codon position (D) data sets. The posterior probabilities are indicated in each principal node of the tree. The scale bar units are the nucleotide substitutions per site. #, WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).

DISCUSSION

Microbiologists often need to determine the relatedness of bacterial isolates to define the network of relationships of an infectious outbreak and effectively assist epidemiological investigations. Standard protocols for typing *Salmonella* rely on internationally accepted methods, like PFGE and MLVA, which a few decades ago flanked the more limited serotyping. The possibility of accessing the vast amount of information provided by WGS of bacterial isolates promises to be the next frontier of subtyping methods, probably capable of surpassing PFGE and MLVA for molecular epidemiological purposes. In this study, we reanalyzed a well-defined *Salmonella* Manhattan outbreak detected in the summer of 2009 in the province of Modena (Italy) using WGS in order to test the power of this approach for resolving the ambiguities left by PFGE. The epidemic episode involved 15 human cases from June to July 2009, with all presenting the same PFGE profile (SXB_BS.0003). The molecular epidemiological investigation of the outbreak involved several isolates, some from the infectious episode and others from the historic collection of the regional surveillance system of the food chain. As expected, PFGE analysis attributed the same pulsotype (SXB_BS.0003) to all the outbreak-related isolates, but the same pulsotype was shared by many historic isolates as well. On the contrary, the WGS-based phylogeny inferred from the total core SNPs clearly showed the presence of four distinct groups of isolates (Fig. 2) within the outbreak pulsotype. The first branch of the tree, within the outbreak pulsotype, separates nonoutbreak historic isolates recovered from swine stool at different locations and times. Among these, we find isolate SM48, which was originally suspected of being implicated in the infectious episode, based on PFGE, and eventually cleared by WGS. Interestingly, isolate SM36, which does not belong to pulsotype SXB_BS.0003 but to the highly similar (95% similarity) pulsotype SXB_PR.0752, is included in this clade. This is a clear discrepancy between WGS and the more limited PFGE that relies on only few genomic loci (rare restriction sites) for its typing inferences. By placing SM36 together with pulsotype SXB_BS.0003 isolates, our WGS approach indicates that a limited genomic difference between isolates is able to jeopardize the typing outcome of PFGE. This observation confirms what Tenover et al. (38) already pointed out, the fact that as PFGE may be heavily influenced by a single mutational event (e.g., SNP occurring in a restriction site), isolates should be considered to be possibly related even if they differ by two or three bands. However, according to this conservative interpretation of PFGE results, the vast majority of the isolates of our study should be regarded as potentially belonging to the outbreak. This would have not been sufficiently discriminatory to help the epidemiological investigations. The interpretation criteria of Tenover et al. (38) are derived from logical considerations; as such, they are intrinsically valid, and our observations regarding isolate SM36 confirms their validity. At the same time, their use leaves molecular epidemiologists with considerable uncertainty about how to interpret PFGE results with regard to whether or not different pulsotypes are part of a single outbreak. In our case, WGS removed that uncertainty about SM36.

Moving deeper along the phylogenetic tree based on the total core SNPs, three other groups of isolates are evident. The outermost set of this node includes isolates (SM21, -23, -24, and -27) not related to the outbreak, as they were collected 3 years before (2006). It is interesting, however, to notice that WGS-based phylogeny indicates these strains to be closer to the outbreak node (inner branch) than was the previous set of swine-stool isolates. On a better look, we were struck by the fact that SM23 and SM24 were collected in 2006, within the own-check procedures of the sausage producer involved in the 2009 outbreak. Moreover, SM21, which is subbranched with SM23 and SM24, was routinely recovered from a local pig farm (from swine stool) at the same time as SM23 and SM24. While this specific molecular similarity was not inferred by PFGE, WGS highlighted a possible link between these two commercial entities. Moving one branch forward in the phylogenetic tree, WGS shows another bifurcation actually separating outbreak-related isolates of human origin from those of food origin. While still speculative, based on this WGS-based phylogeny, coupled with epidemiological data, we could argue that this outbreak was due to a persistent *Salmonella* Manhattan clone, which may have infected one or more pig farms and reached the food producer and the retail customers as animals arrived at the slaughterhouse in a nonclinical septic condition. This is a typical mode of transmission of *Salmonella* along the food chain, as it may asymptomatically persist (thus going unnoticed) within a herd of pigs for long periods of time (even years). Sporadically, animals carrying a high level of the pathogen arrive at the slaughterhouse and contaminate a defined set of food products, thus causing an infectious outbreak as the final consumers (39, 40) become exposed to it. In this scenario, WGS seems to depict a more detailed and articulated epidemiological story. In fact, the tree inferred from core SNPs (Fig. 2) leaves a certain level of uncertainty relative to the actual causative relationship between the isolates of food origin and of human origin within the outbreak, as they cluster in two distinct groups, although very closely to each other, as evidenced by the limited number of exclusive core SNPs accumulated by the two groups (3 for food and 5 for human isolates). In the absence of epidemiological insights, we argue that the two sets of isolates are very similar to each other but still are separate entities. This substantially contradicts the epidemiological evidence that the two sets of isolates belong to the same outbreak clone. Therefore, we further investigated this apparent inconsistency of the WGS-based results by comparing new alternative phylogenies based on two different subsets of polymorphisms, synonymous and nonsynonymous, instead of the total core SNPs. The trees generated from these two subsets of SNPs were different (Fig. 3A and B). Phylogenetic analysis based on nonsynonymous SNPs (Fig. 3B) still divided the outbreak isolates of food and human origins, as in the approach based on total core SNPs. On the contrary, the tree obtained from synonymous SNPs (Fig. 3A) clustered the human isolates together with the food isolates, indicating that all outbreak-related *Salmonella* Manhattan strains constituted a single clone, in line with epidemiological evidence. While intriguing, this new outcome may have been the misleading effect of the smaller amount of data present in these new subsets than that with the total set of core SNPs, of which there were 953, whereas the number of synonymous and nonsynonymous SNPs were 224 and 467, respectively. Therefore, to confirm these results, we took a step forward in this approach by considering not just synonymous versus nonsynonymous SNPs but also taking into account the different codon position of each SNP in the core genome. *Salmonella* Manhattan synonymous SNPs were at the 3rd codon position 94% of the time, while nonsynonymous SNPs were at the 2nd 42% and at the first position 43% of the time (total, 85%). In this study, 1st, 2nd, and 3rd position SNPs accounted for 214, 194, and 283 nucleotide substitutions, respectively. The comparison of subsets of SNPs based on their codon site would then not be impaired by too-large differences in the amount of data processed by the phylogenetic algorithms. The tree obtained from second codon position (Fig. 4B) was comparable to that of the nonsynonymous SNPs, as expected, whereas the tree obtained from third codon position showed human isolates as a subgroup of the food isolates (Fig. 4C), essentially confirming the tree based on synonymous SNPs. These results show that at least limited to our outbreak, synonymous and third-position SNPs were the only ones able to describe the causal relationship between food (source of the outbreak) and clinical isolates in a way that was consistent with the epidemiological evidence. At the same time, our results indicate that nonsynonymous and total core SNPs may have led to misleading conclusions about the relationships between the human and food isolates of the outbreak. One last aspect that caught our attention by deciphering topologies of this WGS-based retrospective analysis was that SNP-based clustering of isolates separated human from food outbreak-related isolates when considering total core SNPs (Fig. 2). As we just discussed, this topology was mainly influenced by nonsynonymous mutations, which means it is possible to find distinctive nonsynonymous SNPs for each group of isolates (human versus food). Using progressiveMauve, we identified a set of 953 core SNPs, among which we selected those that were exclusive to specific clusters of interest: six SNPs exclusive to all outbreak isolates (human and food origin), three exclusive to all food origin outbreak isolates, and only five exclusive to all human origin outbreak isolates (Table 2). The extremely limited number of exclusive SNPs in food and human isolates within the outbreak is an additional compelling element indicative of the fact that these two groups of isolates did not have enough evolutionary time to significantly differentiate, indicating they belong to the same clone. A BLAST analysis of these SNPs against the Virulence Factors Database revealed three genes of particular interest: (i) *fliK*, coding for a flagellar hook-length control protein (41), (ii) *sthD*, a gene coding for a fimbrial outer membrane usher protein (42), and (iii) *rfbD*, coding for a UDP-galactopyranose mutase precursor involved in the synthesis of the O antigen of the lipopolysaccharide (LPS). All three proteins are virulence determinants in *Salmonella* (43–46). WGS has already proved its usefulness for elucidating the evolutionary diversity of large populations of bacterial isolates (11, 47, 48). In the specific case of *Salmonella*, WGS was successfully applied to illuminate the diversity of the pathogen within a vast epidemic episode, allowing highly efficient traceback of clinical and food isolates (4, 13). The results obtained in this study underscore the power of WGS-based methods, when applied together with the most appropriate phylogenetic tools, to resolve small outbreaks characterized by few and highly clonal bacterial isolates. Our comparative genomics approach was able to correctly cluster the clinical isolates within the composite scenario of outbreak-related and collection isolates. Accurate backtracking to the source of infection at the retail and industrial levels was made possible while flagging an originally overlooked suspicious correlation with a farm supplier and clearing an originally suspect food operator. Moreover, by selectively choosing the different types of detected nucleotide variations, we were able to read the message hidden within neutral mutations as opposed to the general use of total core SNPs. Further use of the differential analysis of synonymous and nonsynonymous mutations will test the validity of this approach in deciphering the details of infection transmission in the context of other outbreaks caused by *Salmonella* and, potentially, other pathogens.

REFERENCES

1. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM, International Collaboration on Enteric Disease ‘Burden of Illness’ Studies. 2010. The global burden of nontyphoidal Salmonella gastroenteritis. Clin Infect Dis 50:882–889. 10.1086/650733.

2. European Food Safety Authority (EFSA), European Centre for Disease Prevention and Control (ECDC). 2013. Scientific report of EFSA and ECDC: the European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2011. EFSA J 11:3129–3378. 10.2903/j.efsa.2013.3129.

3. Wattiau P, Boland C, Bertrand S. 2011. Methodologies for Salmonella enterica subsp. enterica subtyping: gold standards and alternatives. Appl Environ Microbiol 77:7877–7885. 10.1128/AEM.05527-11.

4. Allard MW, Luo Y, Strain E, Pettengill J, Timme R, Wang C, Li C, Keys CE, Zheng J, Stones R, Wilson MR, Musser SM, Brown EW. 2013. On the evolutionary history, population genetics and diversity among isolates of Salmonella Enteritidis PFGE pattern JEGX01.0004. PLoS One 8:e55254. 10.1371/journal.pone.0055254.

5. Urwin R, Maiden MCJ. 2003. Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol 11:479–487. 10.1016/j.tim.2003.08.006.

6. Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, Krauland MG, Hale JL, Harbottle H, Uesbeck A, Dougan G, Harrison LH, Brisse S, S. enterica MLST Study Group. 2012. Multilocus sequencing typing as a replacement for serotyping in Salmonella enterica. PLoS Pathog 8:e1002776. 10.1371/journal.ppat.1002776.

7. Fakhr MK, Nolan LK, Logue CM. 2005. Multilocus sequence typing lacks the discriminatory ability of pulsed-field gel electrophoresis for typing Salmonella enterica serovar Typhimurium. J Clin Microbiol 43:2215–2219. 10.1128/JCM.43.5.2215-2219.2005.

8. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474. 10.1126/science.1182395.

9. Gardy JL, Johnston JC, Sui SJH, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJM, Brinkman FSL, Brunham RC, Tang P. 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364:730–739. 10.1056/NEJMoa1003176.

10. Desai PT, Porwollik S, Long F, Cheng P, Wollam A, Clifton SW, Weinstock GM, McClelland M. 2013. Evolutionary genomics of Salmonella enterica subspecies. mBio 4:e00579-12. 10.1128/mBio.00579-12.

11. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, Van Kessel JS, Karns JS, Musser SM, Brown EW. 2013. Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. Genome Biol Evol 5:2109–2123. 10.1093/gbe/evt159.

12. Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, Ayers SL, Cinar HN, Muruvanda T, Li C, Allard MW, Whichard J, Meng J, Brown EW, McDermott PF. 2014. Comparative genomic analysis and virulence differences in closely related Salmonella enterica serotype Heidelberg isolates from humans, retail meats, and animals. Genome Biol Evol 6:1046–1068. 10.1093/gbe/evu079.

13. Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW. 2012. High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 13:32. 10.1186/1471-2164-13-32.

14. Lienau EK, Strain E, Wang C, Zheng J, Ottesen AR, Keys CE, Hammack TS, Musser SM, Brown EW, Allard MW, Cao G, Meng J, Stones R. 2011. Identification of a salmonellosis outbreak by means of molecular sequencing. N Engl J Med 364:981–982. 10.1056/NEJMc1100443.

15. Cao G, Meng J, Strain E, Stones R, Pettengill J, Zhao S, McDermott P, Brown E, Allard M. 2013. Phylogenetics and differentiation of Salmonella Newport lineages by whole genome sequencing. PLoS One 8:e55687. 10.1371/journal.pone.0055687.

16. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, Brown DJ, Coia JE, Mulvey MR, Gilmour MW, Petrovska L, De Pinna E, Kuroda M, Akiba M, Izumiya H, Connor TR, Suchard MA, Lemey P, Mellor DJ, Haydon DT, Thomson NR. 2013. Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. Science 341:1514–1517. 10.1126/science.1240578.

17. Pang S, Octavia S, Feng L, Liu B, Reeves PR, Lan R, Wang L. 2013. Genomic diversity and adaptation of Salmonella enterica serovar Typhimurium from analysis of six genomes of different phage types. BMC Genomics 14:718. 10.1186/1471-2164-14-718.

18. Leekitcharoenphon P, Friis C, Zankari E, Svendsen CA, Price LB, Rahmani M, Herrero-Fresno A, Fashae K, Vandenberg O, Aarestrup FM, Hendriksen RS. 2013. Genomics of an emerging clone of Salmonella serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. J Infect Dev Ctries 7:696–706. 10.3855/jidc.3328.

19. Noël H, Dominguez M, Weill FX, Brisabois A, Duchazeaubeneix C, Kerouanton A, Delmas G, Pihier N, Couturier E. 2006. Outbreak of Salmonella enterica serotype Manhattan infection associated

with meat products, France, 2005. Euro Surveill 11:270–273. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=660.

20. Fisher I, Crowcroft N. 1998. Enter-net/EPIET investigation into the multinational cluster of Salmonella Livingstone. Euro Surveill 2:pii=1271. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=1271.

21. European Food Safety Authority. 2012. Technical report: manual for reporting of food-borne outbreaks in accordance with Directive/99/EC from the year 2011. Supporting publication 2012:EN-265. European Safety Food Authority, Parma, Italy. http://www.efsa.europa.eu/en/supporting/doc/265e.pdf.

22. PulseNet. 2010. One-day (24–28 h) standardized laboratory protocol for molecular subtyping of Escherichia coli O157:H7, non-typhoidal Salmonella serotypes, and Shigella sonnei, by pulsed field gel electrophoresis (PFGE). Centers for Disease Control and Prevention, Atlanta, GA. http://www.cdc.gov/pulsenet/protocols/ecoli_salmonella_shigella_protocols.pdf.

23. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. 10.1093/bioinformatics/btr507.

24. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. In Computer science and biology. Proceedings of the German Conference on Bioinformatics, GCB '99. GCB, Hannover, Germany.

25. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart Consortium, Detter JC. 2009. Genome Project standards in a new era of sequencing. Science 326:236–237. 10.1126/science.1180614.

26. Sassera D, Gaiarsa S, Scaltriti E, Morganti M, Bandi C, Casadei G, Pongolini S. 2013. Draft genome sequence of Salmonella enterica subsp. enterica serovar Manhattan strain 111113, from an outbreak of human infections in northern Italy. Genome Announc 1:e00632-13. 10.1128/genomeA.00632-13.

27. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. 10.1371/journal.pone.0011147.

28. Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia M, Landini MP, Pongolini S, Sambri V, Bandi C, Marone P, Sassera D. 2014. Genomic epidemiology of Klebsiella pneumoniae in Italy and novel insights into the origin and global

evolution of its resistance to carbapenem antibiotics. Antimicrob Agents Chemother 59:389–396. 10.1128/AAC.04224-14.

29. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:D325–D328. 10.1093/nar/gki008.

30. Yang J, Chen L, Sun L, Yu J, Jin Q. 2007. VFDB 2008 release: an enhanced Web-based resource for comparative pathogenomics. Nucleic Acids Res 36:D539–D542. 10.1093/nar/gkm951.

31. Chen L, Xiong Z, Sun L, Yang J, Jin Q. 2011. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res 40:D641–D645. 10.1093/nar/gkr989.

32. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. 10.1186/1471-2105-11-119.

33. Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772. 10.1038/nmeth.2109.

34. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758–771. 10.1080/10635150802429642.

35. Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755. 10.1093/bioinformatics/17.8.754.

36. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. 10.1093/bioinformatics/btg180.

37. Bofkin L, Goldman N. 2006. Variation in evolutionary processes at different codon positions. Mol Biol Evol 24:513–521. 10.1093/molbev/msl178.

38. Tenover FC, Arbeit RD, Goering RV, Mickelsen PA, Murray BE, Persing DH, Swaminathan B. 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J Clin Microbiol 33:2233–2239.

39. Rostagno MH. 2009. Can stress in farm animals increase food safety risk? Foodborne Pathog Dis 6:767–776. 10.1089/fpd.2009.0315.

40. Rostagno MH, Callaway TR. 2012. Pre-harvest risk factors for Salmonella enterica in pork production. Food Res Int 45:634–640. 10.1016/j.foodres.2011.04.041.

41. Uchida K, Aizawa SI. 2014. The flagellar soluble protein FliK determines the minimal length of the hook in Salmonella enterica serovar Typhimurium. J Bacteriol 196:1753–1758. 10.1128/JB.00050-14.

42. Waters RC, O'Toole PW, Ryan KA. 2007. The FliK protein and flagellar hook-length control. Protein Sci 16:769–780. 10.1110/ps.072785407.

43. Suez J, Porwollik S, Dagan A, Marzel A, Schorr YI, Desai PT, Agmon V, McClelland M, Rahav G, Gal-Mor O. 2013. Virulence gene profiling and pathogenicity characterization of non-typhoidal Salmonella accounted for invasive disease in humans. PLoS One 8:e58449. 10.1371/journal.pone.0058449.

44. Weening EH, Barker JD, Laarakker MC, Humphries AD, Tsolis RM, Baumler AJ. 2005. The Salmonella enterica serotype Typhimurium lpf, bcf, stb, stc, std, and sth fimbrial operons are required for intestinal persistence in mice. Infect Immun 73:3358–3366. 10.1128/IAI.73.6.3358-3366.2005.

45. Komoriya K, Shibano N, Higano T, Azuma N, Yamaguchi S, Aizawa S-I. 1999. Flagellar proteins and type III-exported virulence factors are the predominant proteins secreted into the culture media of Salmonella Typhimurium. Mol Microbiol 34:767–779. 10.1046/j.1365-2958.1999.01639.x.

46. Köplin R, Brisson J-R, Whitfield C. 1997. UDP-galactofuranose precursor required for formation of the lipopolysaccharide O antigen of Klebsiella pneumoniae serotype O1 is synthesized by the product of the rfbDKPO1 gene. J Biol Chem 272:4121–4128. 10.1074/jbc.272.7.4121.

47. Leekitcharoenphon P, Lukjancenko O, Friis C, Aarestrup F, Ussery D. 2012. Genomic variation in Salmonella enterica core genes for epidemiological typing. BMC Genomics 13:88. 10.1186/1471-2164-13-88.

48. Lienau EK, Blazar JM, Wang C, Brown EW, Stones R, Musser S, Allard MW. 2013. Phylogenomic analysis identifies gene gains that define Salmonella enterica subspecies I. PLoS One 8:e76821. 10.1371/journal.pone.0076821.

ARTICLE 5

Genomic characterization helps dissecting an outbreak of Listeriosis in northern Italy

PLOS | CURRENTS OUTBREAKS 2017. Edition 1

doi: 10.1371/currents.outbreaks.633fd8994e9f06f31b3494567c7e504c

Genomic Characterization Helps Dissecting an Outbreak of Listeriosis in Northern Italy

Francesco Comandatore[a], Marta Corbella[b], Giuseppina Andreoli[c], Erika Scaltriti[d], Massimo Aguzzi[e], Stefano Gaiarsa[b], Bianca Mariani[b], Marina Morganti[d], Claudio Bandi[f], Massimo Fabbi[c], Piero Marone[b], Stefano Pongolini[d], Davide Sassera[g].

Author Affiliations
[a] Pediatric Clinical Research Center, Università degli Studi di Milano, Milan, Italy.
[b] S.C. Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italia.
[c] Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna, Pavia, Italy.
[d] Servizio di Analisi del Rischio, Direzione Sanitaria, Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia-Romagna (IZSLER), Parma, Italy.
[e] Dipartimento di Prevenzione Veterinaria, Agenzia della Salute di Pavia, Pavia, Italy.
[f] Dipartimento di Bioscienze, Università degli Studi di Milano, Milano, Italy.
[g] Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy.

Address correspondence to Davide Sassera, **davide.sassera@unipv.it**.

This article can be found at: **http://currents.plos.org/outbreaks/article/genomic-characterization-helps-dissecting-an-outbreak-of-listeriosis-in-northern-italy/**

Please scan this QR code to access the website from the printed version

FUNDING STATEMENT. The authors received no specific funding for this work.

COMPETING INTERESTS. I, Davide Sassera, on behalf of all the authors of the manuscript 'Genomic characterization helps dissecting an outbreak of Listeriosis in Northern Italy' submitted to Plos Current Outbreaks DECLARE that no competing interests exist.

DATA AVAILABILITY STATEMENT. I, Davide Sassera, on behalf of all the authors of the manuscript 'Genomic characterization helps dissecting an outbreak of Listeriosis in Northern Italy' submitted to Plos Current Outbreaks [OBK-17-0002] declare that the minimal dataset underlying the findings in the manuscript is publicly available to other researchers. The data were submitted to the European nucleotide database (ENA), which is the database of the European Bioinformatics Institute (EBI) which belongs to the European Molecular Biology Laboratory (EMBL). Specifically, the sequencing reads are made available under the accession number PRJEB20581 and the genome assemblies are made available under the accession numbers ERS1607070-ERS1607078.

ABSTRACT

Introduction. *Listeria monocytogenes* (Lm) is a bacterium widely distributed in nature and able to contaminate food processing environments, including those of dairy products. Lm is a primary public health issue, due to the very low infectious dose and the ability to produce severe outcomes, in particular in elderly, newborns, pregnant women and immunocompromised patients.

Methods. In the period between April and July 2015, an increased number of cases of listeriosis was observed in the area of Pavia, Northern Italy. An epidemiological investigation identified a cheesemaking small organic farm as the possible origin of the outbreak. In this work we present the results of the retrospective epidemiological study that we performed using molecular biology and genomic epidemiology methods. The strains sampled from patients and those from the target farm's cheese were analyzed using PFGE and whole genome sequencing (WGS) based methods. The performed WGS based analyses included: a) *in-silico* MLST typing; b) SNPs calling and genetic distance evaluation; c) determination of the resistance and virulence genes profiles; d) SNPs based phylogenetic reconstruction.

Results. Three of the patient strains and all the cheese strains resulted to belong to the same phylogenetic cluster, in Sequence Type 29. A further accurate SNPs analysis revealed that two of the three patient strains and all the cheese strains were highly similar (0.8 SNPs of average distance) and exhibited a higher distance from the third patient isolate (9.4 SNPs of average distance).

Discussion. Despite the global agreement among the results of the PFGE and WGS epidemiological studies, the latter approach agree with epidemiological data in indicating that one the patient strains could have originated from a different source. This result highlights that WGS methods can allow to better.

INTRODUCTION

Listeria monocytogenes (Lm), widely distributed in the environment including soil, plants, and water, is a foodborne bacterial pathogen that can contaminate different kinds of food among which milk and dairy products (1). Lm is capable of adapting to and growing at refrigeration temperatures and, moreover, it can form biofilm to help colonization of surfaces. Consequently, Lm can colonize food processing environments, contaminating the finished products (2). Although Lm is an uncommon cause of illness in the general population, it can represent an important public health problem in case of large scale distribution of contaminated food, due to the very low infectious dose (3,4). Listeriosis is a severe disease and it primarily affects the elderly (5), newborns, pregnant women and immunocompromised patients, categories that can be up to 20 times more susceptible to the disease (6). Clinical manifestations are highly variable and host-dependent: from non-specific and mild symptoms, to febrile gastroenteric syndromes or even cases of seps is and meningitis with mortality rates up to 30% (7).

Most reported cases of listeriosis are sporadic, however, outbreaks have been described with increasing incidence worldwide (5,8,9,10,11). ECDC and EFSA report 2,161 confirmed human cases of listeriosis in the EU Summary report on zoonoses, zoonotic agents and food-borne outbreaks 2014 (12). The EU notification rate was 0.52 cases per 100,000 population which represents a 30% increase compared with 2013 (0.40 cases per 100,000 population). ECDC and EFSA report 210 deaths due to listeriosis in 2014, and a fatality rate above 12.5%. This was the highest number of deaths reported since 2009 (annual average: 163). An average of 131 cases with 0.22 cases per 100,000 population were reported in Italy from 2010 to 2013 (12). A regional study regarding the Lombardia region reported 134 isolates in the 2006-2010 period (6). The notification of listeriosis in humans is mandatory in most countries in Europe, in Italy since 1991, as regulated by Italian D.M. 15/12/1990. Since 2009, was established a digital platform, ENTER-NET Italia system (Enteric Pathogen Network) connected to European ENTER-NET network, dedicated to the assessment of microbiological clusters of food-borne diseases (13).

Pulsed-Field Gel Electrophoresis (PFGE) represents the gold standard for subtyping of Lm and other foodborne pathogens (14), however, studies employing other molecular and genomic methods have proliferated recently, allowing to characterize Lm strains not only by pulsotype, but also by multilocus sequence type (MLST) and core genome MLST (cgMLST) (15,16,17,18).

In this study we describe an outbreak of *Listeria monocytogenes* occurred in 2015 in Northern Italy, using a combination of molecular biology and genomics techniques. Despite the results obtained from the two approaches resulted coherent at large scale, the Whole Genome Sequencing (WGS) approach resulted more accurate in the discrimination of the strains involved in the outbreak.

THE OUTBREAK

Between 28th April and 11th July 2015 six patients showing symptoms compatible with Listeriosis (sudden onset of fever, chills, severe headache, vomiting, and other influenza-like symptoms) were admitted to hospitals in Pavia province of Lombardia region, Northern Italy. The first three cases were observed at the Fondazione IRCCS Policlinico San Matteo Hospital in Pavia (Italy), the fourth an the Ospedale SS Annunziata di Varzi, the fifth an the Ospedale unificato di Broni-Stradella and the sixt at the Ospedale Civile di Voghera. We will refer to the patients enumerating them chronologically from 1 to 6 (Table 1).

Patient 1, upon admission, informed the hospital personnel of having recently consumed goat cheese produced by a small organic farm. The patient provided the leftover cheese (~ 20g), that was tested and Lm was not detected. It must be noted that the small amount of available cheese could have influenced the sensitivity of the test, as the standard protocol indicates 25g as the correct amount for the analysis.

During the following two weeks, two apparently unrelated listeriosis cases were observed at Pavia hospital, involving a drug-abusing subject (patient 2) and a 1.5 year old child (patient 3). The child's parents informed the hospital personnel that the child had recently consumed home-made cheese, and provided a cheese sample (>25g) which was tested, and no Lm was detected. During the previous three years an average of 3.3 cases per year were observed in Pavia province, thus, three cases in 16 days were considered a possible outbreak, and an epidemiological investigation was performed.

The farm that produced the raw-milk goat cheese eaten by patient 1 was subjected to a first inspection during which goat cheese and raw milk, as well as the food contact surfaces in the processing plant19 were sampled. Lm was isolated from two samples from a single cheese shape (Cheese_1 and Cheese_2 in Table 1), while one of three samples collected from the plant (wood ripening surface) was found to be positive by PCR. In order to monitor the possible persistence of Lm contamination three additional sampling were performed during the two months after the first inspection. One wood ripening surface resulted positive by PCR in June, and Lm was isolated from a cheese shape in July. (Cheese_3 in Table 1). All samples collected in the farm are showed in Table 2.

Sample name	Geographic origin	Sampling date	Source	Age	Gender	Risk factor	Diagnosis and symptoms	Outcome
Patient_1	Pavia	28/04/15	blood and CSF	71	M	no one	Meningitis and sepsis	cured
Patient_2	Pavia	29/04/15	blood	47	M	HIV and HCV infection, hepatic impairment	HIV, HCV and HBV co-infection, jaudice, sepsis	cured
Patient_3	Pavia	14/05/15	CSF	1.5	M	< 2 year aged	Diarrhea, fever and confusion	hydrocephalus
Patient_4	Varzi	21/05/15	blood	78	M	no one	Sepsis	cured
Patient_5	Broni - Stradella	11/07/15	blood	61	F	cancer	Sepsis	death
Patient_6	Voghera	28/07/15	blood	70	M	Parkinson disease	Diarrhea, Meningitis	cured
Cheese_1	Target farm	26/05/15	Cheese (inner)	na	na	na	na	na
Cheese_2	Target farm	26/05/15	Cheese (rind)	na	na	na	na	na
Cheese_3	Target farm	17/07/15	Cheese (rind)	na	na	na	na	na

Table 1. Characteristics of the patients and of the cheese samples where *Listeria monocytogenes* was detected.

Collection date	Cheese samples	Milk samples	Food-contact surface samples	Non food-contact surface samples
22/05/15	2/2/2		3/1/0	
15/05/15		1/0/0		
19/06/15	2/0/0		6/01/0	4/0/0
17/07/15	4/1/1		5/0/0	2/0/0
23/09/15			7/0/0	8/0/0

Table 2. Samples from the putative target farm. Numbers indicate collected samples / PCR positive samples / isolation positive samples

During the epidemiological investigation three other cases of listeriosis were diagnosed in the province of Pavia. None of the three patients reported to have eaten organic cheese, or other useful information for investigators to trace the origin of the infections. Four patients out of the six completely recovered, while patient 3 developed a hydrocephalus and patient 5 died. All the isolates were subjected to molecular characterization and, subsequently, a retrospective genomic investigation was performed.

METHODS

Ethics statement. The study was designed and conducted in accordance with the Helsinki declaration. This study was performed according to the guidelines of the IRCCS Foundation Policlinico San Matteo Hospital in Pavia Institutional Review Board of on the use of biological specimens for scientific purposes in keeping with Italian law (art.13 D.Lgs 196/2003). The work described here is a retrospective study performed on bacterial isolates from human samples that were obtained as part of hospital routine. No extra human samples were obtained for this research. Therefore, informed consent (either written or verbal) was not required.

Strain isolation. Blood and Cerebrospinal Fluid (CSF) samples obtained from the six patients were inoculated in aerobic or pediatric broth, and incubated in BACTEC FX (Becton Dickinson, Heidelberg, Germany). Positive broths from blood and CSF were analyzed by Gram staining method and culture, matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) MicroflexTM LT (Bruker Daltonik GmbH, Bremen, Germany) was used for species identification through the Bruker biotyper 3.1 database. Antibiotic susceptibility tests of each isolate was performed via standard disk diffusion on Mueller-Hinton agar incubated at 37°C for 24 h using the Kirby-Bauer method (20). The results were interpreted with standardized criteria from breakpoint committee EUCAST (21). All isolates were then stocked at -80° C. The cheese samples provided by patient 1 and 3 were subjected in parallel to molecular diagnosis and isolation protocols, respectively PCR Real Time – iQCheckTM *L. monocytogenes* II kit (BIORAD) AFNOR BRD 07/10 – 04/05 and ISO 11290-1:116/Amd 1:2004 (22). eight cheese, one raw milk and 35 environmental samples were collected from the putative origin dairy processing plant. All these samples were subjected to the above PCR method and PCR positive samples were subjected to standard Lm isolation protocol, according to ISO 11290-1:1996/Amd 1:2004 (22**).**

Molecular characterization. All isolates were subjected to DNA extraction using the Qiagen DNeasy kit according to manufacturer's instructions, and to Lm specific PCR using an accredited protocol (PCR real-time – iQ-CheckTM *L. monocytogenes* II kit (BIO-RAD) AFNOR BRD 07/10 – 04/05). All isolates were genotyped by PFGE according to the Pulsenet protocol (23). Genomic DNA underwent restriction with *AscI* and *ApaI* enzymes before electrophoresis in a CHEF Mapper® XA System (Bio-Rad, California, USA). PFGE patterns were analyzed using Bionumerics Software ver. 7.0 (Applied-Maths, Sint-Martens-Latem, Belgium) and associated to strain information in our surveillance database. Clustering of the PFGE profiles was generated using the Unweighted Paired Group Method with arithmetic averages (UPGMA) based on the Dice Similarity Index (Optimization=1% and Band Matching Tolerance=1%). Following comparison of the electrophoretic profiles, a PFGE pattern (pulsotype) was assigned to each isolate within the database of the laboratory of the Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (Sezione Diagnostica di Parma). Two isolates were indicated as belonging to the same pulsotype if the band pattern differed by less than two bands.

Genomics. Whole-genome DNA was extracted from each isolate using a QIAamp DNA minikit (Qiagen) following the manufacturer's instructions, and sequenced using an Illumina Miseq platform with a 2 by 250 paired-end run after Nextera XT paired-end library preparation. Genome assemblies were obtained using Mira software (24) and subjected to open reading frame (ORF) calling using Prodigal (25). The MLST profiles of the sequenced strains were determined in silico (using an in-house Perl script), on the basis of the MLST profiles defined in the Institut Pasteur MLST database (http://bigsdb.pasteur.fr/listeria/ (26)). The 713 *Lm* genomes available in the Patric database (in date 11th July 2016) were retrieved and subjected to in silico MLST profile determination and the genomes belonging to the same clonal complexes of our strains were selected for further analyses. The selected strains were subjected to core genome SNP-based phylogeny. The analysis was performed on a robust dataset of core genes selected from the cgMLST1748 genes (27). The cgMLST1748 genes were extracted from BIGSdb-Lm platform and searched, using Blastn, in the genomes of the selected strains. Bidirectional Best Hit (BBH) method (28) was then used to group the genes into clusters of orthologous genes. For each cgMLST1748-ortholog gene present in single copy in all the genomes, the sequences of all isolates were retrieved, aligned and translated using in-house Perl scripts and Muscle software (29). The cgMLST gene alignments were then screened and the genes with the following features were selected: a) all the aligned sequences begin with a start codon; b) all the aligned sequences finish with a stop codon; c) all the aligned sequences have a single stop codon; d) for each aligned sequence, the gaps cover less than 10% of the alignment length. The nucleotide alignments of the selected cgMLST1748 orthologs were then concatenated and subjected to phylogenetic analysis using Maximum Likelihood approach, with RaxML 8 software (30), setting GTRGAMMA model and 100 pseudo bootstrap replicates.

Four databases of Lm sequences, namely virulence genes, antibiotic resistance genes, genes for resistance to Benzalkonium and genes for resistance to metals and detergents were retrieved from the ListeriaMLST database. For each genome sequenced in this study, the obtained paired-end reads were aligned against the curated genes databases using Bowtie2. Genes with >10X coverage for >95% of the sequence length were considered as present in the isolate.The genomes were searched for presence of phages using Phast (31).

RESULTS

Case characteristics. The median age of the six patients involved in the study was 54 years (range 1-78), five out of the six were males and all of them lived in the Pavia province. In four out of six patients Lm was isolated from blood cultures, in one patient from cerebrospinal fluid (CSF) and in another one from both blood and CSF. For full details on patients and symptoms see Table 1. An epidemiological investigation identified the cheesemaking small organic farm that possibly originated the outbreak, where sampling of milk, cheese and food processing environment was performed. Lm isolation was achieved from two cheese shapes, in the first case from both crust and paste, in the second case from the crust only. PCR positivity was obtained for 2 farm environment samples.

Isolate characterization. All isolates were susceptible to ampicillin, erythromicin, meropenem, cotrimoxazole, penicillin. The result of the clustering analysis based on the PFGE patterns obtained with *ApaI* and *AscI* enzymes resulted congruent, grouping the isolates collected from patients 1, 2, 4, together with those from the three cheese samples in both analyses, indicating a clear relationship between the six. Isolates from the three other patients showed a clearly different PFGE pattern, excluding their belonging to the outbreak (Figure 1).

Whole genome sequencing analysis. Whole genome sequencing was performed for the nine strains, six from patients and three from cheese. Genome assemblies, submitted to the EMBL-EBI database, resulted to be on average of high quality (Table 3). In-silico MLST was performed on the genome assemblies, revealing that the three isolates obtained from the cheese samples and three of the six isolates from patients belong to sequence type 29 (ST29), and the remaining three isolates belong to ST1, ST7 and the ST398 (see Table 3 for genome statistics and STs).

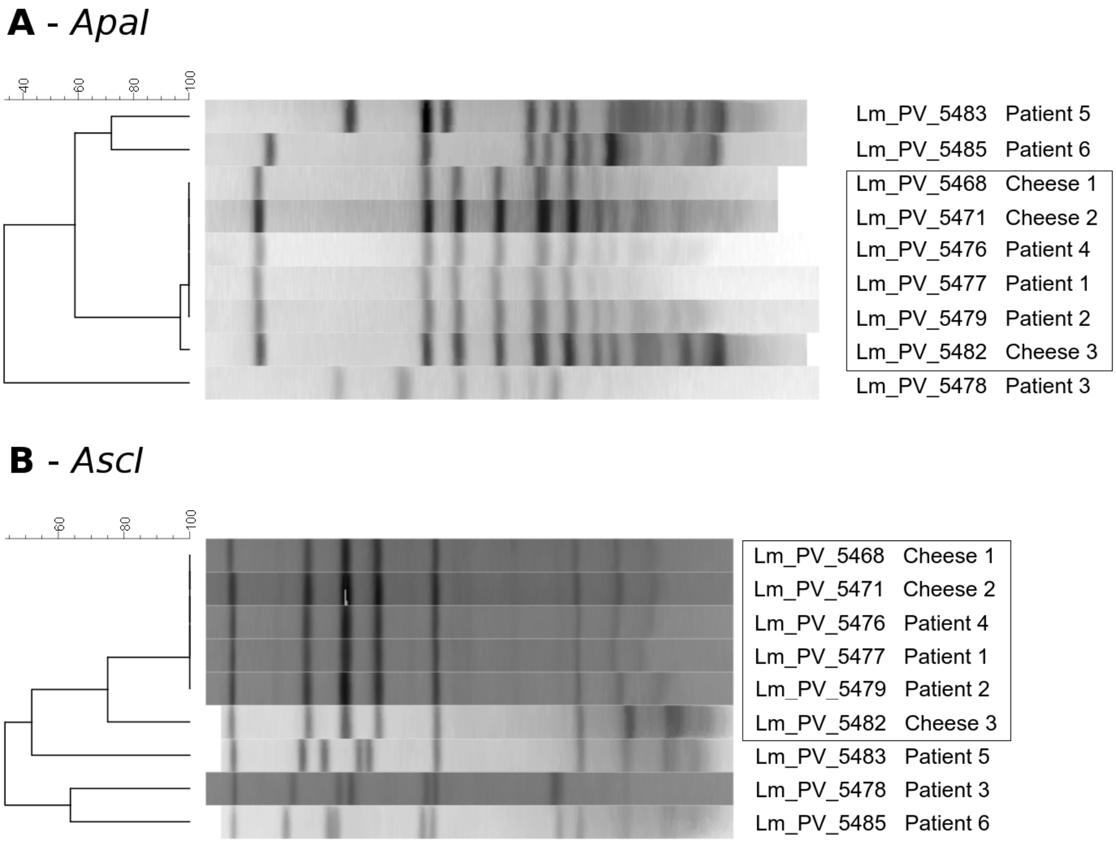


Fig. 1: PFGE profiles. Pulse Field Gel Electrophoresis analysis of the nine strains, obtained using the AscI and ApaI restriction enzymes.

Sample name	Strain	Assembly length	Contig number	Contig average length	N50	Sequence Type	Clonal Complex	Lineage	Genome Accession Number
Patient_1	Lm_PV_5477	2905832	247	11764.50	21441	29	CC29	II	ERS1607073
Patient_2	Lm_PV_5479	2890180	297	9731.25	16094	29	CC29	II	ERS1607075
Patient_3	Lm_PV_5478	2899148	253	11459.08	18959	1	CC1	I	ERS1607074
Patient_4	Lm_PV_5476	2876997	505	5697.02	8596	29	CC29	II	ERS1607072
Patient_5	Lm_PV_5483	2819177	215	13112.45	22547	398	CC398	II	ERS1607077
Patient_6	Lm_PV_5485	2948384	57	51726.04	223238	7	CC7	II	ERS1607078
Cheese_1	Lm_PV_5468	2927109	40	73177.73	246805	29	CC29	II	ERS1607070
Cheese_2	Lm_PV_5471	2930407	118	24833.96	50022	29	CC29	II	ERS1607071
Cheese_3	Lm_PV_5482	2922490	102	28651.86	57506	29	CC29	II	ERS1607076

Table 3. Statistics of the genomes assemblies obtained from nine Listeria monocytogenes strains and MLST profiles.

The 713 Lm genomes present in the Patric database were retrieved, in-silico MLST typed, and the 81 genomes belonging to the clonal complexes of the study strains (i.e. CC1, CC7, CC29 or CC398) were selected. A cgMLST-based phylogenetic reconstruction was performed using a subset of the cgMLST1748 scheme genes, including only the 928 genes present in single copy in all the strains and giving a good quality alignment. The cgMLST-based phylogeny shows that the isolates from patient 1, patient 2, patient 4, cheese 1, cheese 2 and cheese 3 are tightly related (Figure 2), while the other isolates are scattered on the tree. The six closely related strains were then investigated more in depth in order to reconstruct the outbreak structure, using whole genome sequencing (WGS) typing, and data from the epidemiological investigation. In particular, the following evidence was considered:

a) Single nucleotide polymorphism (SNP) distance revealed that the strains from patient 1, patient 2, cheese 1, cheese 2 and cheese 3 differ by 0.8 SNPs on average (values ranging from 0 to 2), while their average distance from patient 4 isolate is ten times higher, at 9.4 SNPs (values ranging from 9 to 10) (see Figure 3).

b) Patient 1 reported to have eaten the cheese produced by the suspect farm, while patient 4, referred to have never bought cheese from the farm. No information on whether patient 2 ate the cheese became available. A potential, albeit unlikely, link would be that patient 4 could have eaten foods prepared with raw materials in common with the contaminated cheese, such as salt solution.

The combination of the higher SNP distance, and the absence of an epidemiological link, led us to consider patient 4 as not associated to the outbreak.

Regarding the presence of resistance genes, all the strains collected in this study showed the same profile of antibiotic resistance genes, harboring the *fos*X, *lmo*1708, *nor*B, and sul genes. This genetic uniformity is in accordance with the results obtained in the antibiograms, which were identical for all strains. Conversely, the virulence genes profiles resulted less conserved among the lineages: the isolates belonging to ST29 (collected from patient 1, patient 2, patient 4, cheese 1, cheese 2, cheese 3 samples) and ST7 (patient 6) presented the same virulence gene profile, while the isolate from patient 5 (ST398) also possessed the vip gene. The isolate from patient 3 (ST1) had multiple additional virulence genes: *aut* IVb, *glt*A, *glt*B, *mdr*M, *vip*and genes of the cluster LIPI-3 (Figure 4). This gene cluster have been reported in the literature to be one of the three major virulence factors (LIPI-1, LIPI-2, LIPI-3) (3). Seven phages were detected, showing an identical pattern of presence/absence in all the strains belonging to ST29. See Figure 4 for a list of the detected phages.

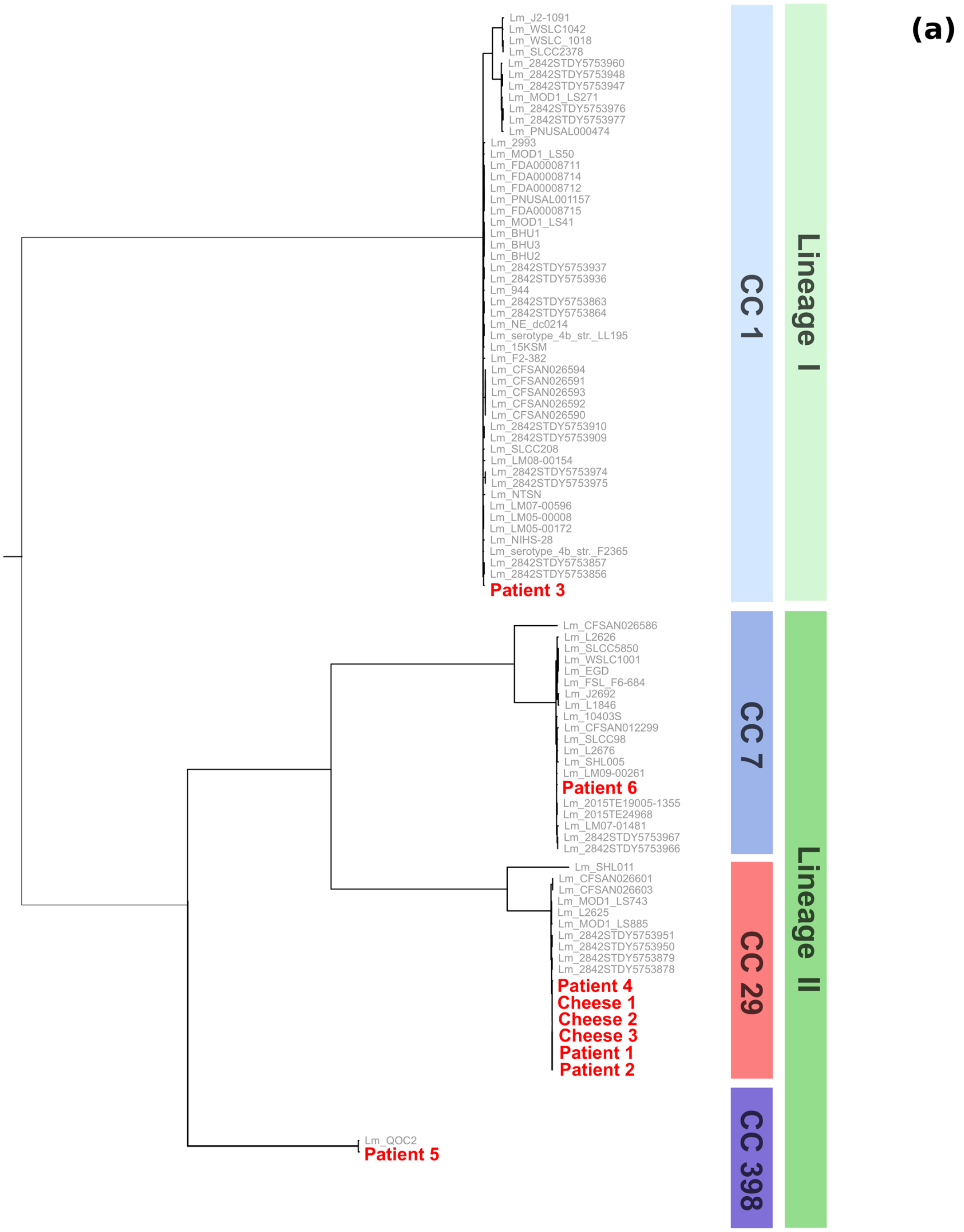


Fig. 2 (a): Phylogeny. (a) Phylogenetic reconstruction of the relationships between the study isolates and database isolates of the corresponding clonal complexes. Tree obtained using Maximum Likelihood approach, with RAxML 8 software, setting GTRGAMMA model and 100 pseudo bootstrap replicates on an alignment of 928 conserved core genome MLST genes. (b) Sub-tree including only the CC29 strains.

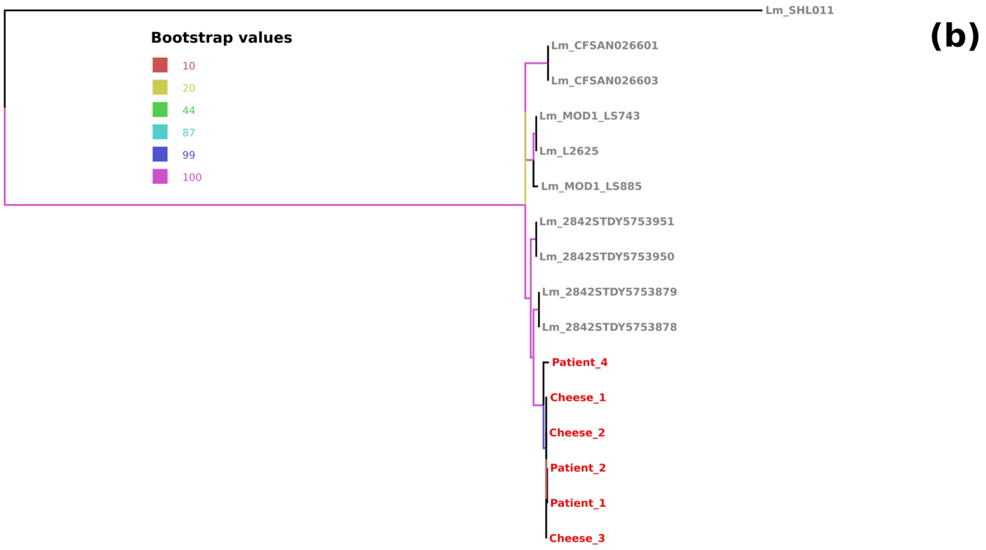


Fig. 2 (b): Phylogeny. (a) Phylogenetic reconstruction of the relationships between the study isolates and database isolates of the corresponding clonal complexes. Tree obtained using Maximum Likelihood approach, with RAxML 8 software, setting GTRGAMMA model and 100 pseudo bootstrap replicates on an alignment of 928 conserved core genome MLST genes. (b) Sub-tree including only the CC29 strains.

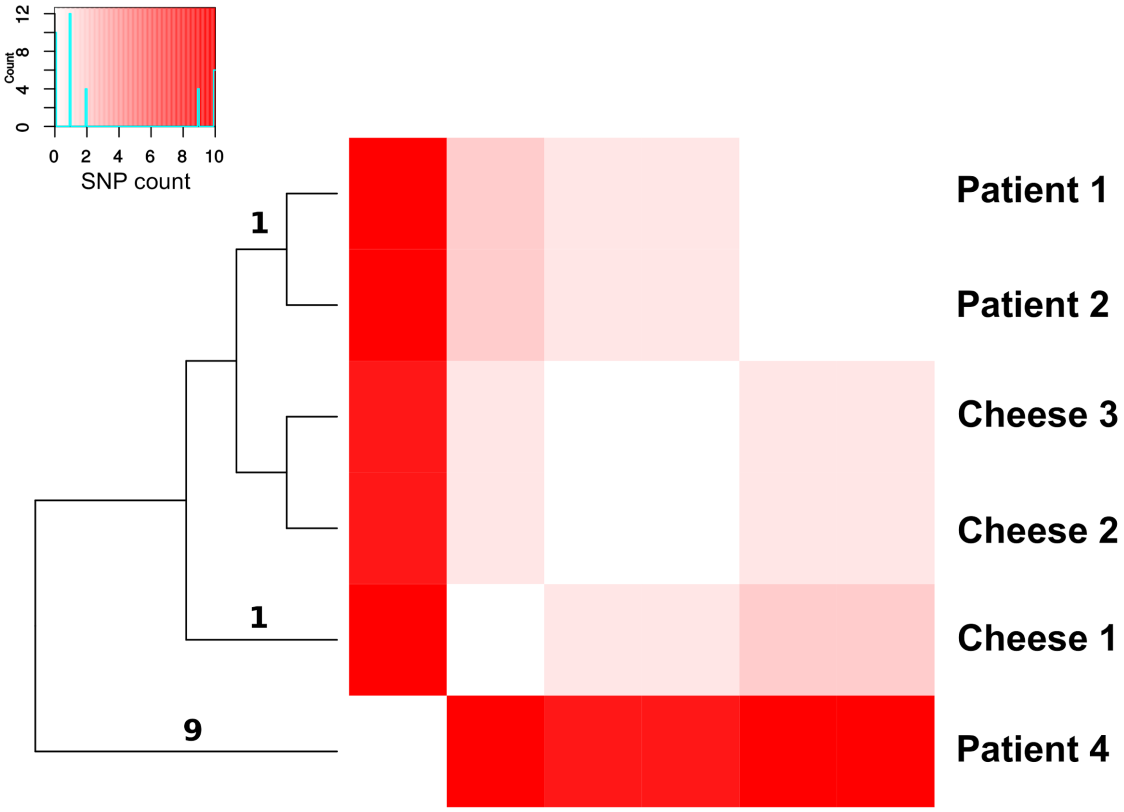


Fig. 3: Heatmap of the SNP distances. Heatmap showing the single nucleotide polymorphism between the isolates obtained in this study. Bright red corresponds to the highest number of SNPs. The number of SNPs supporting tree branched are reported on the relative branch

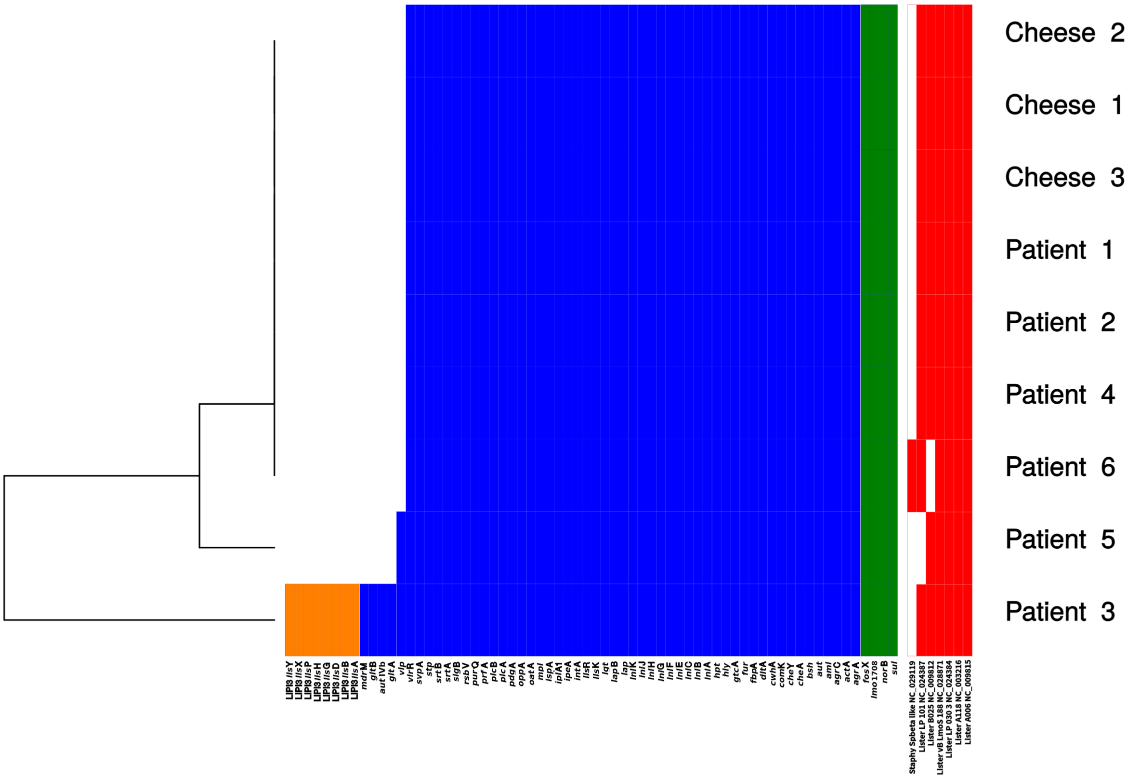


Fig. 4: Presence/absence of resistance genes, virulence genes and phages. Profiles of presence of genes of interest, including genes for antibiotic resistance and virulence, and phages for each genome. LIPI genes are reported in orange, virulence genes in blue, resistance genes in green and phages in red.

DISCUSSION

Six cases of listeriosis occurred between 28th April 2015 and 28th July 2015 in four hospitals of the province of Pavia, Northern Italy. This represented an important increase of the incidence in the area, from an average of 0.28 per month in the three previous years to 2 per month in the examined period. This suggested that an *Lm* strain could be emerging in the area, and an epidemiological investigation was performed. In particular, a first investigation was carried out using molecular techniques and patient interviews, and, after one year, a WGS investigation followed. The results of the two reconstructions were then compared.

PFGE clustered together the strains from three patients (patient 1, 2 and 4) and from all the cheese samples collected from the farm identified as the outbreak origin, indicating that the outbreaking strain originated from that farm and then infected the three patients. Patients 3, 5, and 6 resulted to be unrelated to the outbreak. Additionally, the farm environment was *Lm* positive by PCR, prompting the owner to refurbish the structure. The main inconsistency of this reconstruction was patient 4, who declared with confidence to have never eaten the cheese produced at the farm, while he stated to have consumed raw meat, but could not indicate the origin.

The results of the retrospective WGS investigation allowed to better investigate this point. The core genes SNPcgMLST phylogenetic reconstruction clustered the strains from patient 1, 2, 4 and the cheese strains together, in accordance with the PFGE clustering. We then calculated the number of SNPs between each pair of strains of the PFGE cluster. This analysis showed that the isolate from patient 4 presents an average SNP distance ten times higher than the average distance within the cluster (Figure 3). This pattern suggests that patient 4 could had been not part of the outbreak but, having a sole outlier strain, it was not possible to statistically test this hypothesis. Despite this, epidemiological data resulted coherent to the scenario we inferred from WGS data: since these six strains were collected in a span of three months, and the isolate from patient 4 was obtained in the middle of this period, such difference is unlikely to have arisen from multiple mutations of the isolate from patient 4.

On the basis of the collected data we propose the following epidemiological scenario: patient 1 and 2 were infected by the cheese from the target farm, while patient 4 acquired the bacterium from an unidentified source. Furthermore, we suggest that a WGS-based surveillance program could allow to detect this unidentified source, and solve similar cases in the future.

In summary, WGS allowed to characterize the six human isolates of Lm showing that they represent five different clonal clades that circulate in the studied area, all belonging to STs that were previously reported in the region (6). ST 29 is commonly described in *Lm* outbreaks in USA and Europe, and it was previously described as capable of causing invasive illness (32,33). Two closely related clones, both belonging to ST29, were discriminated through genomics leading to accurate assignment of cases to the outbreak source. The close relatedness of the two clones in absence of a demonstrated epidemiological link opens a question about their possible common ancestry and its associated shared environmental niche. Prospective genomic epidemiology investigations focused on ST29 in the area could allow to understand whether these clones are still circulating in the human population and potentially find clues about their environmental niche.

REFERENCES

1. Stessl B, Fricker M, Fox E, Karpiskova R, Demnerova K, Jordan K, Ehling-Schulz M, Wagner M. 2014. Collaborative survey on the colonization of different types of cheese-processing facilities with Listeria monocytogenes. Foodborne Pathog Dis. 11(1):8-14. PubMed PMID:24138033.

2. Ferreira V, Barbosa J, Stasiewicz M, Vongkamjan K, Moreno Switt A, Hogg T, Gibbs P, Teixeira P, Wiedmann M. 2011. Diverse geno- and phenotypes of persistent Listeria monocytogenes isolates from fermented meat sausage production facilities in Portugal. Appl Environ Microbiol. 77(8):2701-15. PubMed PMID:21378045.

3. Allerberger F, Bagó Z, Huhulescu S, Pietzka A. 2015. Listeriosis: The Dark Side of Refrigeration and Ensiling. In: Zoonoses–Infections Affecting Humans and Animals Springer Science+Business Media Dordrecht 2015 A. Sing (ed.). 249-286

4. Pouillot R, Hoelzer K, Chen Y, Dennis SB. 2015. Listeria monocytogenes dose response revisited--incorporating adjustments for variability in strain virulence and host susceptibility. Risk Anal. 35(1):90-108. PubMed PMID:24975545.

5. Muñoz P, Rojas L, Bunsow E, Saez E, Sánchez-Cambronero L, Alcalá L et al. 2012. Listeriosis: an emerging public health problem especially among the elderly. J Infect. 64:19–33.

6. Mammina C, Parisi A, Guaita A, Aleo A, Bonura C, Nastasi A et al. 2013. Enhanced surveillance of invasive listeriosis in the Lombardy region, Italy, in the years 2006-2010 reveals major clones and an increase in serotype 1/2a. BMC Infect Dis. 13:152.

7. Thønnings S, Knudsen JD, Schønheyder HC, Søgaard M, Arpi M, Gradel KO et al. 2016. Antibiotic treatment and mortality in patients with Listeria monocytogenes meningitis or bacteraemia. Clin Microbiol Infect. 22(8):725-30.

8. Cartwright EJ, Jackson KA, Johnson SD, Graves LM, Silk BJ, Mahon BE. 2013. Listeriosis outbreaks and associated food vehicles, United States, 1998–2008. Emerg Infect Dis. 19(1): 1–9.

9. Doorduyn Y, de Jager CM, van der Zwaluw WK, Wannet WJ, van der Ende A, Spanjaard L et al. 2006. First results of the active surveillance of Listeria monocytogenes infections in the Netherlands reveal higher than expected incidence. Euro Surveill. 11(4) E060420.4.

10. Goulet V, Hedberg C, Le Monnier A, de Valk H. 2008. Increasing incidence of listeriosis in France and other European countries. Emerg Infect Dis. 14:734–740.

11. Koch J, Stark K. 2006. Significant increase of listeriosis in Germany. Epidemiological patterns 2001-2005. Euro Surveill.11:85–88.

12. EFSA (European Food Safety Authority) and ECDC (European Centre for Disease Prevention and Control). 2015. The European Union Summary Report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014. EFSA J. 13 (12): 74-84.

13. Pontello M, Guaita A, Sala G, Cipolla M, Gattuso A, Sonnessa M, Gianfranceschi MV. 2012. Listeria monocytogenes serotypes in human infections (Italy, 2000-2010). Ann Ist Super Sanita. 48(2):146-50. PubMed PMID:22751557.

14. Allerberger F. 2012. Molecular typing in public health laboratories: from an academic indulgence to an infection control imperative. J Prev Med Public Health. 45:1-7.

15. Ebner R, Stephan R, Althaus D, Brisse S. 2015. Phenotypic and genotypic characteristics of Listeria monocytogenes strains isolated during 2011e2014 from different food matrices in Switzerland. Food Control. 57: 321-326.

16. Kvistholm JA, Nielsen EM, Björkman JT, Jensen T, Müller L, Persson S et al. 2016. Whole-genome Sequencing Used to Investigate a Nationwide Outbreak of Listeriosis Caused by Ready-to-eat Delicatessen Meat, Denmark, 2014. Clin Infect Dis. 63(1): 64-70.

17. Schmid D, Allerberger F, Huhulescu S, Pietzka A, Amar C, Kleta S et al. 2014. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013. Clin Microbiol Infect. 20(5): 431-6.

18. A Lomonaco S, Nucera D, Filipello V. 2015. The evolution and epidemiology of Listeria monocytogenes in Europe and the United States. Infection, Genetics and Evolution. 35: 172-183.

19. FSIS Compliance Guideline: Controlling Listeria monocytogenes in Post-lethality Exposed Ready-to-Eat Meat and Poultry Products –January 2014.

20. Bauer A, Kirby W, Sherris JC, Turck M. 1966. Antibiotic susceptibility testing by a standardized single disk method. Am J Clin Pathol. 45: 493-6.

21. European Committee on Antimicrobial Susceptibility Testing Breakpoint tables for interpretation of MICs and zone diameters Version 5.0, valid from 2015-01-01.

22. ISO 11290-1 1996/Amd. 1:2004 Microbiology of Food and Animal Feeding Stuffs — Horizontal Method for the Detection and Enumeration of Listeria monocytogenes. Part.1).

23. Centers of Disease Control and Prevention. 2013. Standard Operating procedure for PulseNet PFGE of Listeria monocytogenes. Centers for Disease Control and prevention, Atlanta, GA. http://www.cdc.gov/pulsenet/PDF/listeria-pfge-protocol-508c.pdf

24. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. Proceedings of the 1999 German Conference on Bioinformatics.

25. Hyatt D, G Chen G, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification BMC Bioinformatics. 8: 11-119.

26. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, Brisse S. 2008. A new perspective on Listeria monocytogenes evolution. PLoS Pathog. 4(9):e1000146. PubMed PMID:18773117.

27. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Björkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EP, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. Nat Microbiol. 2:16185. PubMed PMID:27723724.

28. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96(6): 2896-901.

29. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput Nucl. Acids Res. 32 (5): 1792-7.

30. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312-3.

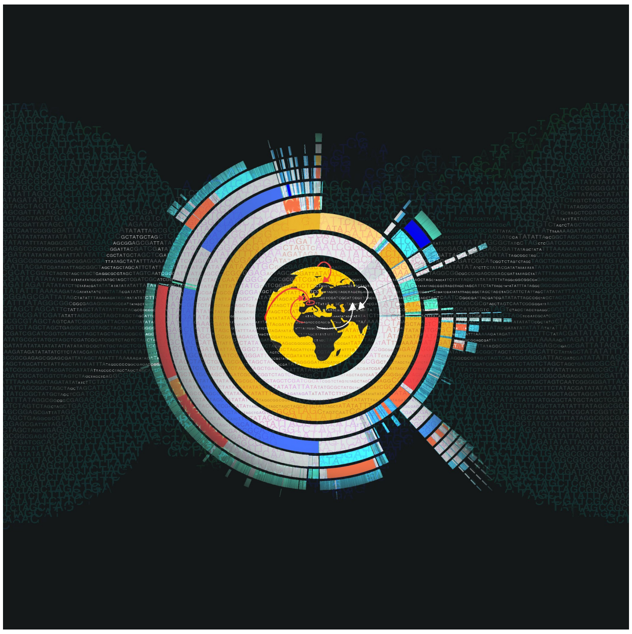
31. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: a fast phage search tool. Nucleic Acids Res. 39(Web Server issue):W347-52. PubMed PMID:21672955.

32. Aureli P, Fiorucci GC, Caroli D, Marchiaro G, Novara O, Leone L, Salmaso S. 2000. An outbreak of febrile gastroenteritis associated with corn contaminated by Listeria monocytogenes. N Engl J Med. 342(17):1236-41. PubMed PMID:10781619.

33. den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M. 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evol Biol. 8:277. PubMed PMID:18842152.

ARTICLE 6

Bacterial genomic epidemiology, from local outbreak characterization to species-history reconstruction



Pathogens and Global Health

2015. 109:319-327
doi:10.1080/20477724.2015.1103503

Cover of the volume dedicated to the present manuscript.
Artwork by Iacopo Leardini

Review

Bacterial genomic epidemiology, from local outbreak characterization to species-history reconstruction

Stefano Gaiarsa[a,b], Leone De Marco[c,d], Francesco Comandatore[b,d], Piero Marone[a], Claudio Bandi[b] and Davide Sassera[d].

Author Affiliations
[a] Struttura Complessa di Microbiologia e Virologia, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy
[b] Dipartimento di Scienze Veterinarie e Sanità Pubblica, Università degli Studi di Milano, Milan, Italy,
[c] Scuola di Bioscienza e Medicina Veterinaria, Università di Camerino, Camerino, Italy
[d] Dipartimento di Biologia e Biotecnologie, Università degli Studi di Pavia, Pavia, Italy

Address correspondence to Davide Sassera, **davide.sassera@unipv.it**

Published online: 16 Feb 2016.
The authors declare no conflict of interest

This article can be found at:
http://www.tandfonline.com/doi/full/10.1080/20477724.2015.1103503

Please scan this QR code to access the website from the printed version

Keywords: Genomic epidemiology, Bacteria, Next-generation sequencing, Mycobacterium tuberculosis, Salmonella Typhi, Carbapenem-resistant Enterobacteriaceae, Staphylococcus aureus, Streptococcus pneumoniae

ABSTRACT

Bacteriology has embraced the next-generation sequencing revolution, swiftly moving from the time of single genome sequencing to the age of genomic epidemiology. Hundreds and now even thousands of genomes are being sequenced for single bacterial species, allowing unprecedented levels of resolution and insight in the evolution and epidemic diffusion of the main bacterial pathogens. Here, we present a review of some of the most recent and groundbreaking studies in this field.

INTRODUCTION

It seems that lately all scientific articles presenting results based on next-generation sequencing start with slight variations of the same formula 'In recent years the advent of novel sequencing technologies has revolutionized the field of …'. This uniformity can teach us a couple of lessons. First of all that scientists do not apply their unquestionable creativity to the writing of introductions, but more importantly that maybe we are actually really facing a scientific revolution. These technologies allow to obtain unprecedented levels of resolution and standardization in genomic data at affordable prices and turnaround times.

Many researchers quickly understood the power of these novel sequencing approaches, generating wealth of data for a number of different biological systems, and designing novel methods to exploit these information. Among them, many bacteriologists fully embraced the revolution, understanding that the generation of high numbers of genomes from a single species, strain or clonal group would allow to reconstruct the history of a bacterium in time and space, to trace its movements and relevant evolutionary events, and to understand the success of specific strains. This genomic epidemiology approach has been applied to a plethora of bacteria, first and foremost to pathogens, with the final goal of obtaining novel strategies to effectively deal with diseases. In this review, we present a small collection of some of the most novel and groundbreaking results obtained in this field, to provide the reader with basic knowledge of the new advances and hopefully to inspire others to pursue innovative lines of research.

LOCAL STUDIES

Third-generation sequencing to tackle the plasmid issue. Multidrug-resistant (MDR) bacteria are currently considered a health problem of primary importance both in Europe and USA (1). Enterobacteriaceae are among the most common MDR bacteria involved in nosocomial infection worldwide. These bacteria, *Escherichia coli* and *Klebsiella pneumoniae,* in primis are often commensals in healthy subjects, as part of the gut flora composition. When colonizing subjects with depressed or weakened immune system, they can turn into dangerous pathogens. These bacteria are able to rapidly develop antibiotic resistance, both by acquiring resistant factors from other bacteria and by developing favourable chromosome mutations. Among the most feared are carbapenem-resistant Enterobacteriaceae (CRE), capable of surviving even treatments with the most recently discovered molecules.

Due to antimicrobial resistance and opportunistic pathogenicity, many of the fatal infections caused by CRE occur within the hospital intensive care units, hosting patients in precarious health conditions, where antibiotic pressure is continuous, often resulting in nosocomial outbreaks. Indeed, during 2013, MDR Enterobacteriaceae were responsible for over 9000 nosocomial infections in the USA, causing over 600 deaths (2). Multiple genomic epidemiology studies have been performed on Enterobacteriaceae global evolution, in particular focusing on the most common antibiotic-resistant clones (3–6). Additionally, the application of genomic approaches has allowed to characterize nosocomial outbreaks in detail, identifying dangerous clones and detecting transmission patterns (7–9). These studies used the ‘standard’ next-generation technologies for genome sequencing (Illumina and 454). These now established technologies, sometimes now referred to as ‘second-generation technologies’ (10), output very large sets of short sequences, suitable to obtain draft genome assemblies. Such methods, however, do not allow to fully characterize the genomic structure, that is to completely detect genomic rearrangements and fully reconstruct complete plasmid sequences. In order to overcome these issues, we can turn to the latest advancements in DNA sequencing, the so called third-generation sequencing platforms (10). Among them is the single molecule real time (SMRT) sequencing technology which allows to obtain very long sequences (> 10,000 nt) (11) and thus to perform genome assemblies that easily result in the full reconstruction of ‘closed’ chromosomes and plasmids.

Recently, Conlan and co-workers (12) exploited this technology to bring local genomic epidemiology to a higher level. Twenty strains of CRE were sampled in the NIH Clinical Center in the time span between 2011 and 2013 both from patients and from the hospital environment. Initial sequencing, performed with Illumina or 454 technologies was, as expected, not precise enough to characterize the genomic content on a structural basis. For this reason, genomes were sequenced a second time using SMRT technology and polished on two levels. Illumina MiSeq reads allowed to ensure the highest precision in base content, while OpGen physical maps (13) gave fundamental information on chromosome and plasmid structures. The effort allowed to obtain complete genomes at a very high level of accuracy, which were used together with epidemiological information to reconstruct the transmission routes of the pathogens. In five cases, the aid of complete genomes and plasmids was essential to detect the exact pattern of transmission. Among the 20 isolates analysed, four showed novel genomic features, never described before. A *Klebsiella oxytoca* strain carrying a plasmid with two copies of the *blaKPC* gene, an isolate of *K. pneumoniae* with three plasmids each carrying a copy of the *blaKPC* gene and two isolates with a chromosomal copy of the resistance gene. Is such an high-percentage of novel features (4/20) surprising? Are the bacteria isolated at the NIH Clinical Center so unique? The more parsimonious explanation is actually that these features are more common than we think, and that we had never seen them before because the quality of our sequencing was previously insufficient.

The approach by Conlan and colleagues also allowed to detect the movement of KPC plasmids between isolates of different species through the environment. A p55-like plasmid of *K. pneumoniae* isolated from a patient was the same found in *Citrobacter freundii* and *Enterobacter cloacae* sampled from the sink in the patient’s room, differentiating only for two adaptive changes in the entire sequence. This suggested that the transmission happened in the sink, as a consequence of a transient presence of *K. pneumoniae* in the local biofilm. This finding underlines, once more, the importance of environmental controls in the hospital wards and suggests specific inspections that should be included in the safety protocols. Lastly, thanks to the finesse of the sequencing, it was possible to analyse a case of co-colonization in a 2011 patient who was found to carry both a carbapenem-resistant *K. pneumoniae* and a *E. cloacae.* Since the patient resulted not colonized at admission, and the two pathogens present two different copies of KPC plasmids, the authors concluded that there was no plasmid transfer and that both resistant strains were circulating in the institution prior of the patient’s arrival.

Hubs of gene flow in *Streptococcus pneumoniae.* *Streptococcus pneumoniae* is a gram-positive bacterium that resides, usually as an innocuous commensal, in the nasopharynx of healthy carriers, with prevalence varying from 5 to 90% (14, 15). When *S. pneumoniae* colonizes immunocompromised individuals, children or elderly, it can act as an opportunistic pathogen, leading not only to pneumonia, but also to a variety of other important diseases, such as meningitis and febrile bacteraemia (16). Multiple vaccines are currently available. For example, conjugate vaccine PCV7 was introduced in the United States in 2000, rapidly proving its effectiveness. Indeed, by 2003, the occurrence of infant pneumococcal disease in Massachusetts was 69% lower (17). Nevertheless, *S. pneumoniae* remains a leading infectious disease worldwide, with higher impact in developing countries. Multiple studies have used genomics to study this important pathogen, starting with the first complete genome sequence presented in 2001 (18). Subsequent comparative works allowed to discover that this species exhibits a high frequency of genomic recombination, which in turn favours the rapid acquisition of novel genetic features. This is often followed by rapid diffusion of those characteristics that give a distinct selective advantage, such as traits conferring resistance to antibiotics and vaccines (19–22). Such behaviour was described even within the course of a single-patient chronic paediatric polyclonal infection (23).

Recently a study presented the sequencing of over 3000 genomes of *S. pneumoniae* isolated during the course of 4 years from the population of a single refugee camp in Thailand (24). This is, to our knowledge, the largest study of bacterial genomic epidemiology to date, structured so that it allows an unprecedented level of sampling density, providing novel insight in the genomic evolution of *S. pneumoniae* and of bacterial populations in general. The authors used the Bayesian approach BAPS (25, 26) to investigate the structure of the bacterial population of the refugee camp, identifying 33 clusters, that could be divided in 183 secondary, mostly clonal, subclusters. An analysis focused on the capsule biosynthesis locus and detected a strong presence of non-typeable (NT) isolates, lacking the capsule, but also the impressive number of 191 ‘plausible capsule switching events’, with numerous of these recombinations causing switches between the capsulated and non-capsulated states.

While nucleotide substitution rate did not differ between clusters, rate of recombination was variable. This ratio, intended as the ratio between recombination and mutation events (*r/m*), resulted to be very different among population clusters and significantly higher in NT isolates. The more striking result of this analysis was, however, the consistently higher rate of recombination of six genomic loci, specifically antigens and antibiotic-resistance genes. Among these ‘recombination hotspots’ were genes providing resistance to beta-lactams and cotrimoxazole. Interestingly, the recombination histories of genes providing resistance to the two antibiotics were different, and the authors found them to match the variable use of beta-lactams and cotrimoxazole, respectively, increasing and decreasing during the time of the study within the refugee camp area.

The extreme genome density of this study allowed also to tackle one important issue of recombination studies, the detection of the donor isolates. Indeed, the authors identified 443 blocks that were identical between the recipient and the donor, nine of which were the results of recombination between nine donors and one single recipient. These blocks were not uniformly distributed in the population, and the NT isolates resulted to be not only good recipients, but good donors as well. In summary, this study reported the presence of 'recombination hotspots' that can easily move among isolates, but also the presence of specific lineages that can act as 'hubs of gene flow'. Since these lineages are not necessarily those that present higher virulence, such as NT isolates in *S. pneumoniae*, these results provide novel insights that can help to change the way we look at the dynamics of bacterial populations.

GLOBAL TRENDS

The explosion of *Salmonella* Typhi H58. *Salmonella enterica* is a widely studied pathogenic agent, of such importance that has caused multiple health and economic crises worldwide (27). The three *S. enterica* clusters of greater importance for human health are serovar Typhi (or *S.* Typhi), serovar Typhimurium (or *S.* Typhimurium) and serovar Enteritidis (or *S.* Enteritidis). *S.* Typhi is the causative agent of typhoid fever, a disease particularly diffused in Asia, and endemic in the Indian subcontinent (28). It was estimated that, just during 2010, typhoid fever affected 26.9 million people worldwide, with fatality rates ranging from 1 to 30% (29). *S.* Typhi is able to infect human blood and intestine, producing an array of symptoms including nausea, vomiting, fever and death. A portion of colonized people usually remain asymptomatic for long periods of time (up to years), shedding the bacterium into their stool and urine (27), with the effect of sustaining the pathogen transmission.

Blantyre is a ~1.3 million people district, localized in the south of Malawi. In this area, between 1998 and 2010, several cases of bloodstream infection (BSI) caused by nontyphoidal serovars of *Salmonella* were reported, while the typhoidal *S.* Typhi serovar was rare (30). Starting from 2011, an increase in the number of BSI caused by MDR S. Typhi has been reported. In particular, the *S.* Typhi lineage H58 resulted to be predominant in the infected population. Feasey and colleagues (30) used a genomic epidemiology approach to reconstruct the origin of the emergence of S. Typhi H58 in this area: they collected epidemiological data from the Blantyre district (covering the period 1998–2011) and sequenced the genome of 112 S. Typhi strains, isolated there from 2004 to 2011. Whole-genome analysis was performed using a genome mapping approach, and the obtained SNPs were subjected to maximum-likelihood phylogenetic analysis. Merging the information from the resulting tree and the epidemiological data, the authors were able to reconstruct that the increase of typhoid fever reported from 2011 was due to the rapid diffusion of a monophyletic S. Typhi lineage, the aforementioned H58-haplotype. H58 resulted to be associated with MDR with a much higher frequency (89.3%) than the other circulating S. Typhi types (21.4%).

Strong phylo-geographical clusters were described within the H58 lineage, indicating it to be endemic in the areas included in the study. H58 isolates, collected in the same areas, resulted to be clustered on the phylogenetic tree, independently of the date of isolation. This geographical clusterization is consistent with the existence of reservoirs. Furthermore, genomes of the H58 strain result to be very conserved within the lineages, in contrast with the other S. Typhi monophyla. Indeed, the H58 tree branch lengths are shorter than the other S. Typhi lineages. These data can be explained hypothesizing that a strong purifying selective pressure affects the H58 lineage, and/or that frequent genomic recombinations occurred among the H58 strains.

Due to the undeniable importance of this haplotype, a second study was performed to describe its emergence at the global level. Wong and colleagues (31) considered the impressive collection of 1,832 S. Typhi isolates collected in the period 1905–2014, from 63 countries spanning 6 continents. Whilst the most ancient isolate included in the study was collected in 1905, the first S. Typhi H58 isolate was from 1992, indicating a very recent origin of this lineage. Since 1992, the H58 haplotype represents ~40% of all the isolates collected each year, a remarkable explosive diffusion. Indeed, H58 genomes differ by a mean of only six SNPs, with 93% of them having less than five isolate-specific SNPs. These data show that H58 isolates are very closely related, consistent with the hypothesis of an impressive recent clonal expansion. It must be noted that the number of isolates obtained before 1992 is limited (*n* = 50), and this may skew the perception of the H58 diffusion, nevertheless the result is remarkable. The authors then inferred the date of the H58 origin to be between 1985 and 1992. After 1993, they observed a drastic increase of the H58 effective population size. Furthermore, on the basis of the phylogenetic reconstruction, the authors traced the major geographical transfers of the *S.* Typhi H58 haplotype: the strain originated in India, and through independent events reached Southeast Asia, Fuji, Western Asia, East Africa and Malawi. In Africa, it then invaded Malawi a second time through East Africa and then diffused from Malawi to South Africa.

Novel insights into the genomic evolution of *Staphylococcus aureus*. *Staphylococcus aureus* is among the most important antibiotic-resistant pathogens worldwide. Methicillin-resistant strains (MRSA), in particular, are spread in all continents and can be up to 70% of all *S. aureus* isolates in the most affected countries (32). The first report of MRSA was an hospital-acquired infection in 1960 (33), but the pathogen has since then developed endemic status and can be transmitted outside of the nosocomial environment (the first cases were reported in the mid-1990) (34). The terms HA-MRSA (health care-associated MRSA) and CA-MRSA (community-associated MRSA) reflect this distinction. And if this was not enough, LA-MRSA (livestock-associated) is the zoonotic variant, which is common in farms (35). Resistance to methicillin is encoded in the staphylococcal cassette chromosome *mec* (SCC*mec*) which contains the *mecA* gene. Several variants of the cassette have been discovered and found to be able to transmit and move between strains. The spread of MRSA strains has been the focus of a number of high-profile studies that used genomic approaches to investigate their diffusion and evolution, starting from the pioneering study of Harris (36), already discussed in previous reviews (e.g. (37)).

Holden and colleagues (38) used genomics to reconstruct the evolution of EMRSA-15, a strain belonging to sequence type 22, which is considered the most rapidly spreading and tenacious *S. aureus* in Europe, currently invading other continents. Genomes were obtained from 193 ST22 strains of *S. aureus* isolated from 1990 to 2009 and a SNP-based phylogeny was reconstructed. A molecular clock analysis allowed to distinguish and date different clades with variable virulence levels, corresponding to different stages in the epidemic diffusion. Genomic variability among the clades was analysed both at the SNP level and the gene content level, with the aim of correlating genomic changes with fitness and virulence.

The study concluded that sequence type 22-A (ST22-A) was the first of this lineage to obtain resistance to methicillin, from the primitive community-associated methicillin-sensitive ST22, and dated this event to before 1977. This led to a health care-associated MRSA epidemic that spread in England in the 1980s (ST22-A1). In the mid 1980s, a sublineage acquired resistance to fluoroquinolones, and EMRSA-15 (also called ST22-A2) was born. The authors performed a bayesian analysis and detected a considerable difference in population size between ST22-A1 and ST22-A2, putative consequence of a fitness boost which caused the worldwide diffusion of the latter strain. Lastly, genomic traits were correlated with antimicrobial resistance profiles, thus highlighting the potential of genome sequencing as a diagnostic tool. Appearance on the tree of genetic variants responsible for antimicrobial resistance was found to agree with the variations of antibiotic prescription policy-making in the different regions during the years. Indeed, EMRSA-15 spread through the UK when fluoroquinolones where highly used, while a subsequent spread in Germany was a consequence of the development of yet another resistance, to clindamycin, which was heavily used in that country. Recently, other similar works reconstructed and dated the origin of other MRSA epidemic clones. Genomic variants were mapped on the trees and correlated with phenotypic changes. For example, Planet and coworkers (39) worked on USA300 and

USA300-LV clones, while Stinear and colleagues (40) on CA-MRSA ST93. Stegger *et al. (41)* studied CC80 and Baines *et al.* (42) focused their efforts on HA-MRSA ST239.

In addition to these phylogeny-based works, the genomics of *S. aureus* has been used to investigate variations in genome structure. Indeed recombinations, transmission of plasmids and pathogenicity islands represent big adaptive steps in the history of MRSA, as they do for CRE and *S. pneumoniae*. Recently, Méric and coworkers (43) studied the evolution and genomic flow between two species that share the same niche: *S. aureus* and *Staphylococcus epidermidis.* These species are indeed both common commensals on the human skin and in the nasal pharynx. The authors selected and sequenced 324 isolates from archives and databases, in order to represent the global diversity of the two species, choosing among different genomic variants, location and sources of isolation. Shared genes and alleles were searched between each pair of isolates, and the two species resulted to share a maximum of nine core genome alleles between them, thus suggesting that genomic recombination is very rare between individuals of the two species. On the contrary, mobile elements were highly shared, in particular genes associated with the SaPIn1 pathogenicity island, metal detoxification and the methicillin-resistance island *SCCmec*. The authors use these interesting results to discuss the concept of evolution in relation with the host as a niche, that is two strains or species that share the same host, also share the same selective pressure. Recombination can be driven by direct contact between donor and receiver, but also by evolution of niches, as genomic material can be shared in these enclosed environments.

In this review, we report multiple examples of how exchange of genomic material can involve the accessory parts of the genomes, but also homologous recombinations in core genome loci. The latter is a common thread in global genomic epidemiology as it is commonly found in most bacterial species involved in nosocomial infections. This was recently pointed out by Croucher and Klugman (44) who observed that large recombinations (even bigger than one megabase) seem to be an evolutive weapon that pathogens use to rapidly gain fitness and survive in the hospital environment. The two authors compare recombined bacteria to the 'hopeful monsters' of the Cambrian period, citing the use by Stephen Gould of the term introduced earlier by Richard Goldschmidt(45).

The presence of recombinations should be taken into account because they need to be removed from the genomic alignment in order to obtain resolved and correct phylogenies, which represent the real evolutive history of the pathogen.

HISTORICAL PERSPECTIVES

***Mycobacterium tuberculosis* through history.** *Mycobacterium tuberculosis* is an obligate aerobic pathogenic bacterium and the causative agent of tuberculosis (TB) (46). The presence of mycolitic acids in *M. tuberculosis* coating confers the bacterium resistance to weak disinfectants and dehydration and prevents the effective activity of hydrophobic antibiotics. Additionally, it allows the bacterium to grow inside of macrophages, effectively hiding it from the host's immune system (47). All these characteristics contribute to the ease with which it is transmitted, despite its extremely slow replication time compared to other bacteria. Tuberculosis is a disease that accompanied human populations since antiquity, it has been prevalent worldwide and, if left untreated, causes death in 50% of cases. In the last two centuries, progress has been made both in diagnosis and treatment with the advent of screening programs, antibiotics and vaccines, relegating the emergency to third world countries. Nevertheless, deaths are increasing after an almost 40 years decline (46), and the emergence of multiple antibiotic-resistant strains (48) makes *M. tuberculosis* one of the most important re-emerging bacterial pathogens to date, and the leading bacterial killer worldwide with 1.3 million deaths a year. The so called *M. tuberculosis* Beijing family is a heterogeneous group of strains, among which hypervirulent subtypes stand out, equipped with multiple antibiotic resistances and the ability to cause disease outbreaks (49). The whole family, considered the predominant genotype in East Asia and still currently spreading, can be accounted for more than a quarter of the total tuberculosis cases worldwide. Despite previous epidemiology studies showed high genetic similarity even among strains isolated in different geographic areas (49), pathobiological characteristics appeared heterogeneous (50). The increasing availability of standard genotyping leads to the identification of several Beijing sublineages (51). This approach, however, proved itself limited for fully understanding the diversity of this family due to the insufficient amount of nucleotide variation detected by this technique. Once again, genomic epidemiology can come to our rescue.

Given the relevance of this family for public health globally, several studies focused on reconstructing the origin and spread of *M. tuberculosis* Beijing strains (52–56). Merker *et al.* (57) focused on the biogeographical structure of strains belonging to the Beijing family, with a in-depth analysis on the association between sublineages and antibiotic resistance. The authors assembled a huge data set, comprising almost five thousand genotyped isolates plus 110 whole-genome sequences, the biggest and broadest collection of Beijing family strains to date, both in terms of sheer size and variety of geographical origins. Initially, a minimum-spanning tree (MStree) was constructed using genotyping data, grouping the genomic diversity into 6 major clonal complexes (CCs) and 3 distant branches which were collectively designated as basal sublineage 7 (BL7). CC1 through CC5 were classified as typical/modern Beijing while CC6 and BL7 comprised typical ancestral Beijing variants. The shape of the Mstree and the mean allelic richness confirmed this hypothesis, suggesting that CC1, CC2 and CC5 are in a state of population expansion, while CC6 and BL7 are more ancient and/or in a situation of milder expansion.

The authors then used the information for all the five thousand genotyped isolates to estimate past expansions and time to the most recent common ancestor (TMRCA). CC6 and BL7 were confirmed once again as the oldest sublineages with, respectively, a TMRCA of ~6,000 and 5,000 years, while CC5 resulted the youngest with a TMRCA of ~1500 years. For all sublineages, an estimate of the time elapsed since the beginning of the latest expansion was computed. This analysis showed a much recent timeframe, roughly 200 years ago, for CC1, CC2 and CC5, compared to the estimate detected for the more ancient lineanges, CC6 and BL7, which dated back to the middle age. Once again, genome information, available for 110 isolates, made it possible to obtain a more sensitive estimate of population changes in the recent past, by means of a Bayesian skyline plot. Two significant population growth phases were detected in conjunction with the Industrial Revolution and the First World War. The only decrease in population size was observed contemporary with the spreading of anti-tuberculosis drug usage, while a slight expansion coincides with the advent of the HIV epidemics and the first tuberculosis outbreaks in the former Soviet Union and the United States in the 1990s. A subset of roughly one thousand clinical isolates with known drug resistance profiles was analysed to shed light on possible association between the identified CCs and antibiotic resistance, resulting in CC2 having the highest proportions of MDR strains. It is important to note that CC1, while having a similar proportion of MDR strains to the ancient lineages, showed a high clustering rate (95%) meaning that almost all isolates shared a single resistant haplotype, in contrast with CC6 and BL7 (42% and 57%, respectively). Additionally, strains from CC1 (central Asian outbreak) and CC2 (Russian-European outbreak) showed a higher similarity between them than with other clonal complexes, supporting the MDR outbreak hypothesis and a recent specific expansion of these two clonal complexes.

It is interesting to note that *M. tuberculosis* population growths are closely related to historical events and human migrations. Genotype data allowed to estimate the start of the last expansion events for the recent sublineages to 200–250 years ago, roughly around the time of the industrial revolution and matching known episodes of Chinese immigration towards Pacific islands, Americas and Russia. It is both interesting and not surprising that Beijing family strains experienced increases in population size in conjunction with the

Industrial revolution and with the First World War. This is consistent both with Chinese immigration episodes, as noted above, and with the deprivations caused by war conditions and the co-mortality due to the influenza pandemics of that time (58). The only decrease in population size detected coincides with the advent of antibiotics and mass vaccinations around the 1960s. It has to be kept in mind that the expansion of the two sublineages more associated with antibiotic resistance, CC1 and CC2, predates this event. This indicates that MDR is not the reason of the success of these clonal complexes but just the consequence of public health policies implemented on an already growing bacterial population. Finally, the most recent increasing trend is consistent with the onset of the global HIV epidemic and follows it closely.

The strange case of the amphibious *Mycobacterium*. Albeit recent studies such as the one described above are shedding light in the history of *M. tuberculosis*, there are still plenty of dark patches that need to be illuminated. It is clear that the co-evolution of the bacterium with humans started with the shift from the hunter/gatherer behaviour to the onset of agriculture and animal husbandry, especially cattle (59). Until recent years, the most accredited theory had a zoonotic transfer of *Mycobacterium bovis* following animal domestication during the Neolithic (60). However, recent comparative genomic analyses lean towards the opposite theory. That is, strains adapted to bovines and other animals may have originated from human strains (61, 62). For what concerns the Americas, given that strains currently present in the area are closely related to the European ones, the consensus is that the pathogen was brought by colonizers and settlers after the Columbian discovery (63). This, however, is not consistent with several evidences of skeletal samples dated before 1492 with obvious signs of the disease. If tubercolosis was not brought on caravels, who or what carried it to the New World? And how can we explain the genomic similarity between American and European strains of *M. tuberculosis*? Bos *et al*. (64) try to give us an answer.

Progresses in protocols for isolation of ancient DNA made it possible to collect three *M. tubercolosis* genomes from skeletal samples with signs of tuberculosis infection collected in Peru and dated back to roughly one thousand years ago. They completed a the data set by adding 259 modern *M. tuberculosis* complex (MTBC) genomes, 14 animal isolates and an additional ancient genome originated from an eighteen century Hungarian mummy. An alignment of 22,480 variable positions was the input for a phylogenetic analysis. The resulting tree came with a surprise: the ancient Peruvian samples did not cluster with other human isolates; they were closer to the animal strains, in particular to the modern *Mycobacterium pinnipedii* sample. As the latin suggests, *M. pinnipedii* has been isolated from seals and sea lions. Bayesian dating analysis, using radiocarbon dates as tip calibration, was implemented for dating the most recent common ancestor (MRCA). Using a relaxed molecular clock model, the MRCA was dated between 4000 and 4500 years ago. Since the Bering land bridge closed some 15,000 years ago, 10000 years before the estimated time of the MRCA, the researchers discarded a human migration hypothesis for the appearance of TB in the New World. The remaining, unexpected hypothesis is the amphibious one: seals contracted the disease on the coasts of Africa and carried it to South America where populations living in the seaside contracted by exploiting the marine mammals. This is consistent with similar cases in literature for other pathogens (65). The later eradication of this strain and the almost total substitution with European-like lineages could be accounted by a spread-after-contact of the latter following favourable conditions (66).

CONCLUSIONS: FROM EPIDEMIOLOGY TO DIAGNOSTICS

The examples presented here testify how the use of genome sequencing in bacteriology for epidemiological purposes is now widespread. Favoured by the wealth of data that are being generated and by the continued advances in sequencing technologies, the next step will be to branch into microbiological diagnostics (67–69). Efforts are being made in multiple directions, two of the most promising being direct sequencing from clinical samples and the use of genomic data to predict phenotypic characteristics.

The utility of direct sequencing of clinical samples for diagnosis purposes is obvious, and particularly important for bacteria that are difficult, slow or impossible to culture. A possible approach, when investigating a single pathogen, is the use of specific baits to sequence the target bacterium starting from a mixed clinical sample, an approach that was used for example to detect *M. tubercolosis* from sputum (70). In many clinical cases, it is, however, impossible to know a priori which bacterium is causing a pathological state. In these situations, a whole metagenomic sequencing approach could allow to identify the causative agent/s quickly and without bias. Ad-hoc bioinformatic methods are being developed to tackle the problematic issue of analysing the complex metagenomes that can be obtained from clinical samples, with the goal of correctly sorting and identifying bacterial populations (71, 72)(73)(71, 72).

Some phenotypic characteristics of a pathogen can be readily inferred from its genome sequence. This is the case for example of antibiotic resistance traits that are determined by the acquisition of one single gene. Many other phenotypes are, however, multi-factorial, making such correlations more complex. These difficulties have not discouraged pioneering projects that use genome-wide association studies to link genotype to phenotype, in an effort to obtain diagnostic informations from the now quick and cheap whole-genome sequences. Laabei and colleagues (74) applied such an approach to MRSA, developing a model that can predict with a high degree of accuracy the toxicity of an isolate based on the sequence of signature sites. Another study integrated a genomic approach with gene expression analysis, performed with RNA-seq, to determine antibiotic resistance profiles in *E. coli* (75). These are not the only examples, as other approaches are being tried and validated on multiple pathogens (76, 77).

Multiple issues will need to be addressed to allow the transition from genomic epidemiology to genomic diagnostics in bacteriology, and one that needs the most concerted effort is standardization of genomes, related metadata and bioinformatic analysis. Indeed, the full potentiality of genomics will only be exploited in clinical microbiology when all the wealth of data generated worldwide will be fully compatible, allowing real time evaluation and comparison of the characteristics of sequenced isolates, in a single global database. This will in turn allow to fully correlate genotype with phenotype, to optimize diagnostic and therapeutic approaches and to monitor movement of dangerous strains worldwide. Efforts towards this goal are already being made from the setting of standards for genome qualities(78) to the establishment of platforms using standarized bioinformatic protocols for genome analyses (79, 80).

It is only fitting to conclude this review how it started: Next-generation sequencing is revolutionizing bacteriology… and the best is yet to come.

REFERENCES

1. Nordmann P, Cuzon G, Naas T. 2009. The real threat of Klebsiella pneumoniae carbapenemase-producing bacteria. Lancet Infect Dis 9:228–236.

2. 2014. National Strategy for Combating Antibiotic-resistant Bacteria.

3. Bohlin J, Brynildsrud OB, Sekse C, Snipen L. 2014. An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli. BMC Genomics 15:882.

4. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan M-D, Gomes Moriel D, Peters KM, Davies M, Rogers BA, Dougan G, Rodriguez-Baño J, Pascual A, Pitout JDD, Upton M, Paterson DL, Walsh TR, Schembri MA, Beatson SA. 2014. Global dissemination of a multidrug resistant Escherichia coli clone. Proc Natl Acad Sci U S A 111:5694–5699.

5. Deleo FR, Chen L, Porcella SF, Martens CA, Kobayashi SD, Porter AR, Chavda KD, Jacobs MR, Mathema B, Olsen RJ, Bonomo RA, Musser JM, Kreiswirth BN. 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae. Proc Natl Acad Sci U S A 111:4988–4993.

6. Gaiarsa S, Comandatore F, Gaibani P, Corbella M, Dalla Valle C, Epis S, Scaltriti E, Carretto E, Farina C, Labonia M, Landini MP, Pongolini S, Sambri V, Bandi C, Marone P, Sassera D. 2015. Genomic epidemiology of Klebsiella pneumoniae in Italy and novel insights into the origin and global evolution of its resistance to carbapenem antibiotics. Antimicrob Agents Chemother 59:389–396.

7. Onori R, Gaiarsa S, Comandatore F, Pongolini S, Brisse S, Colombo A, Cassani G, Marone P, Grossi P, Minoja G, Bandi C, Sassera D, Toniolo A. 2015. Tracking Nosocomial Klebsiella pneumoniae Infections and Outbreaks by Whole-Genome Analysis: Small-Scale Italian Scenario within a Single Hospital. J Clin Microbiol 53:2861–2868.

8. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA, Program NCS. 2012. Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella pneumoniae with Whole-Genome Sequencing. Sci Transl Med 4:148ra116–148ra116.

9. Stoesser N, Sheppard AE, Shakya M, Sthapit B, Thorson S, Giess A, Kelly D, Pollard AJ, Peto TEA, Walker AS, Crook DW. 2015. Dynamics of MDR Enterobacter cloacae outbreaks in a neonatal unit in Nepal: insights using wider sampling frames and next-generation sequencing. J Antimicrob Chemother 70:1008–1015.

10. Pareek CS, Smoczynski R, Tretyn A. 2011. Sequencing technologies and genome sequencing. J Appl Genet 52:413–435.

11. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299:682–686.

12. Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, Tsai Y-C, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW, NISC Comparative Sequencing Program, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre JA. 2014. Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. Sci Transl Med 6:254ra126.

13. Schwartz D, Li X, Hernandez L, Ramnarain S, Huff E, Wang Y. 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science 262:110–114.

14. Website. Pneumococcal Disease. Available from: http://www.cdc.gov/vaccines/pubs/pinkbook/downloads/pneumo.pdf

15. Thummeepak R, Leerach N, Kunthalert D, Tangchaisuriya U, Thanwisai A, Sitthisak S. 2015. High prevalence of multi-drug resistant Streptococcus pneumoniae among healthy children in Thailand. J Infect Public Health 8:274–281.

16. Elliott T. 1986. Medical microbiology: Edited by S. BARON. 1986, 2nd ed. Addison-Wesley Publishers Ltd, Wokingham, Berks. Pp. xxvi and 1262. 19.95. J Med Microbiol 22:284–284.

17. Hsu K, Pelton S, Karumuri S, Heisey-Grove D, Klein J, Massachusetts Department of Public Health Epidemiologists. 2005. Population-based surveillance for childhood invasive pneumococcal disease in the era of conjugate vaccine. Pediatr Infect Dis J 24:17–23.

18. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, Durkin AS, Gwinn M, Kolonay JF, Nelson WC, Peterson JD, Umayam LA, White O, Salzberg SL, Lewis MR, Radune D, Holtzapple E, Khouri H, Wolf AM, Utterback TR, Hansen CL, McDonald LA, Feldblyum TV, Angiuoli S, Dickinson T, Hickey EK, Holt IE, Loftus BJ, Yang F, Smith HO, Venter JC, Dougherty BA, Morrison DA, Hollingshead SK, Fraser CM. 2001. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science 293:498–506.

19. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. 2009. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science 324:1454–1457.

20. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. 2011. Rapid pneumococcal evolution in response to clinical interventions. Science 331:430–434.

21. Donkor ES, Bishop CJ, Gould KA, Hinds J, Antonio M, Wren B, Hanage WP. 2011. High levels of recombination among Streptococcus pneumoniae isolates from the Gambia. MBio 2:e00040–11.

22. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656–663.

23. Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, Janto B, Boissy RJ, Hogg J, Barbadora K, Sampath R, Lonergan S, Post JC, Hu FZ, Ehrlich GD. 2010. Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog 6:e1001108.

24. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD. 2014. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet 46:305–309.

25. Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539.

26. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224–1228.

27. Watson CH, Edmunds WJ. 2015. A review of typhoid fever transmission dynamic models and economic evaluations of vaccination. Vaccine 33 Suppl 3:C42–54.

28. Chugh TD, Kothari A, Pruthi A. 2008. The Burden of Enteric Fever. J Infect Dev Ctries 2.

29. Buckle GC, Walker CLF, Black RE. 2012. Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. J Glob Health 2:010401.

30. Feasey NA, Gaskell K, Wong V, Msefula C, Selemani G, Kumwenda S, Allain TJ, Mallewa J, Kennedy N, Bennett A, Nyirongo JO, Nyondo PA, Zulu MD, Parkhill J, Dougan G, Gordon MA, Heyderman RS. 2015. Rapid emergence of multidrug resistant, H58-lineage Salmonella typhi in Blantyre, Malawi. PLoS Negl Trop Dis 9:e0003748.

31. Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley RA, Thomson NR, Keane JA, Weill F-X, Edwards DJ, Hawkey J, Harris SR, Mather AE, Cain AK, Hadfield J, Hart PJ, Thieu NTV, Klemm EJ, Glinos DA, Breiman RF, Watson CH, Kariuki S, Gordon MA, Heyderman RS, Okoro C, Jacobs J, Lunguya O, Edmunds WJ, Msefula C, Chabalgoity JA, Kama M, Jenkins K, Dutta S, Marks F, Campos J, Thompson C, Obaro S, MacLennan CA, Dolecek C, Keddy KH, Smith AM, Parry CM, Karkey A, Mulholland EK, Campbell JI, Dongol S, Basnyat B, Dufour M, Bandaranayake D, Naseri TT, Singh SP, Hatta M, Newton P, Onsare RS, Isaia L, Dance D, Davong V, Thwaites G, Wijedoru L, Crump JA, De Pinna E, Nair S, Nilles EJ, Thanh DP, Turner P, Soeng S, Valcanis M, Powling J, Dimovski K, Hogg G, Farrar J, Holt KE, Dougan G. 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. Nat Genet 47:632–639.

32. Website. MRSA infection rates by country. Available from: www.cddep.org/tool/mrsa_infection_rates_country.

33. Jevons M. 1963. METHICILLIN RESISTANCE IN STAPHYLOCOCCI. Lancet 281:904–907.

34. David MZ, Daum RS. 2010. Community-associated methicillin-resistant Staphylococcus aureus: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev 23:616–687.

35. Hetem DJ, Bootsma MCJ, Troelstra A, Bonten MJM. 2013. Transmissibility of livestock-associated methicillin-resistant Staphylococcus aureus. Emerg Infect Dis 19:1797–1802.

36. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474.

37. Parkhill J, Wren BW. 2011. Bacterial epidemiology and biology--lessons from genome sequencing. Genome Biol 12:230.

38. Holden MTG, Hsu L-Y, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlickova H, Coombs G, Kearns AM, Hill RLR, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramirez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, Nubel U. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res 23:653–664.

39. Planet PJ, Diaz L, Kolokotronis S-O, Narechania A, Reyes J, Xing G, Rincon S, Smith H, Panesso D, Ryan C, Smith DP, Guzman M, Zurita J, Sebra R, Deikus G, Nolan RL, Tenover FC, Weinstock GM, Ashley Robinson D, Arias CA. 2015. Parallel Epidemics of Community-Associated Methicillin-ResistantStaphylococcus aureusUSA300 Infection in North and South America. J Infect Dis 212:1874–1882.

40. Stinear TP, Holt KE, Chua K, Stepnell J, Tuck KL, Coombs G, Harrison PF, Seemann T, Howden BP. 2014. Adaptive change inferred from genomic population analysis of the ST93 epidemic clone of community-associated methicillin-resistant Staphylococcus aureus. Genome Biol Evol 6:366–378.

41. Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simões PM, Tristan A, Petersen A, Aziz M, Kiil K, Cirković I, Udo EE, del Campo R, Vuopio-Varkila J, Ahmad N, Tokajian S, Peters G, Schaumburg F, Olsson-Liljequist B, Givskov M, Driebe EE, Vigh HE, Shittu A, Ramdani-Bougessa N, Rasigade J-P, Price LB, Vandenesch F, Larsen AR, Laurent F. 2014. Origin and evolution of European community-acquired methicillin-resistant Staphylococcus aureus. MBio 5:e01044–14.

42. Baines SL, Holt KE, Schultz MB, Seemann T, Howden BO, Jensen SO, van Hal SJ, Coombs GW, Firth N, Powell DR, Stinear TP, Howden BP. 2015. Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant Staphylococcus aureus. MBio 6:e00080.

43. Méric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, Mikhail J, Harris LG, Wilkinson TS, Rolo J, Lamble S, Bray JE, Jolley KA, Hanage WP, Bowden R, Maiden MCJ, Mack D, de Lencastre H, Feil EJ, Corander J, Sheppard SK. 2015. Ecological Overlap and Horizontal Gene Transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol 7:1313–1328.

44. Croucher NJ, Klugman KP. 2014. The Emergence of Bacterial "Hopeful Monsters." MBio 5:e01550–14–e01550–14.

45. Gould SJ. 1977. The return of hopeful monsters.

46. World Health Organization. 2010. Global Tuberculosis Control: WHO Report 2010. World Health Organization.

47. Flynn JL, Chan J, Lin PL. 2011. Macrophages and control of granulomatous inflammation in tuberculosis. Mucosal Immunol 4:271–278.

48. Fonseca JD, Knight GM, McHugh TD. 2015. The complex evolution of antibiotic resistance in Mycobacterium tuberculosis. Int J Infect Dis 32:94–100.

49. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. 2002. Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. Trends Microbiol 10:45–52.

50. Coscolla M, Gagneux S. 2014. Consequences of genomic diversity in Mycobacterium tuberculosis. Semin Immunol 26:431–444.

51. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbón MH, Bobadilla del Valle M, Fyfe J, García-García L, Rastogi N, Sola C, Zozio T, Guerrero MI, León CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendón A, Sifuentes-Osornio J, Ponce de León A, Cave MD, Fleischmann R, Whittam TS, Alland D. 2006. Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. J Bacteriol 188:759–772.

52. Wang W, Hu Y, Mathema B, Jiang W, Kreiswirth B, Xu B. 2012. Recent transmission of W-Beijing family Mycobacterium tuberculosis in rural eastern China. Int J Tuberc Lung Dis 16:306–311.

53. Liu M, Jiang W, Liu Y, Zhang Y, Wei X, Wang W. 2014. Increased genetic diversity of the Mycobacterium tuberculosis W-Beijing genotype that predominates in eastern China. Infect Genet Evol 22:23–29.

54. Zanini F, Carugati M, Schiroli C, Lapadula G, Lombardi A, Codecasa L, Gori A, Franzetti F. 2014. Mycobacterium tuberculosis Beijing family: analysis of the epidemiological and clinical factors associated with an emerging lineage in the urban area of Milan. Infect Genet Evol 25:14–19.

55. Li D, Dong C-B, Cui J-Y, Nakajima C, Zhang C-L, Pan X-L, Sun G-X, Dai E-Y, Suzuki Y, Zhuang M, Ling H. 2014. Dominant modern sublineages and a new modern sublineage of Mycobacterium tuberculosis Beijing family clinical isolates in Heilongjiang Province, China. Infect Genet Evol 27:294–299.

56. Luo T, Comas I, Luo D, Lu B, Wu J, Wei L, Yang C, Liu Q, Gan M, Sun G, Shen X, Liu F, Gagneux S, Mei J, Lan R, Wan K, Gao Q. 2015. Southern East Asian origin and coexpansion of Mycobacterium tuberculosis Beijing family with Han Chinese. Proc Natl Acad Sci U S A 112:8136–8141.

57. Merker M, Blin C, Mona S, Duforet-Frebourg N, Lecher S, Willery E, Blum MGB, Rüsch-Gerdes S, Mokrousov I, Aleksic E, Allix-Béguec C, Antierens A, Augustynowicz-Kopeć E, Ballif M, Barletta F, Beck HP, Barry CE 3rd, Bonnet M, Borroni E, Campos-Herrero I, Cirillo D, Cox H, Crowe S, Crudu V, Diel R, Drobniewski F, Fauville-Dufaux M, Gagneux S, Ghebremichael S, Hanekom M, Hoffner S, Jiao W-W, Kalon S, Kohl TA, Kontsevaya I, Lillebæk T, Maeda S, Nikolayevskyy V, Rasmussen M, Rastogi N, Samper S, Sanchez-Padilla E, Savic B, Shamputa IC, Shen A, Sng L-H, Stakenas P, Toit K, Varaine F, Vukovic D, Wahl C, Warren R, Supply P, Niemann S, Wirth T. 2015. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet 47:242–249.

58. Drolet GJ. 1945. World War I and Tuberculosis. A Statistical Summary and Review. Am J Public Health Nations Health 35:689–697.

59. Hershkovitz I, Donoghue HD, Minnikin DE, May H, Lee OY-C, Feldman M, Galili E, Spigelman M, Rothschild BM, Bar-Gal GK. 2015. Tuberculosis origin: The Neolithic scenario. Tuberculosis 95 Suppl 1:S122–6.

60. Cockburn TA. 1964. THE EVOLUTION AND ERADICATION OF INFECTIOUS DISEASES. Perspect Biol Med 7:498–499.

61. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176–1182.

62. Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, van Soolingen D, Cole ST. 2002. A new evolutionary scenario for the Mycobacterium tuberculosis complex. Proc Natl Acad Sci U S A 99:3684–3689.

63. Hershberg R, Lipatov M, Small PM, Sheffer H, Niemann S, Homolka S, Roach JC, Kremer K, Petrov DA, Feldman MW, Gagneux S. 2008. High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. PLoS Biol 6:e311.

64. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ, Campbell TJ, Majander K, Wilbur AK, Guichon RA, Wolfe Steadman DL, Cook DC, Niemann S, Behr MA, Zumarraga M, Bastida R, Huson D, Nieselt K, Young D, Parkhill J, Buikstra JE, Gagneux S, Stone AC, Krause J. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514:494–497.

65. Patrucco R, Tello R, Bonavia D. 1983. Parasitological Studies of Coprolites of Pre-Hispanic Peruvian Populations. Curr Anthropol 24:393–394.

66. Herring DA, Sattenspiel L. 2007. Social contexts, syndemics, and infectious disease in northern Aboriginal populations. Am J Hum Biol 19:190–202.

67. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8:e1002824.

68. Bertelli C, Greub G. 2013. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect 19:803–813.

69. Fournier P-E, Dubourg G, Raoult D. 2014. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med 6:114.

70. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer J. 2015. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol 53:2230–2237.

71. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk K-C, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Grard G, Leroy E, Schneider BS, Fair JN, Martinez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J, Miller S, Chiu CY. 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res 24:1180–1192.

72. Hasan NA, Young BA, Minard-Smith AT, Saeed K, Li H, Heizer EM, McMillan NJ, Isom R, Abdullah AS, Bornman DM, Faith SA, Choi SY, Dickens ML, Cebula TA, Colwell RR. 2014. Microbial community profiling of human saliva using shotgun metagenomic sequencing. PLoS One 9:e97699.

73. Takeuchi F, Sekizuka T, Yamashita A, Ogasawara Y, Mizuta K, Kuroda M. 2014. MePIC, Metagenomic Pathogen Identification for Clinical Specimens. Jpn J Infect Dis 67:62–65.

74. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL, Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L, Peacock SJ, Holden

M, Wilson D, Read TD, van den Elsen J, Priest NK, Feil EJ, Hurst LD, Josefsson E, Massey RC. 2014. Predicting the virulence of MRSA from its genome sequence. Genome Res 24:839–849.

75. Suzuki S, Horinouchi T, Furusawa C. 2014. Prediction of antibiotic resistance by gene expression profiles. Nat Commun 5:5792.

76. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo Elias C, Johnson JR, Walker AS, Peto TEA, Crook DW. 2013. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. J Antimicrob Chemother 68:2234–2244.

77. Bradley P, Claire Gordon N, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N, Omar SV, Grace Smith E, Buck D, McVean G, Sarah Walker A, Peto T, Crook D, Iqbal Z. 2015. Rapid antibiotic resistance predictions from genome sequence data for S. aureus and M. tuberculosis.

78. Bristow F, Adam J, Carriço JA, Courtot M, Dhillon B, Dooley D, Griffiths E, Isaac-Renton J, Keddy A, Kruczkiewicz P, Laird M, Matthews T, Petkau A, Schriml L, Shay J, Taboada E, Tang P, Thiessen J, Winsor G, Beiko RG, Van Domselaar G GM, Hsiao W, The IRIDA Consortium, Brinkman, F. IRIDA: Canada's federated platform for genomic epidemiologyThe 8th Meeting of the Global Microbial Identifier.

79. Magalhães WCS, Rodrigues MR, Silva D, Soares-Souza G, Iannini ML, Cerqueira GC, Faria-Campos AC, Tarazona-Santos E. 2012. DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology studies. Genet Epidemiol 36:360–367.

80. Land ML, Hyatt D, Jun S-R, Kora GH, Hauser LJ, Lukjancenko O, Ussery DW. 2014. Quality scores for 32,000 genomes. Stand Genomic Sci 9:20.

CONCLUSIONS

Different evolutionary scenarios. In this thesis, I have presented five research papers and one review article. Two of the research articles are focused on the genomic evolution of a nosocomial pathogen strain. These works are focused on *Klebsiella pneumoniae* CC258 (Gaiarsa et al., 2015; see Article 1) and *Acinetobacter baumannii* ST78 (Gaiarsa et al., submitted; see Article 2). In the *K. pneumoniae* paper, the main finding is the presence of a 1.3 million base recombination in the genomes of the CC258 isolates. This discovery was used to complete the picture of the evolutionary history that led to the current genome of this worldwide spread pathogen. Moreover, it was possible to date the newly detected and the previously reported recombination events. The so-depicted scenario is reminiscent of the evolutionary model of punctuated equilibrium in which fast evolutionary events (in this case recombinations) lead to the establishment of an highly fit variant. The model of punctuated equilibrium was theorized by Stephen Gould to explain the high morphologic variability of the fossils of the Cambrian era. Such species were the result of a very quick and diverging evolutionary process, but only a few of them were fit enough to survive in the ages. Gould named those species “hopeful monsters” after a name invented by evolutionist Richard Goldschmidt (1, 2). Indeed, the isolates of CC258 can be compared to the “hopeful monsters”. This evolutionary strategy is not exclusive of the CC258: as a matter of facts, it is known that recombination of wide portions of genomes are a common feature in bacterial pathogens, such as *Vibrio cholerae*, *Clostridium difficile*, *Salmonella enterica* and *Streptococcus pneumoniae*. These and other examples are listed in a commentary article published by Nicholas Croucher in 2014, called ‘The Emergence of Bacterial “Hopeful Monsters”’ (3).

Not all pathogens evolve in this way, though. In fact, the bacterial strain analyzed in Article 2 showed to have a very low recombination lifestyle and a low gene content variability. Indeed, in this work about *A. baumannii* ST78, we were able to identify two different clusters of isolates, some of them (ST78A) with a low gene content variability and a very high number of copies of insertion sequences (ISs). In our evolutionary model, the acquisition of exogenous DNA is slowed down by the inactivation of the *comEC/rec2* gene, which codes for a protein used for the importation of DNA through transformation and thus involved in the pathway of homologous recombination. ST78 has a “strong” phenotype (e.g. high production of biofilm) which grants it a very high persistence. Yet, the loss of genomic plasticity in this bacterium limits its ability to adapt to environmental changes and could be the cause of its low incidence.

Almost forensic genomics. The remaining three research articles (Article 3: Onori et al., 2015; Article 4: Scaltriti et al., 2015; Article 5: Comandatore et al., 2017) are focused on the reconstruction of epidemic events of bacterial infections. Article 3 reconstructs the chain of contagion of *K. pneumoniae* in an intensive care unit of an Italian hospital. In this case, global phylogeny and isolation dates helped in identifying what isolates were involved in the outbreak event. Then, an approach based on core SNPs and isolation dates allowed to reconstruct the spreading route, which resulted to be a star-like radiation from the patient zero to all the other ones and suggested that the hospital staff may be responsible of the spread, through negligence in respecting the safety procedures. In Article 4, phylogeny allowed to identify the source of a food poisoning by *Salmonella enterica.* In this case, only synonymous mutations were used as input to the phylogeny, in order to filter out pathoadaptative mutations. Lastly, in Article 5, epidemiological data, molecular typing and SNP-based phylogeny were used to reconstruct the dynamics of infection of nine *Listeria monocytogenes* isolates, which were believed to be part of the same outbreak and in the end proved to be genomically unrelated, and thus indicating different epidemic events.

All three articles underline the high potential of genomic epidemiology, which is able to provide the resolution to reconstruct the chain of events happened in very short time spans and small spaces. In all three cases it was possible to identify the source and responsibility for the contagion. No result of these papers was used as evidence in a trial, but the level of precision in this kind of investigations is so high to suggest a possible forensic application (4).

Future developments in the projects. Bacterial “hopeful monsters”, such as *K. pneumoniae* CC258, are pathogens with a very high rate of HGT. Thus, besides studying the evolutionary history of the strains, it is necessary to understand how these “superbugs” are formed: i.e. in what environments and conditions do recombinations happen. This could be achieved by studying the patterns of coinfections of the same patient by more than one variant of the same species. Of course the copresence in the same environment is the *conditio sine qua non* DNA exchange can happen. One other way of studying this phenomenon would be to use phylogenetic approaches to track the movements of transferred DNA portions instead of those of entire genomes.

Moreover, the human body is not the only environment where “superbugs” should be searched. *K. pneumoniae*, for example, is almost ubiquitous and can be normally found in a wide array of other niches, such as other animals (either pets, farm animals or bugs) but also vegetation and water. Thus, it would be important to understand in what environments can resistant and/or virulent variants be found, i.e. what the reservoirs of pathogens are. With this investigative strategy, it would be also possible to detect what the spreading routes are for dangerous bacteria and what are the hub points of transmission that could be tackled in order to limit the diffusion of the pathogens. Lastly, it would be possible to understand if one of the studied environments can host recombination events. Indeed, the more diverse the environments are, the more diverse is the gene pool that can be mixed when “superbugs” are formed.

Article 2 underlines the importance of selfish DNA, such as ISs, in the evolution of genomes. Indeed the phenomenon of reduced genome plasticity caused by the proliferation of ISs should be investigated further. It should be understood if it is common to other variants of *A. baumannii* or even to other species. Moreover in *A. baumannii* ST78, all genes inactivated by ISs could be detected in order to understand what pathways are targeted by this mechanism.

Potential translational application of microbial genomics. All works presented in this thesis and, more in general the vast majority of articles published in the field of microbial genomics and genomic epidemiology, prove that the resolution power of these techniques is very high. The characterization of an isolate and the detection of molecular determinants for antimicrobial resistance are quick and cheap tasks when performed with genomics. In light of the continuous drop of costs for DNA sequencing, there is a moral duty to exploit these new technologies to develop cheap diagnostic and surveillance tools (5). The new tests will eventually substitute the traditional molecular and microbiological assays, thus cutting the time span between isolation and cure administration. This will lower the chance of death of the patient as well as limit the possibility of spread of the pathogens.

REFERENCES

1. Gould SJ. 1977. The Return of Hopeful Monsters. Nat Hist 86:22–30.

2. Goldschmidt R. 1933. SOME ASPECTS OF EVOLUTION. Science 78:539–547.

3. Croucher NJ, Klugman KP. 2014. The emergence of bacterial “hopeful monsters.” MBio 5:e01550–14.

4. Schmedes SE, Sajantila A, Budowle B. 2016. Expansion of Microbial Forensics. J Clin Microbiol 54:1964–1974.

5. Fournier P-E, Dubourg G, Raoult D. 2014. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med 6:114.

ARTICLE 1

Supplemental Material

TABLE S1

Table of *Klebsiella pneumoniae* isolates sequenced in this study and selected characteristics.

PUBLICATION_NAME	ENTRY	YEAR OF COLLECTION	MLST	HOSPITAL	PHENOTYPE
100SGR	ERS480596	2012	16	SAN GIOVANNI ROTONDO	ESBL
101BO	ERS480597	2012	512	BOLOGNA	KPC
102BO	ERS480598	2012	512	BOLOGNA	KPC
103BO	ERS480599	2012	512	BOLOGNA	KPC
104BO	ERS480600	2012	512	BOLOGNA	KPC
10PV	ERS480601	2012	307	PAVIA	ESBL
11PV	ERS480602	2013	258	PAVIA	KPC
12PV	ERS480603	2013	258	PAVIA	KPC
13PV	ERS480604	2013	258	PAVIA	KPC
14PV	ERS480605	2012	1624	PAVIA	Susceptible
15PV	ERS480606	2012	976	PAVIA	Susceptible
16BO	ERS480607	2011	37	BOLOGNA	ESBL
17PV	ERS480608	2012	307	PAVIA	ESBL
18PV	ERS480609	2013	15	PAVIA	ESBL
19PV	ERS480610	2013	15	PAVIA	ESBL
20PV	ERS480611	2012	1631	PAVIA	Susceptible
21PV	ERS480612	2012	240	PAVIA	Susceptible
22PV	ERS480613	2012	1625	PAVIA	Susceptible
23PV	ERS480614	2013	258	PAVIA	KPC
24PV	ERS480615	2013	258	PAVIA	KPC
25BO	ERS480616	2013	35	BOLOGNA	Susceptible
26BO	ERS480617	2013	35	BOLOGNA	Susceptible
27BO	ERS480618	2011	45	BOLOGNA	ESBL
28BO	ERS480619	2011	37	BOLOGNA	ESBL
29BO	ERS480620	2012	512	BOLOGNA	KPC
30BO	ERS480621	2012	512	BOLOGNA	KPC
31AVR	ERS480622	2013	512	CESENA	KPC
32AVR	ERS480623	2013	466	CESENA	Susceptible
34AVR	ERS480624	2013	405	CESENA	Susceptible
36AVR	ERS480625	2013	37	CESENA	Susceptible
37AVR	ERS480626	2013	323	CESENA	ESBL
39AVR	ERS480627	2013	395	CESENA	Susceptible
40AVR	ERS480628	2013	307	CESENA	ESBL
41AVR	ERS480629	2013	16	CESENA	ESBL
42AVR	ERS480630	2013	512	CESENA	KPC
43AVR	ERS480631	2013	512	CESENA	KPC
44AVR	ERS480632	2013	512	CESENA	KPC
45AVR	ERS480633	2013	160	CESENA	Susceptible
46AVR	ERS480634	2013	395	CESENA	ESBL
47AVR	ERS480635	2013	323	CESENA	ESBL
48AVR	ERS480636	2013	512	CESENA	KPC
49BG	ERS480637	2012	147	BERGAMO	Susceptible
50BG	ERS480638	2011	258	BERGAMO	KPC
51BG	ERS480639	2007	1626	BERGAMO	Susceptible
52BG	ERS480640	2006	268	BERGAMO	Susceptible
53BG	ERS480641	2009	321	BERGAMO	Susceptible
54BG	ERS480642	2011	258	BERGAMO	KPC
55BG	ERS480643	2012	466	BERGAMO	Susceptible
56BG	ERS480644	2011	258	BERGAMO	KPC

PUBLICATION_NAME	ENTRY	YEAR OF COLLECTION	MLST	HOSPITAL	PHENOTYPE
57BG	ERS480645	2011	258	BERGAMO	KPC
58BG	ERS480646	2011	512	BERGAMO	KPC
60BG	ERS480647	2012	45	BERGAMO	ESBL
62BG	ERS480648	2011	147	BERGAMO	ESBL
63BG	ERS480649	2011	1627	BERGAMO	ESBL
65BO	ERS480650	2013	1243	BOLOGNA	Susceptible
66BO	ERS480651	2013	416	BOLOGNA	Susceptible
67BO	ERS480652	2013	1628	BOLOGNA	Susceptible
68BO	ERS480653	2011	37	BOLOGNA	ESBL
69BO	ERS480654	2011	277	BOLOGNA	ESBL
70BO	ERS480655	2011	37	BOLOGNA	ESBL
71RE	ERS480656	2011	258	REGGIO EMILIA	KPC
72RE	ERS480657	2011	258	REGGIO EMILIA	KPC
73RE	ERS480658	2011	258	REGGIO EMILIA	KPC
74RE	ERS480659	2011	512	REGGIO EMILIA	KPC
75RE	ERS480660	2011	258	REGGIO EMILIA	KPC
76RE	ERS480661	2012	1243	REGGIO EMILIA	Susceptible
77RE	ERS480662	2012	1629	REGGIO EMILIA	Susceptible
78RE	ERS480663	2012	1164	REGGIO EMILIA	Susceptible
79RE	ERS480664	2012	35	REGGIO EMILIA	Susceptible
81RE	ERS480665	2011	147	REGGIO EMILIA	ESBL
82RE	ERS480666	2012	405	REGGIO EMILIA	ESBL
83RE	ERS480667	2012	147	REGGIO EMILIA	ESBL
84RE	ERS480668	2012	322	REGGIO EMILIA	ESBL
85RE	ERS480669	2012	37	REGGIO EMILIA	ESBL
86SGR	ERS480670	2011	512	SAN GIOVANNI ROTONDO	KPC
87SGR	ERS480671	2011	512	SAN GIOVANNI ROTONDO	KPC
88SGR	ERS480672	2011	512	SAN GIOVANNI ROTONDO	KPC
89SGR	ERS480673	2011	512	SAN GIOVANNI ROTONDO	KPC
90SGR	ERS480674	2011	512	SAN GIOVANNI ROTONDO	KPC
91SGR	ERS480675	2012	29	SAN GIOVANNI ROTONDO	Susceptible
92SGR	ERS480676	2012	70	SAN GIOVANNI ROTONDO	Susceptible
93SGR	ERS480677	2012	35	SAN GIOVANNI ROTONDO	Susceptible
94SGR	ERS480678	2012	45	SAN GIOVANNI ROTONDO	Susceptible
95SGR	ERS480679	2012	1307	SAN GIOVANNI ROTONDO	Susceptible
96SGR	ERS480680	2012	1630	SAN GIOVANNI ROTONDO	ESBL
97SGR	ERS480681	2012	512	CESENA	ESBL
98SGR	ERS480682	2012	20	SAN GIOVANNI ROTONDO	ESBL
99SGR	ERS480683	2012	15	SAN GIOVANNI ROTONDO	ESBL
9PV	ERS480684	2012	307	PAVIA	ESBL

TABLE S2

Genes with potential effect on virulence or antibiotic resistance phenotype comprised in recombined region of ~1.3 Mb described in this work. Coordinates and strand are referred to the genome of the reference strain NJST258_1, annotation was obtained by BLAST search against a specifically designed database, as reported in the materials and methods section.

START	END	STRAND	PRODUCT NAME
18482	20224	+	Integral membrane protein with trka domains
29471	30706	-	Multidrug resistance protein emrD
69212	72319	-	multidrug transporter
72319	73443	-	acriflavine resistance protein E
138949	140223	-	3-deoxy-D-manno-octulosonic acid transferase WaaA
142431	143558	-	glycosyl transferase family 1 WabG
143555	144631	-	glycosyl transferase family 9 WaaQ
148902	149873	-	ADP-heptose--LPS heptosyltransferase WaaC
149877	150935	-	ADP-heptose--LPS heptosyltransferase WaaF
150945	151877	-	ADP-L-glycero-D-manno-heptose-6-epimerase RfaD
188732	190165	+	Xylose isomerase
291995	293161	+	UDP-4-amino-L-arabinose synthase PmrH
293109	294146	+	Undecaprenyl-phosphate alpha-4-amino-L-arabinosyltransferase ArnC
294143	296128	+	UDP-4-amino-4-deoxy-L-arabinose formyltransferase ArnA
296125	297027	+	4-deoxy-4-formamido-L-arabinose-phospho-UDP deformylase PmrJ
297024	298682	+	Undecaprenyl phosphate-alpha-4-amino-4-deoxy-L-arabinose arabinosyl transferase ArnT
298679	299017	+	4-amino-4-deoxy-L-arabinose-phospho-UDP flippase PmrL
299017	299397	+	4-amino-4-deoxy-L-arabinose-phospho-UDP flippase PmrM
373562	374329	+	Transcriptional regulatory protein ompR
374326	375681	+	osmolarity sensor protein envZ
408428	409060	-	Crp/Fnr family transcriptional regulator
424107	425291	+	Elongation factor Tu
457426	460536	-	multidrug transporter AcrB
460549	461688	-	acrE
462055	462717	+	AcrAB operon repressor
532636	532923	-	yhbH
536245	536811	-	yrbI
540606	541172	+	yrbD/mlad
541191	541826	+	mlac
553941	554792	+	dihydropteroate synthase
4632900	4634141	+	multidrug transporter
4638808	4639326	-	Transcriptional regulator, MarR family protein
4646942	4648030	-	ABC-type_sugar_transport_system,_periplasmic_component
4682645	4683964	-	xylose isomerase
4684337	4685830	+	Xyloside transporter
4685889	4687568	+	beta-xylosidase
4845809	4846771	+	Phosphatidylserine_decarboxylase psd
4853100	4854890	+	fumarate reductase frdA
4854835	4855617	+	frdB
4855628	4856023	+	fumarate reductase frdC
4856004	4856393	+	frdD
4856507	4857040	+	Bacterial_lipocalin
4857037	4857354	-	Membrane_transporter_of_cations_and_cationic_drugs sugE
4874086	4875387	+	C4-dicarboxylate ABC transporter DcuA
4901698	4904016	-	Ferrienterobactin receptor precursor fepA
4996917	5000978	-	rpoB
5005173	5006357	-	elongation factor Tu
5149976	5150674	+	cpxR
5150671	5152044	+	Sensor protein cpxA
5160685	5161671	+	ABC-type_sugar_transport_system,_periplasmic_component rhaS
5256865	5258115	-	Chloramphenicol resistance protein

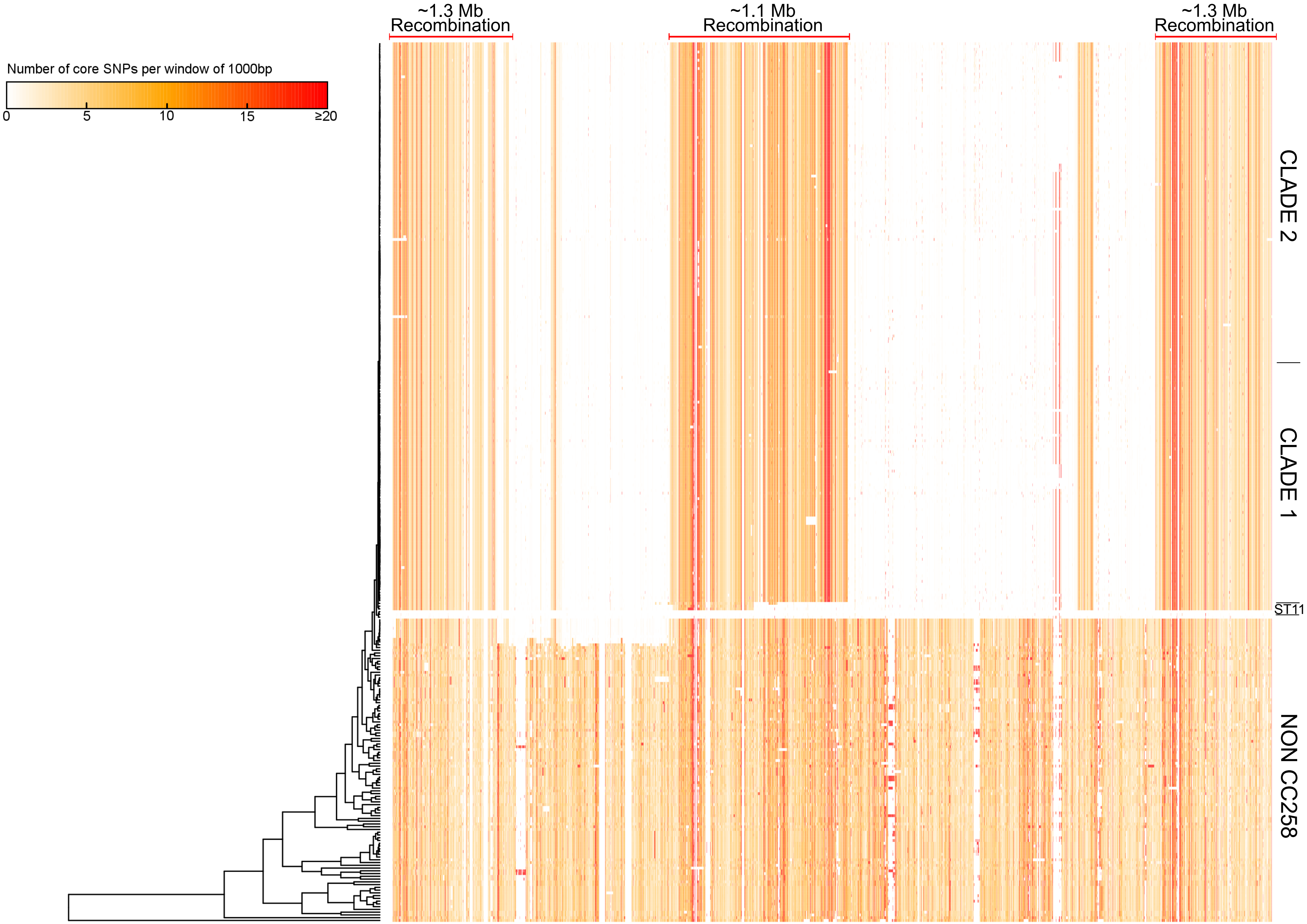


FIGURE S1

Clustering of core SNPs in the 319 Klebsiella pneumoniae genomes. The phylogenetic reconstruction is shown on the left, while the core SNP frequency is shown on the right, in shades of red representing number of core SNPs per 1000bp windows for each genome. Detected recombinations are indicated on the top of the figure, main clades detected in the phylogenetic analysis are indicated on the right side of the figure.

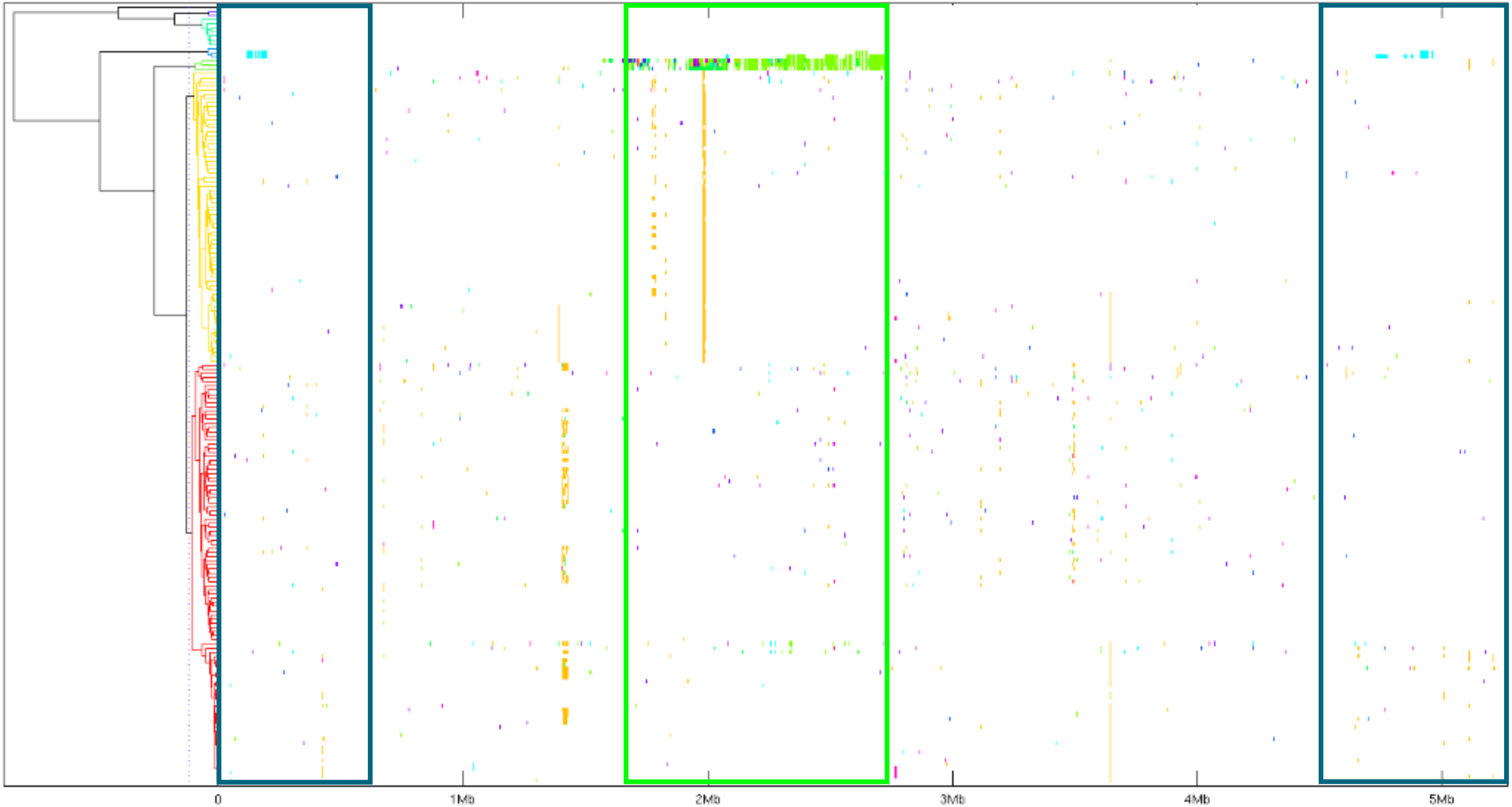


FIGURE S2

Recombination analysis obtained with BRATnextgen. A subdataset of 187 genomes was used as input for a 100 iteration analysis with 100 replicates with the BRATnextgen software. The recombination proposed by Chen and coworkers is detected in green while the recombination proposed in this work is detected in blue. Recombined regions as detected with the SNP-based method are indicated with boxes, using the same colors as those chosen by the BRATnextgen software.

FIGURE S3

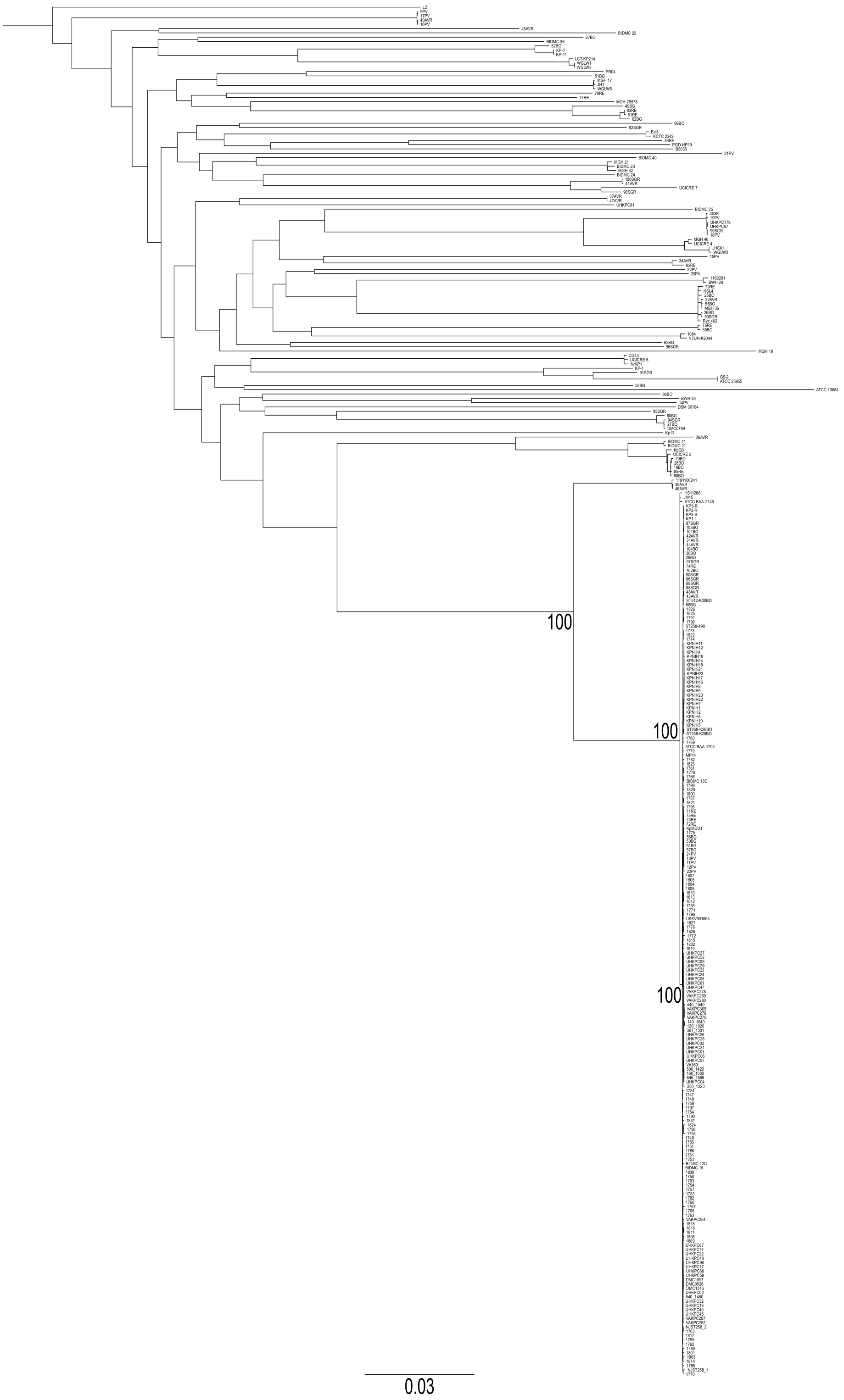


FIGURE S3. Phylogeny of the 319 *Klebsiella pneumoniae* genomes based on core SNPs in non-recombined regions. Phylogeny was reconstructed starting from an alignment of 55,368 core SNPs, located outside of the two main recombined regions of the genome, using the software RAxML, with the Generalised time-reversible (GTR) model and 100 bootstrap replicates. Bootstrap is shown only for the three main nodes of Clonal Complex 258.

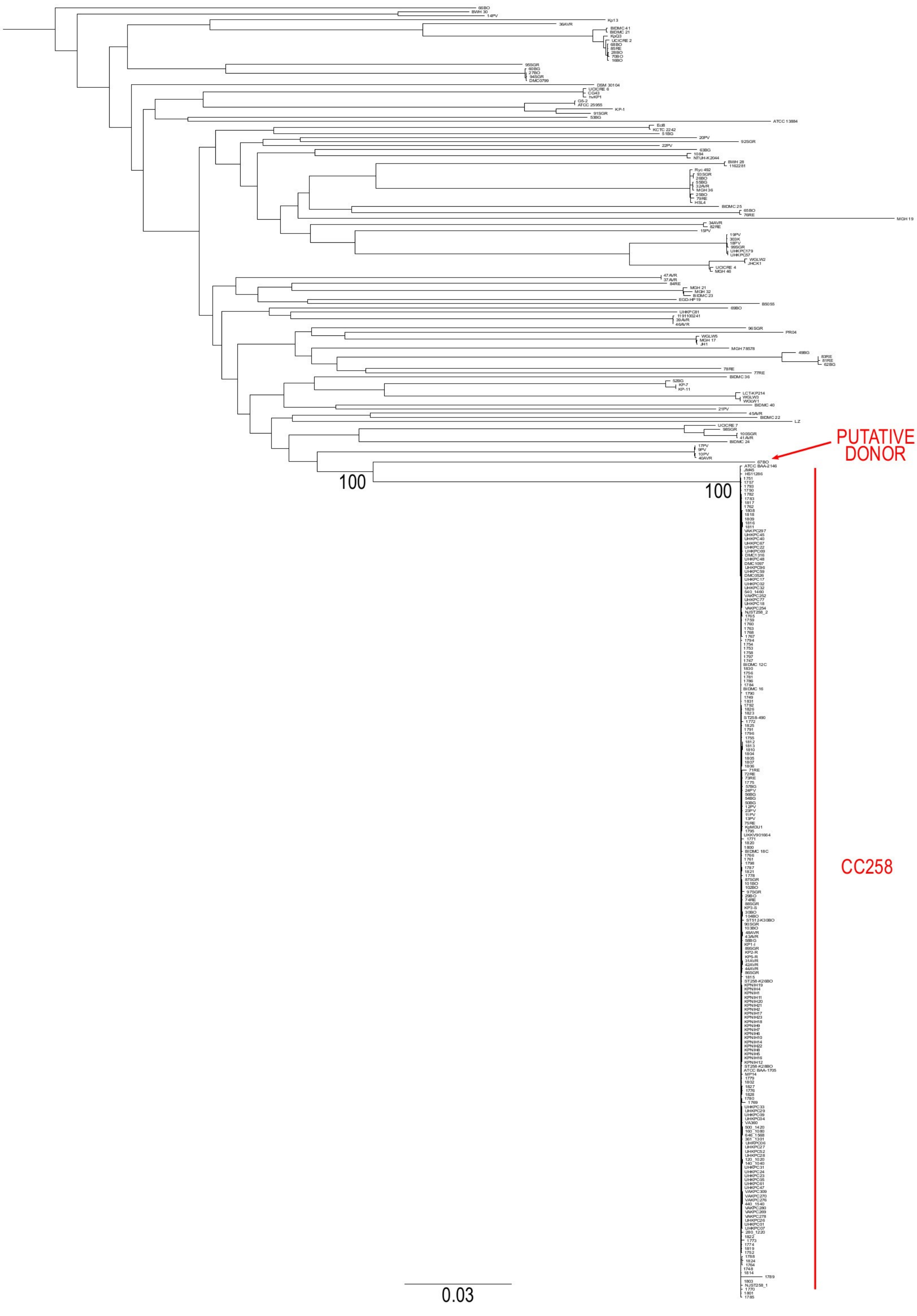


FIGURE S4

phylogeny of the 319 *Klebsiella pneumoniae* genomes based on core SNPs in the ~1.3Mb recombined region. Phylogeny was reconstructed starting from an alignment of 24,537 core SNPs present only in the recombined region located from 4,554,906 to 629,621, spanning the origin of replication, using coordinates of genome NJST258_1. The tree was obtained using the software RAxML, with the Generalised time-reversible (GTR) model and 100 bootstrap replicates, bootstraps are shown only for nodes of interest. The putative donor of the ~1.3Mb recombined region, 67BO, results as sister clade of the recipient CC258.

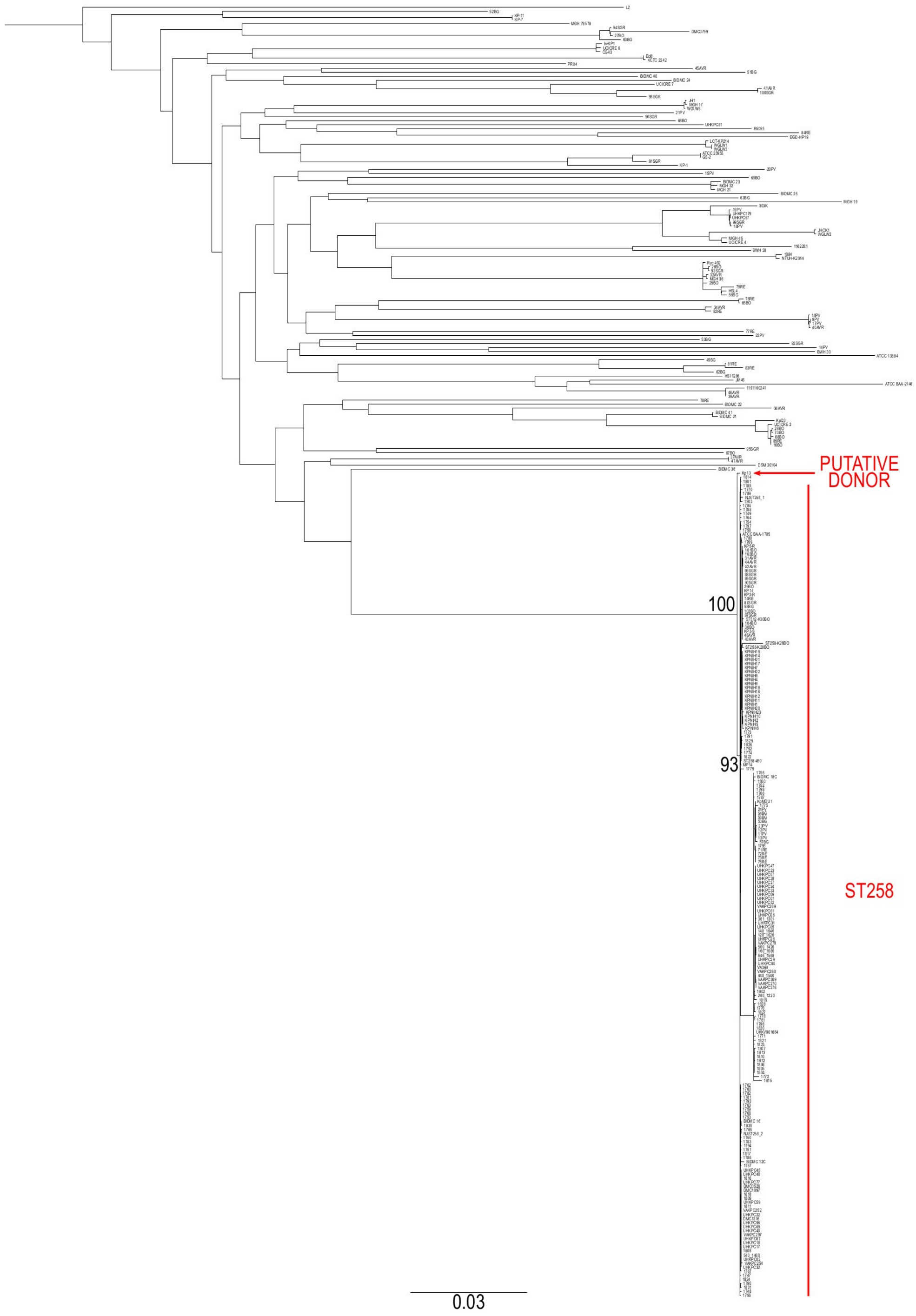
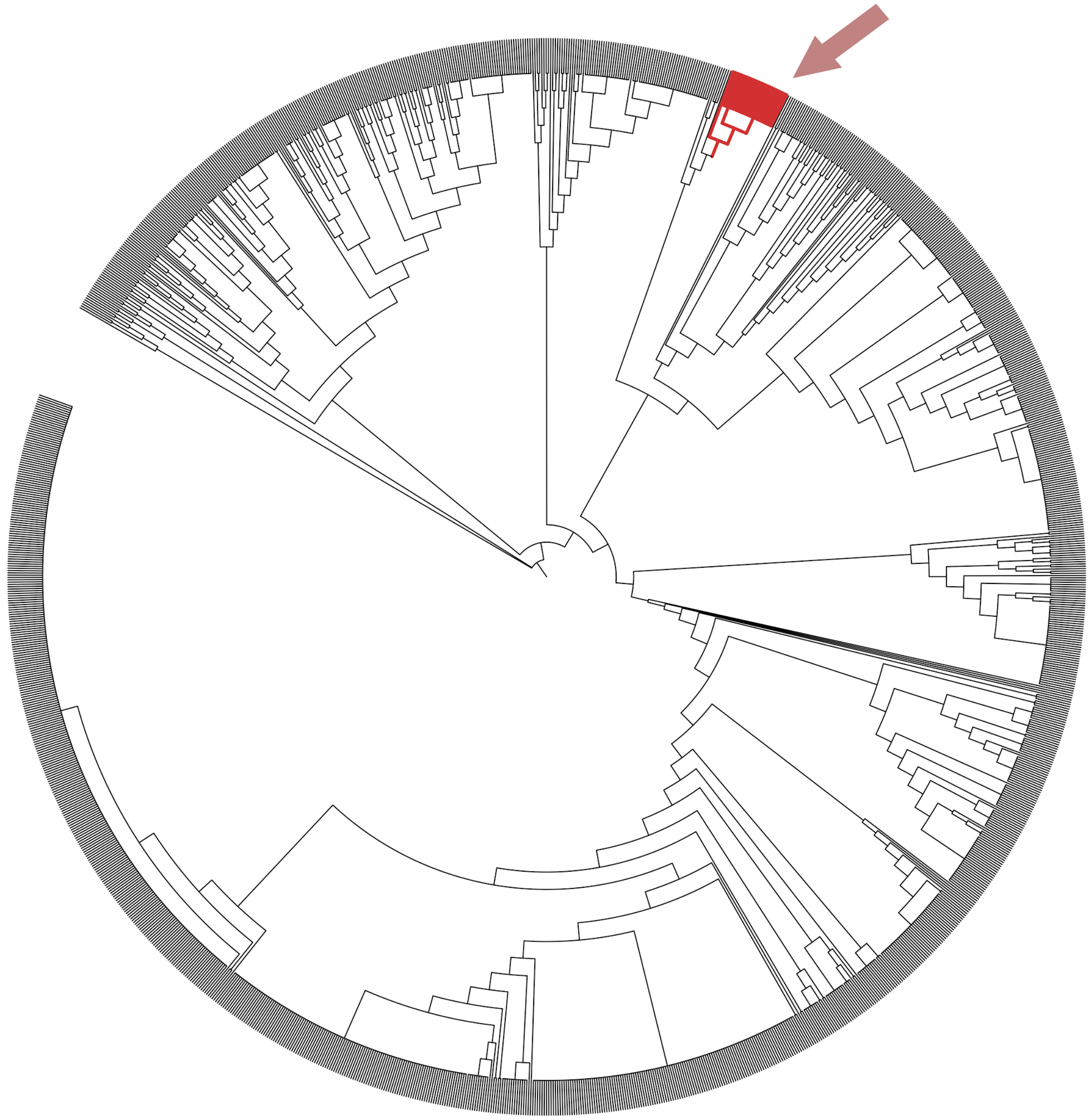


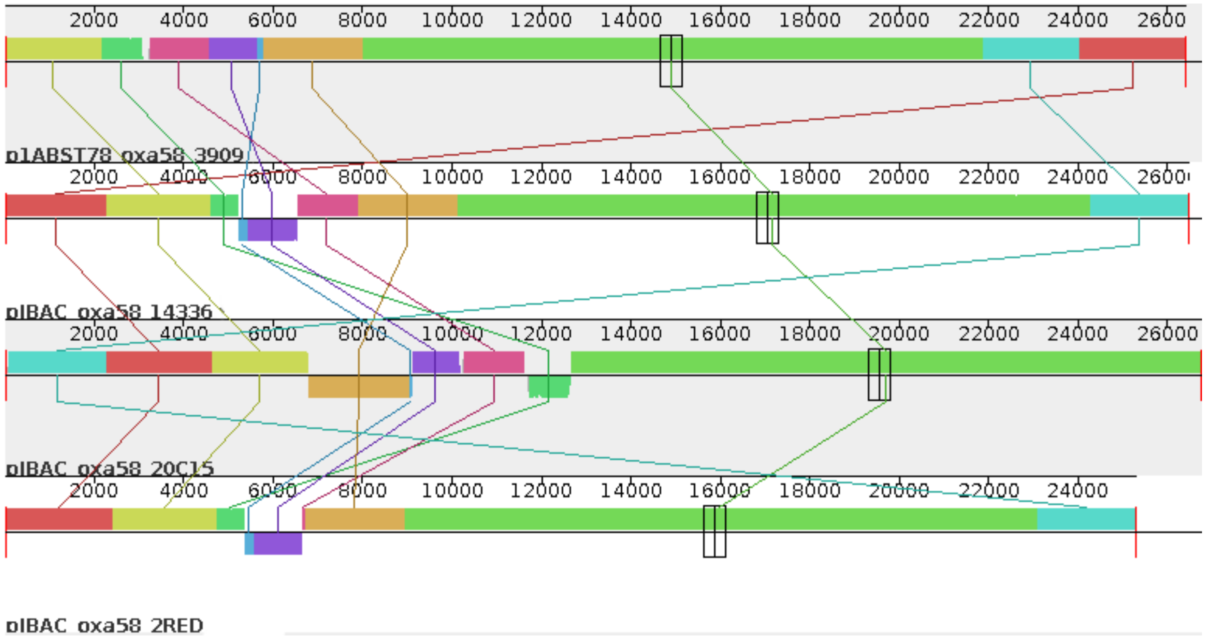
FIGURE S5. phylogeny of the 319 *Klebsiella pneumoniae* genomes based on core SNPs in the ~1.1Mb recombined region. Phylogeny was reconstructed starting from an alignment of 14,905 core SNPs present only in the recombined region located from 1,675,550 to 2,740,033, using coordinates of genome NJST258_1. The tree was obtained using the software RAxML, with the Generalised time-reversible (GTR) model and 100 bootstrap replicates, bootstraps are shown only for nodes of interest. The putative donor of the ~1.1Mb recombined region, Kp13, results as sister clade of the recipient CC258 clade, with the exclusion of the ST11 clade.

ARTICLE 2

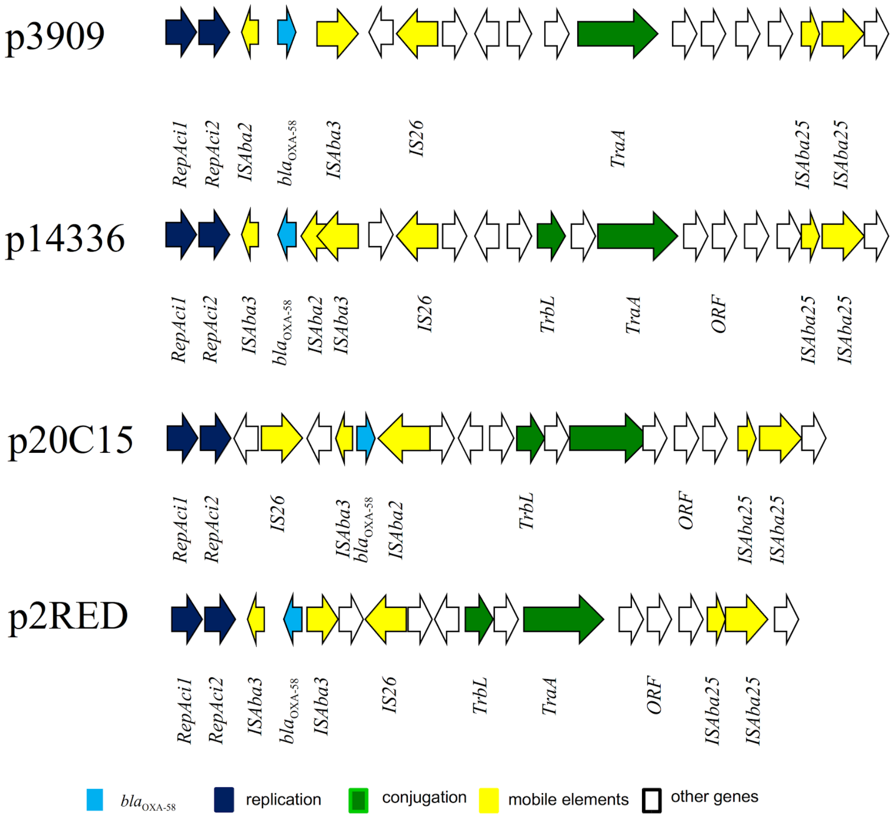
Supplemental Material



Supplementary Figure 1. Global phylogeny of the species *Acinetobacter baumannii* obtained using fasttree on a dataset of core SNPs*.* The genomes of the Sequence Type 78 are highlighted in red and indicated with an arrow; names of the strains are not reported for seek of better visualization.



Supplementary Figure 2. Synteny analysis of the plasmids containing the gene *bla*$_{\text{OXA-58}}$ among all strains of the Sequence Type 78. The analysis was run on the software Mauve.



Supplementary Figure 3. Structure of the genomic locus encompassing the gene *bla*$_{OXA-58}$ in the three newly described *Acinetobacter baumannii* plasmids and the previously described plasmids *p3909* and *p183Eco*

Isolate	MIC (µg/ml)		Biofilm formation capability (OD600)		
	MER	IPM			
2MG	0.75	0.38	0.4	0.4	0.4
65SM01	0.75	0.75	0.4	0.4	0.3
5MO	>32	>32	0.6	0.7	0.5
2RED09	>32	>32	0.4	0.4	0.6
14336	>32	>32	0.4	0.4	0.5
20C15	>32	>32	0.6	0.4	0.6
25C30	>32	>32	0.5	0.5	0.5
96SM	1	0.5	0.6	0.7	0.4
103SM	1	0.38	0.6	0.7	0.5
74SM01	0.75	0.5	0.4	0.4	0.3
MGTN	1	0.5	0.4	0.4	0.4
72SM01	0.5	0.38	0.6	0.7	0.8
68SM01	>32	0.38	0.2	0.7	0.5
MONUR	>32	0.38	0.5	0.5	0.5
3909	32	6	0.4	0.4	0.4
61SM01	>32	0.38	0.5	0.5	0.3

Supplementary Table 1. Phenotypic testing on the 16 Italian strains. A. Results of Minimum Inhibitory Concentration using the E-test method. B. Results of the biofilm formation capability assay

Table 2-A

Strain	*bla*$_{OXA-90}$	*ISAba1*	*bla*$_{OXA-58}$	*bla*$_{OXA-23}$	*bla*$_{ADC-52}$	*bla*$_{CARB-PSE}$	*carO*	*adeR*	*adeS*	*floR*	*sul2*	*aadB*	*aph(3')-Ic*
2MG	x	x			1SNP	x	x	x	x	x	x	x	x
2RED09	x	x	x		1SNP		x	x	x	x	x	x	x
5MO	x	x		x	1SNP		x	x	x	x	x	x	x
20C15	x	x	x	x	1SNP		x	x	x	x	x	x	x
25C30	1SNP	x			1SNP		x	x	x	x	x	x	x
61SM01	x	x			1SNP		x	x	x	x	x	x	x
65SM01	x	x			1SNP		x	x	x	x	x	x	x
68SM01	x	x			1SNP	x	x	x	x	x	x	x	x
72SM01	x	x			1SNP		x	x	x	x	x	x	x
74SM01	x	x			1SNP		x	x	x	x	x	x	x
96SM	x	x			1SNP		x	x	x	x	x	x	x
103SM	x	x			1SNP		x	x	x			x	x
14336	x	x	x		1SNP		x	x	x			x	x
MGTN	x	x			1SNP		x	x	x	x	x	x	x
MONUR	x	x			1SNP		x	x	x	x	x	x	x
3909	x	x	x		1SNP		x	x	INT	x	x	x	x
TG22142	3SNP	x			1SNP	x	x	x	INT	x	x	x	x
TG22146	3SNP	x			1SNP	x	x	x	INT	x	x	x	x
TG22150	3SNP	x			1SNP	x	x	x	INT				x
UH5207	x				x		x	x	x				
1096934	x				x		x	x	x				
831240	x				x		x	x	x				
855125	x				x		x	x	INT				
UH1752	x	x			x		x	x	x				
ABBL025	x				x		x	x	x				
ABBL026	x				x		x	x	x				

Supplementary Table 2-A. Results of the refined analysis of presence of genes of interest of the following categories: A. Resistance. B. Virulence. C. Competence. D. Biofilm formation

Table 2-B

Strain	*pil* genes*	*ptk*	*cap8J*	*zur*	*hcp*	*pmt*	*entE*	*sc1*	*bfmR*	*envZ*	*ompF*	*ostA*	*pbpG*	*ptk*	*rstA*	*epsA*	*bap*
2MG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
2RED09	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
5MO	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
20C15	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
25C30	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
61SM01	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
65SM01	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
68SM01	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
72SM01	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
74SM01	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
96SM	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
103SM	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
14336	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
MGTN	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
MONUR	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
3909	x	x	x	x	x	x	x	x	x	x	INT	x	x	x	x		x
TG22142	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
TG22146	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
TG22150	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
UH5207	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
1096934	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
831240	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
855125	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
UH1752	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
ABBL025	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x
ABBL026	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x

*. *pil* genes are A, B, C, F, M, O, Q, W, R, S, T, and U.

Supplementary Table 2-B. Results of the refined analysis of presence of genes of interest of the following categories: A. Resistance. B. Virulence. C. Competence. D. Biofilm formation

Table 2-C

Strain	*comEC*	*comB*	*comC*	*comD*	*comE*
2MG	INT	x	x	x	x
2RED09	INT	x	x	x	x
5MO	INT	x	x	x	x
20C15	INT	x	x	x	x
25C30	INT	x	x	x	x
61SM01	INT	x	x	x	x
65SM01	INT	x	x	x	x
68SM01	INT	x	x	x	x
72SM01	INT	x	x	x	x
74SM01	INT	x	x	x	x
96SM	INT	x	x	x	x
103SM	INT	x	x	x	x
14336	INT	x	x	x	x
MGTN	INT	x	x	x	x
MONUR	INT	x	x	x	x
3909	INT	x	x	x	x
TG22142	x	x	x	x	x
TG22146	x	x	x	x	x
TG22150	x	x	x	x	x
UH5207	x	x	x	x	x
1096934	x	x	x	x	x
831240	x	x	x	x	x
855125	x	x	x	x	x
UH1752	x	x	x	x	x
ABBL025	x	x	x	x	x
ABBL026	x	x	x	x	x

Supplementary Table 2-C. Results of the refined analysis of presence of genes of interest of the following categories: A. Resistance. B. Virulence. C. Competence. D. Biofilm formation

Table 2-D

Strain	*csuA*	*csuB*	*csuC*	*csuD*	*csuE*	*ompA*	*pgaA*	*pgaB*	*pgaC*	*pgaD*	*bfmS*	*bfmR*
103SM	x	x	x	x	x	x	x	x	x	x	x	x
1096934	x	x	1SNP	x	x	x	1SNP	x	x	x	1SNP	x
14336	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
20C15	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
25C30	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
2MG	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
2RED09	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
3909	x	x	1SNP	x	x	x	1INS (Stop Codon)[#]	x	x	ND*	1SNP	x
5MO	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
61SM01	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
65SM01	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
68SM01	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
72SM01	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
74SM01	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
831240	x	1SNP	2SNP	1SNP	x	x	2SNP	x	x	x	1SNP	x
855125	x	x	1SNP	x	x	x	1SNP	1INS (Stop Codon)[##]	x	x	1SNP	x
96SM	x	x	x	x	x	Triplet INS	x	x	x	x	x	x
ABBL025	x	x	1SNP	x	x	x	1SNP	x	x	x	2SNP	x
ABBL026	x	x	1SNP	x	x	x	ND**	x	x	x	1SNP	x
MGNT	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
MONUR	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
TG22142	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
TG22146	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
TG22150	x	x	1SNP	x	x	x	x	x	x	x	1SNP	x
UH1752	x	x	1SNP	x	x	x	1SNP	x	x	x	1SNP	x
UH5207	x	x	1SNP	x	x	x	1SNP	x	x	x	1SNP	x

* the first 239 bases are not assembled; i.e. the contig ends at base 239 in a +/+ alignment. No evidence of the presence of the other half

** the gene is split in two different contigs, no sign of insertion sequences. Impossible to determine if it is an assembly error

# insertion of an adenine in position 805, Stop codon at base 840

## insertion of a timine in position 1203, Stop codon at base 1227

Please note: all SNPs indicated are mutations of the allele found in strain 103SM

Supplementary Table 2-D. Results of the refined analysis of presence of genes of interest of the following categories: A. Resistance. B. Virulence. C. Competence. D. Biofilm formation

Strain	IS5 ssgr IS903	IS3 ssgr IS51	IS5 ssgr IS427	IS1	ISL3	IS6	IS4 ssgr IS10	IS3 ssgr IS150	IS256	IS3 ssgr IS3	IS91	IS66	IS5 ssgr ISL2	ISNCY
74SM01	28	0	0	0	3	4	9	3	0	0	6	13	0	0
65SM01	18	0	0	0	3	4	2	2	0	0	8	13	0	1
20C15	62	12	0	3	1	26	9	3	0	0	5	6	0	0
72SM01	6	0	0	0	3	25	2	6	0	0	2	5	0	1
14336	20	3	0	2	3	18	10	1	0	0	0	8	0	0
MONUR	5	0	0	0	2	20	11	4	0	0	7	32	0	1
MGTN	20	0	0	0	3	18	10	5	0	0	7	12	0	1
96SM	4	0	0	0	3	32	6	5	0	0	6	41	0	1
103SM	31	0	0	0	3	9	7	2	0	0	0	48	0	1
2MG	11	0	0	0	1	4	4	5	0	1	10	6	0	1
5MO	12	0	0	0	3	5	7	2	0	0	10	10	2	1
3909	2	1	0	2	3	2	2	2	0	0	4	1	0	0
61SM01	17	0	0	0	3	3	5	5	0	0	6	6	0	1
25C30	2	0	0	0	2	19	6	3	0	0	8	9	0	1
68SM01	3	0	0	0	3	12	2	8	0	1	8	5	0	1
2RED09	4	5	0	3	3	32	4	3	0	0	2	6	0	1
TG22146	5	0	0	0	3	7	40	6	0	1	4	4	0	1
TG22142	1	0	0	0	3	7	36	7	0	1	5	43	0	1
TG22150	2	0	0	0	3	1	2	3	0	1	2	2	0	1
831240	3	0	1	0	4	0	0	0	0	2	0	0	0	1
UH5207	3	1	0	0	4	0	0	0	0	1	0	0	1	1
1096934	3	1	0	0	3	0	0	0	0	1	0	0	1	1
855125	1	0	0	0	3	0	0	0	0	4	0	0	1	1
ABBL026	0	0	0	0	3	0	0	0	6	0	0	0	0	0
UH1752	3	1	0	0	4	0	2	0	0	1	0	0	1	1
ABBL025	4	0	0	0	4	1	2	0	0	0	0	0	1	1

Supplementary Table 3. Number of insertion sequences of different classes detected by the platform ISSaga on each genome of the ST78. The number reported is the total number of putative sequence detected by the algorithm

ARTICLE 3

Supplemental Material

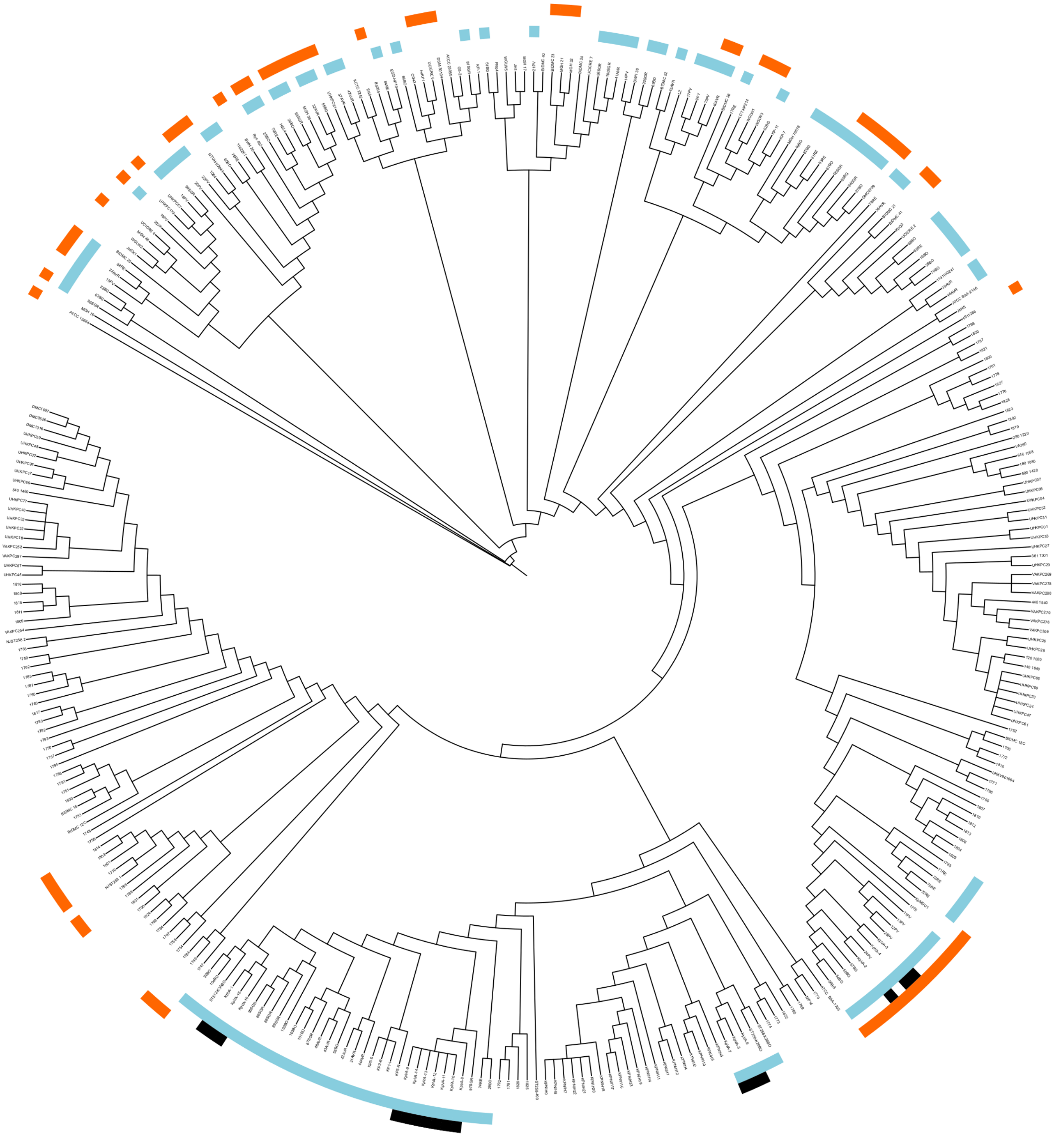


FIG S1. Global Maximum Likelihood phylogeny of 335 genomes of *K. pneumoniae*. Italian strains are indicated in blue, genomes characterized in this study are indicated in black, while strains expressing genes for Yersiniabactin are indicated in orange. Branch lengths are not represented and bootstrap values are not indicated for the sake of image clarity.

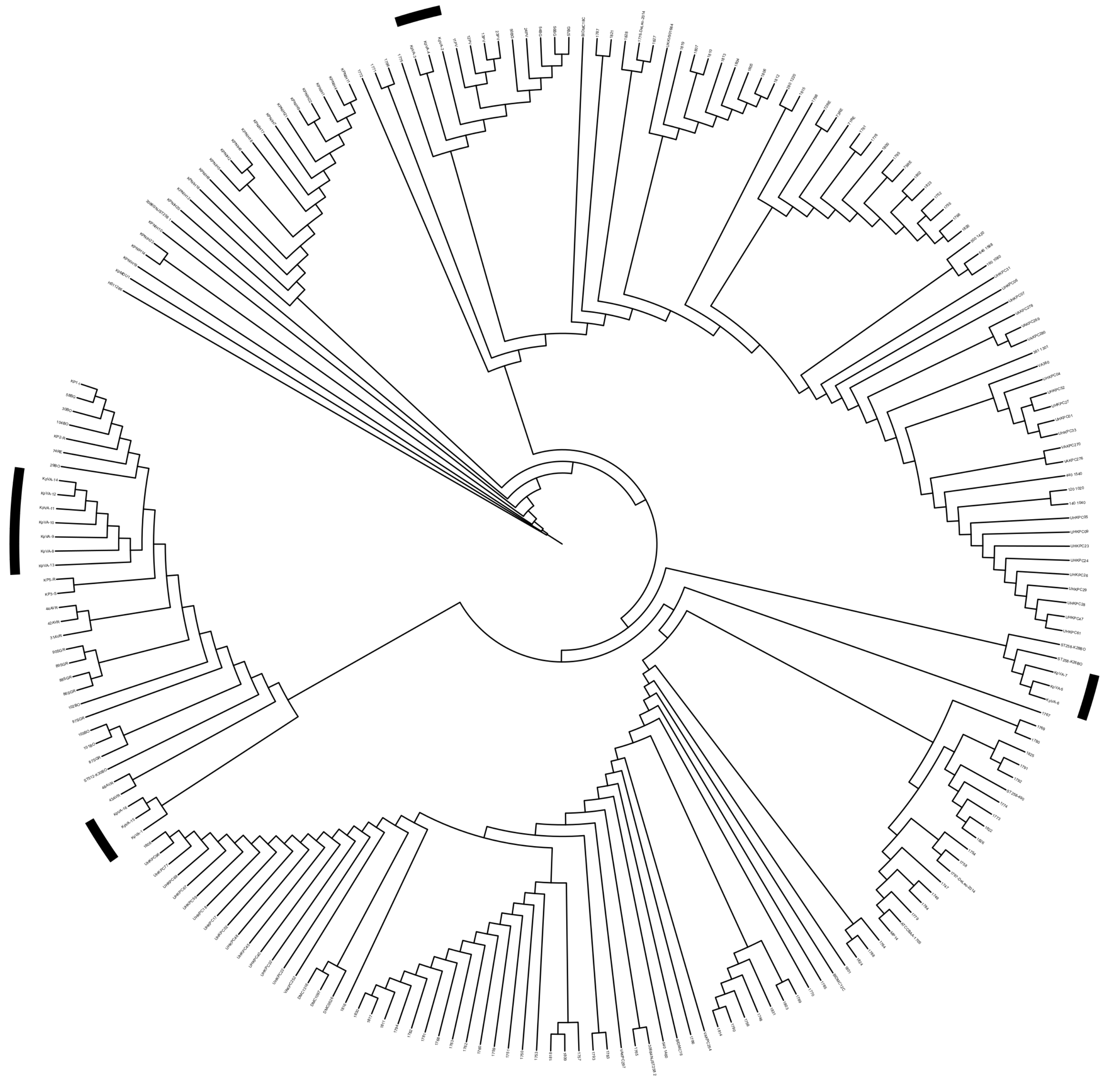


FIG S2. UPGMA tree of Core genome MLST (cgMLST) profiles of 219 genomes belonging to CG258, comprising the 16 novel genomes characterized in this study (indicated in black). Branch lengths are not represented and bootstrap values are not indicated for the sake of image clarity.

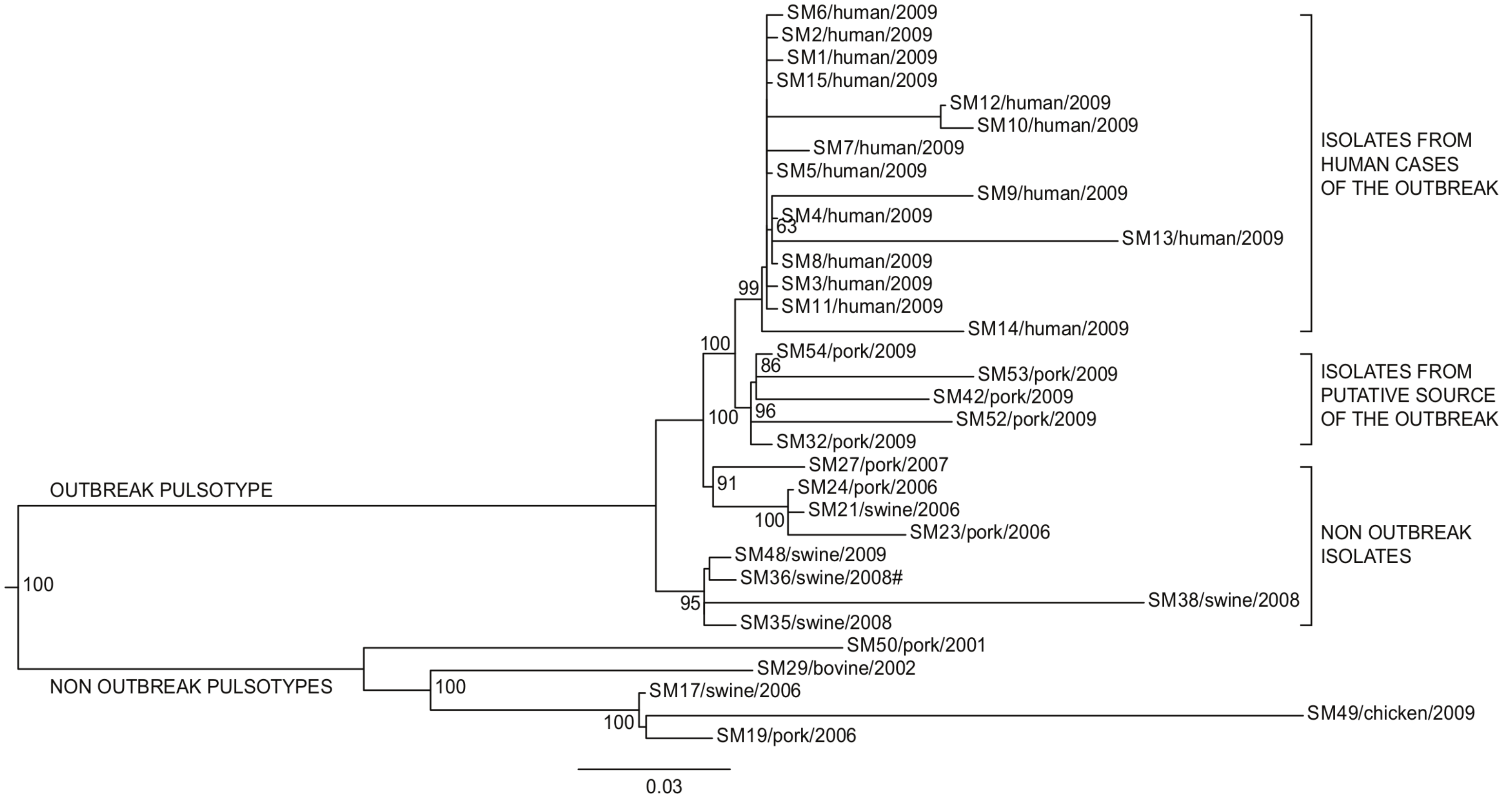
ARTICLE 4

Supplemental Material

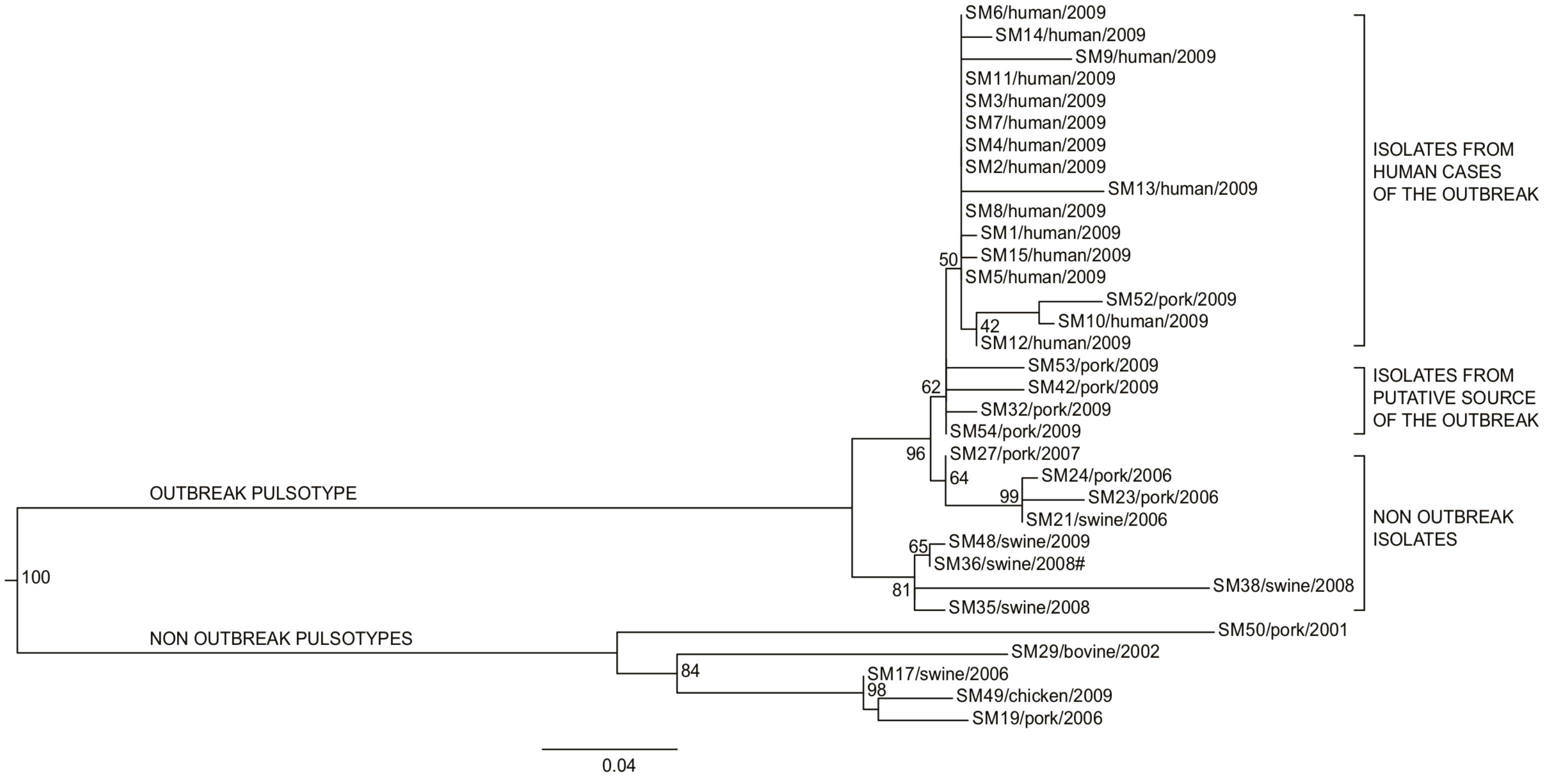
Table S1. Assembly data of *Salmonella* Manhattan isolates sequenced in this study

Lab. N.	Assembly Name	N. of Reads assembled	Avg. Coverage	N. of Contigs	N50	Consensus Sequence (bp)
160969_3	**SM1**	1710982	68.74	156	141280	4719245
160969_5	**SM2**	2535189	93.97	61	276557	4691989
160969_6	**SM3**	2521184	95.83	47	550676	4687862
165051_2	**SM4**	1146836	46.38	141	122404	4721098
165051_3	**SM5**	2819494	104.36	54	432406	4695440
165051_5	**SM6**	830776	32.23	151	76537	4716655
165051_7	**SM7**	2513260	87.78	73	426283	4700219
111113	**SM8**	7263890	237.97	52	653456	4696541
165051_11	**SM9**	705833	22.13	377	22555	4690025
165051_12	**SM10**	4464240	152.62	51	447907	4695242
180073_1	**SM11**	2139784	71.96	70	220086	4695118
180073_2	**SM12**	6155851	228.73	65	502439	4701365
180073_3	**SM13**	550660	19.19	532	15189	4691653
180073_4	**SM14**	635849	20.68	547	17783	4668733
180073_6	**SM15**	1718689	67.07	119	206519	4707922
226963	**SM17**	1115555	45.53	135	94514	4678408
226972	**SM19**	3647977	136.8	132	612314	4692088
226979	**SM21**	2113563	81.76	161	186243	4701558
226985	**SM23**	743857	24.78	370	26335	4647025
226987	**SM24**	8101538	245.95	101	401695	4680260
226998	**SM27**	1083529	43.12	1342	28467	4977060
227009	**SM29**	2745872	115.51	124	239046	4725171
227021	**SM32**	1534324	55.28	135	109153	4703213
227033	**SM35**	4401622	170.21	73	398381	4708271
227039	**SM36**	5195701	198.34	60	441398	4670782
227052	**SM38**	463509	18.26	560	16977	4659968
250920	**SM42**	690442	24.49	353	27141	4693959
188806	**SM48**	2189455	84.38	247	54375	4705944
188795	**SM49**	1110088	38.44	1162	11195	4727662
188781	**SM50**	1887637	61.78	320	51793	4764060
188801	**SM52**	1047164	31.64	166	57534	4690524
216630/1	**SM53**	599806	21.47	464	20152	4696208
216630/2	**SM54**	7964395	244.2	67	472508	4700846

Figure S1. Phylogenetic reconstruction of the 33 *Salmonella* Manhattan isolates based on core (panel A), synonymous (panel B) and non-synonymous (panel C) SNPs datasets, analyzed with Maximum Likelihood method. Bootstrap values are indicated in each principal node of the trees. The scale bar units are nucleotide substitutions per site. # WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).

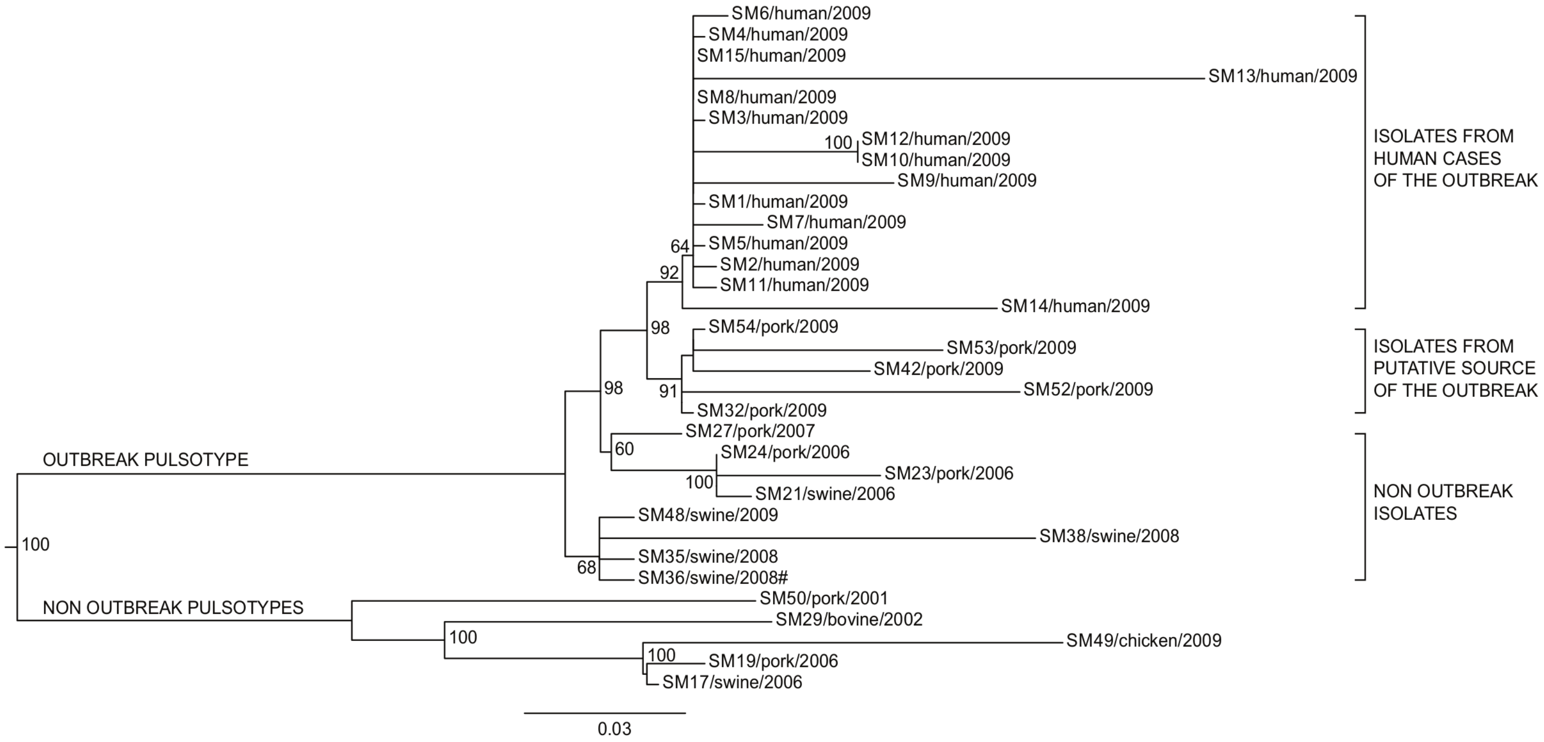


Panel A.



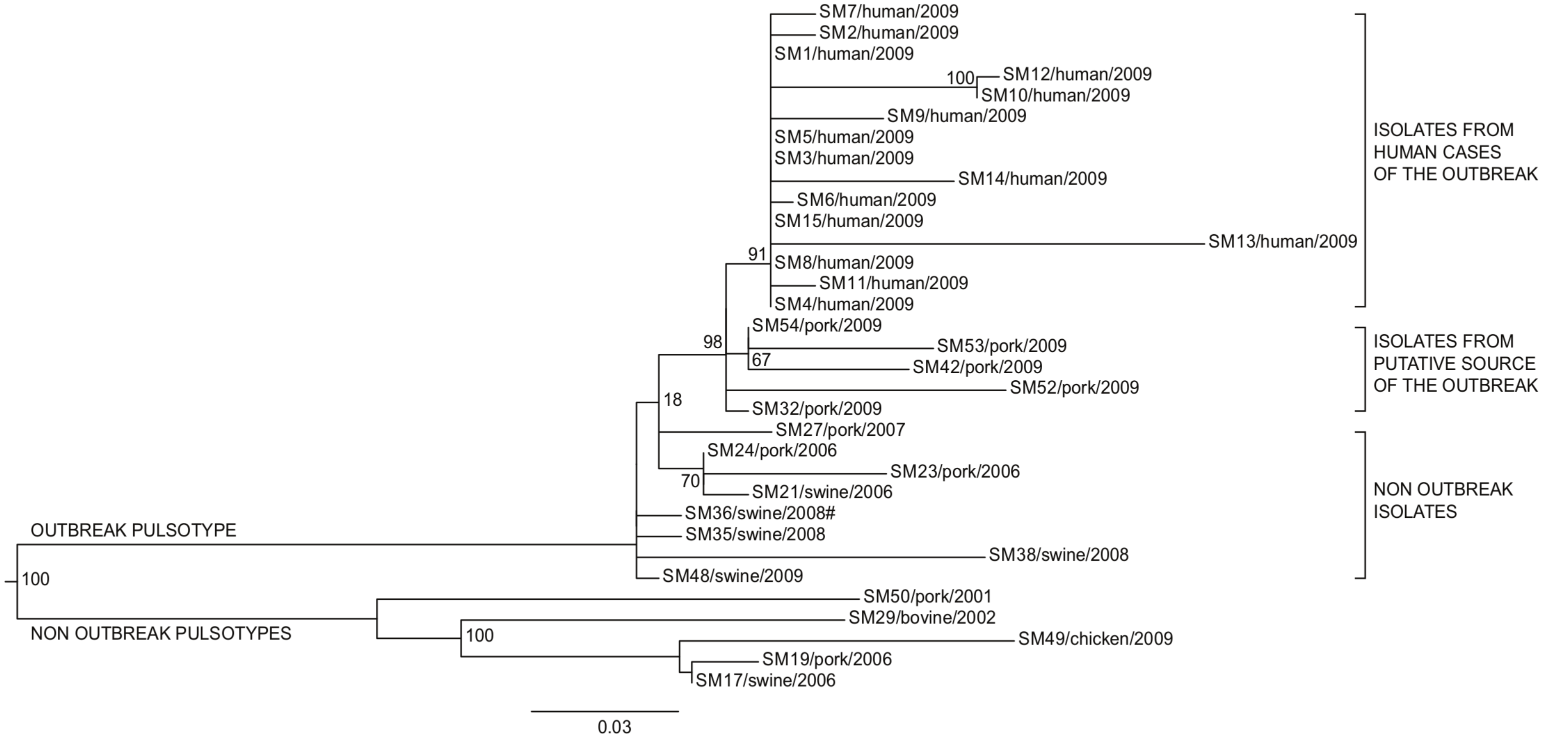
SM6/human/2009
SM14/human/2009
SM9/human/2009
SM11/human/2009
SM3/human/2009
SM7/human/2009
SM4/human/2009
SM2/human/2009
SM13/human/2009
SM8/human/2009
SM1/human/2009
SM15/human/2009
50
SM5/human/2009
SM52/pork/2009
42
SM10/human/2009
SM12/human/2009
SM53/pork/2009
62
SM42/pork/2009
SM32/pork/2009
SM54/pork/2009
96
SM27/pork/2007
64
SM24/pork/2006
99
SM23/pork/2006
SM21/swine/2006
OUTBREAK PULSOTYPE
65
SM48/swine/2009
SM36/swine/2008#
100
81
SM38/swine/2008
SM35/swine/2008
SM50/pork/2001
NON OUTBREAK PULSOTYPES
SM29/bovine/2002
84
SM17/swine/2006
98
SM49/chicken/2009
SM19/pork/2006
0.04
ISOLATES FROM
HUMAN CASES
OF THE OUTBREAK
ISOLATES FROM
PUTATIVE SOURCE
OF THE OUTBREAK
NON OUTBREAK
ISOLATES

Panel B.

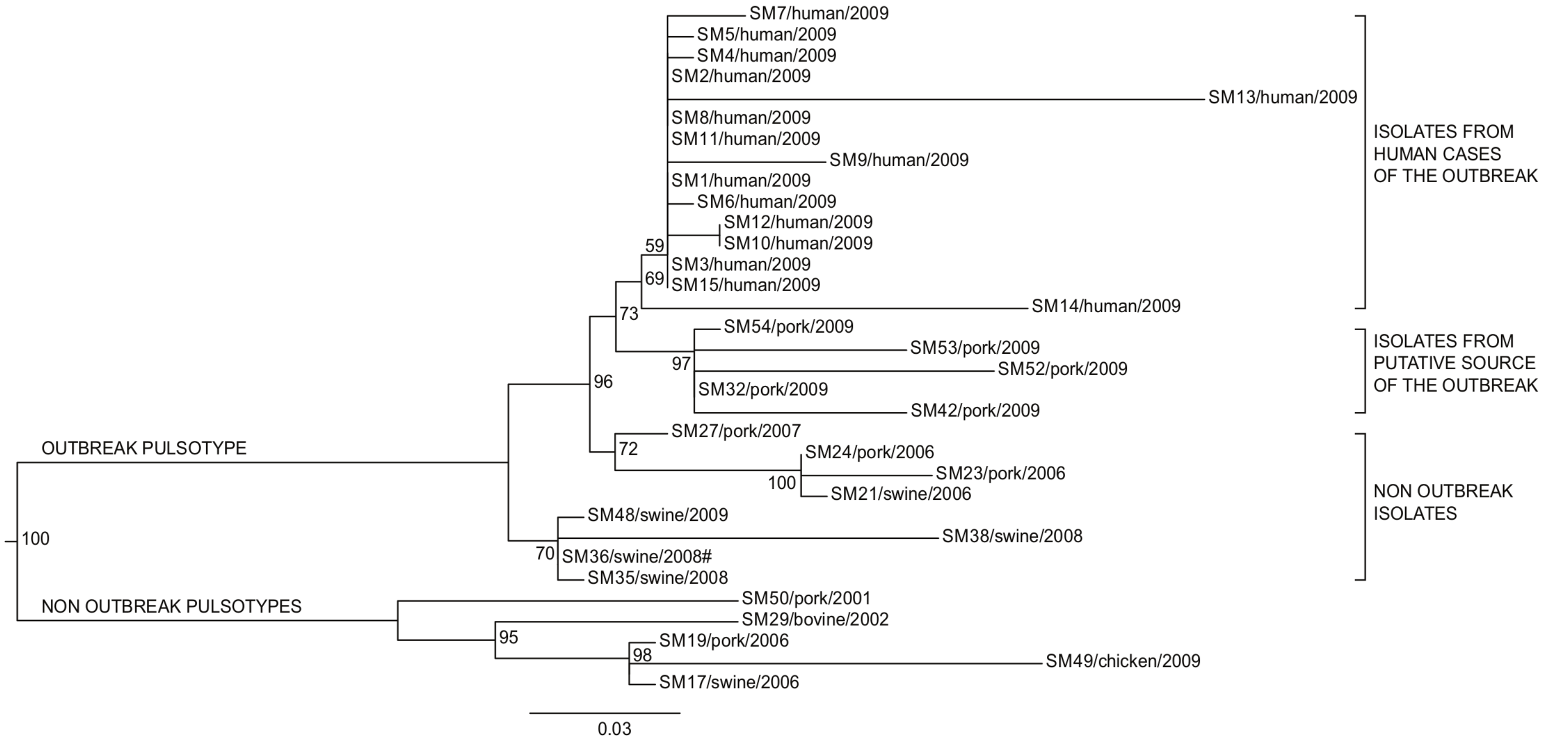

SM6/human/2009
SM4/human/2009
SM15/human/2009
SM13/human/2009
SM8/human/2009
SM3/human/2009
100
SM12/human/2009
SM10/human/2009
SM9/human/2009
SM1/human/2009
SM7/human/2009
64
SM5/human/2009
92
SM2/human/2009
SM11/human/2009
SM14/human/2009
98
SM54/pork/2009
SM53/pork/2009
SM42/pork/2009
98
91
SM52/pork/2009
SM32/pork/2009
SM27/pork/2007
60
SM24/pork/2006
SM23/pork/2006
100
SM21/swine/2006
SM48/swine/2009
SM38/swine/2008
SM35/swine/2008
68
SM36/swine/2008#
OUTBREAK PULSOTYPE
100
NON OUTBREAK PULSOTYPES
SM50/pork/2001
SM29/bovine/2002
100
SM49/chicken/2009
100
SM19/pork/2006
SM17/swine/2006
0.03
ISOLATES FROM HUMAN CASES OF THE OUTBREAK
ISOLATES FROM PUTATIVE SOURCE OF THE OUTBREAK
NON OUTBREAK ISOLATES

Panel C.

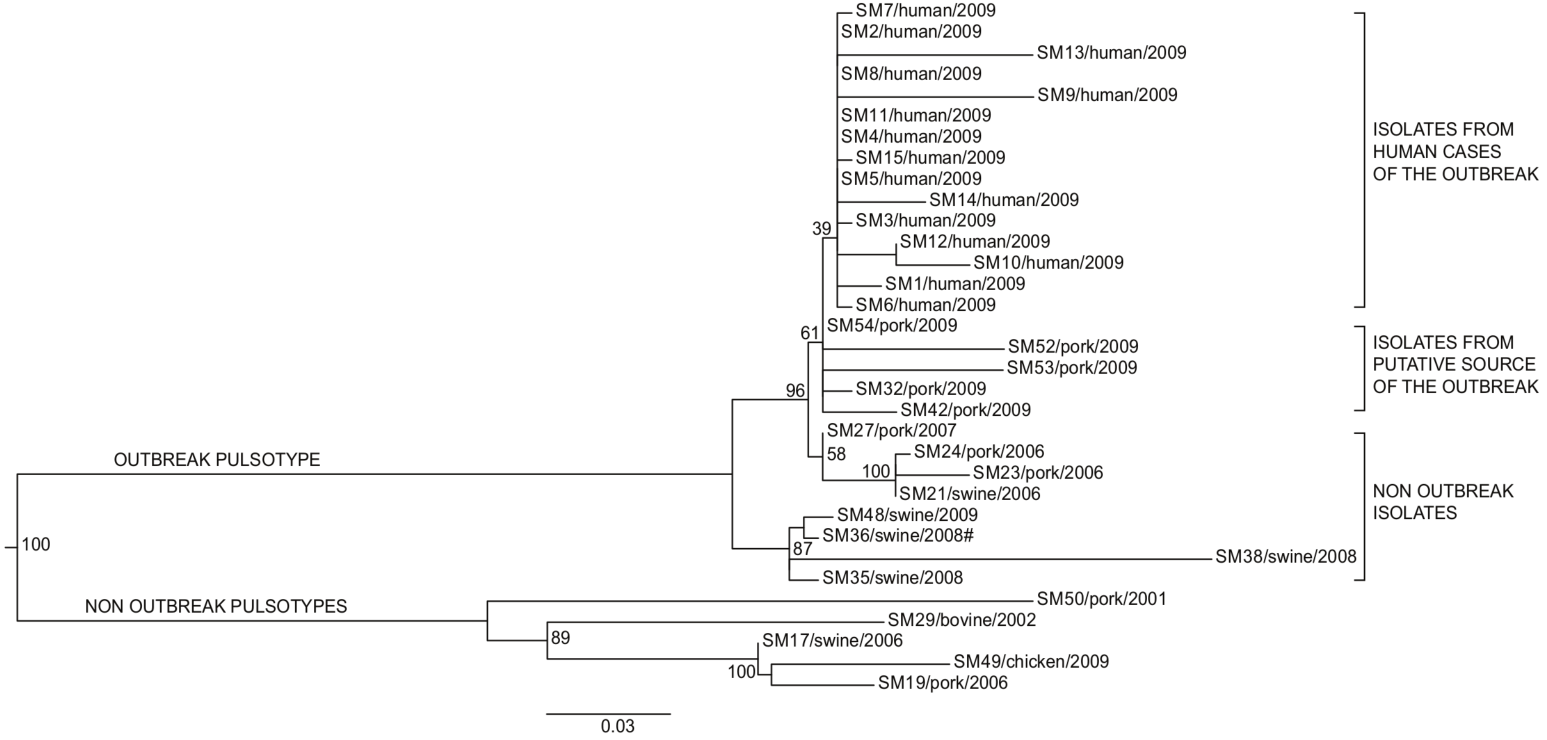
Figure S2: Phylogenetic reconstruction of the 33 *Salmonella* Manhattan isolates analyzed with the Maximum Likelihood algorithm, based on SNPs in first (panel A), second (panel B), third (Panel C) and first+second codon positions (Panel D) datasets. Bootstrap values are indicated in each principal node of the trees. The scale bar units are nucleotide substitutions per site. # WGS analyses clustered isolate SM36 (pulsotype SXB_PR.0752) together with the isolates of the outbreak pulsotype (SXB_BS0003).



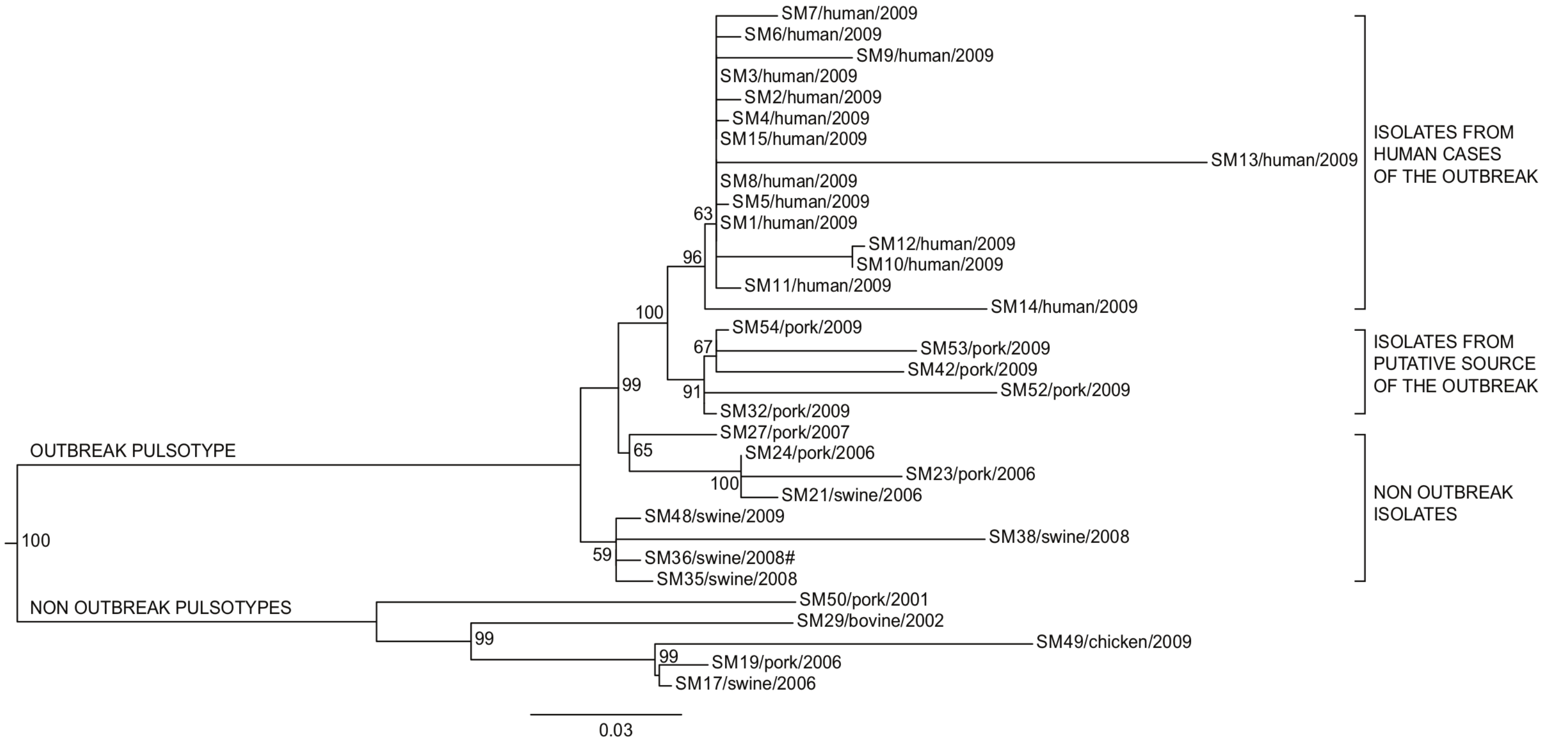
Panel A.


SM7/human/2009
SM5/human/2009
SM4/human/2009
SM2/human/2009
SM13/human/2009
SM8/human/2009
SM11/human/2009
SM9/human/2009
SM1/human/2009
SM6/human/2009
SM12/human/2009
SM10/human/2009
59
SM3/human/2009
69
SM15/human/2009
SM14/human/2009
73
SM54/pork/2009
SM53/pork/2009
97
SM52/pork/2009
96
SM32/pork/2009
SM42/pork/2009
SM27/pork/2007
OUTBREAK PULSOTYPE
72
SM24/pork/2006
SM23/pork/2006
100
SM21/swine/2006
SM48/swine/2009
SM38/swine/2008
100
70
SM36/swine/2008#
SM35/swine/2008
NON OUTBREAK PULSOTYPES
SM50/pork/2001
SM29/bovine/2002
95
SM19/pork/2006
98
SM49/chicken/2009
SM17/swine/2006
0.03
ISOLATES FROM
HUMAN CASES
OF THE OUTBREAK
ISOLATES FROM
PUTATIVE SOURCE
OF THE OUTBREAK
NON OUTBREAK
ISOLATES

Panel B.


SM7/human/2009
SM2/human/2009
SM13/human/2009
SM8/human/2009
SM9/human/2009
SM11/human/2009
SM4/human/2009
SM15/human/2009
SM5/human/2009
SM14/human/2009
SM3/human/2009
39
SM12/human/2009
SM10/human/2009
SM1/human/2009
SM6/human/2009
61
SM54/pork/2009
SM52/pork/2009
SM53/pork/2009
96
SM32/pork/2009
SM42/pork/2009
SM27/pork/2007
OUTBREAK PULSOTYPE
58
SM24/pork/2006
100
SM23/pork/2006
SM21/swine/2006
SM48/swine/2009
100
SM36/swine/2008#
87
SM38/swine/2008
SM35/swine/2008
NON OUTBREAK PULSOTYPES
SM50/pork/2001
SM29/bovine/2002
89
SM17/swine/2006
SM49/chicken/2009
100
SM19/pork/2006
0.03
ISOLATES FROM
HUMAN CASES
OF THE OUTBREAK
ISOLATES FROM
PUTATIVE SOURCE
OF THE OUTBREAK
NON OUTBREAK
ISOLATES

Panel C.


SM7/human/2009
SM6/human/2009
SM9/human/2009
SM3/human/2009
SM2/human/2009
SM4/human/2009
SM15/human/2009
SM13/human/2009
SM8/human/2009
SM5/human/2009
SM1/human/2009
SM12/human/2009
SM10/human/2009
SM11/human/2009
SM14/human/2009
SM54/pork/2009
SM53/pork/2009
SM42/pork/2009
SM52/pork/2009
SM32/pork/2009
SM27/pork/2007
SM24/pork/2006
SM23/pork/2006
SM21/swine/2006
SM48/swine/2009
SM38/swine/2008
SM36/swine/2008#
SM35/swine/2008
SM50/pork/2001
SM29/bovine/2002
SM49/chicken/2009
SM19/pork/2006
SM17/swine/2006
63
96
100
67
91
99
65
100
59
100
99
99
OUTBREAK PULSOTYPE
NON OUTBREAK PULSOTYPES
ISOLATES FROM HUMAN CASES OF THE OUTBREAK
ISOLATES FROM PUTATIVE SOURCE OF THE OUTBREAK
NON OUTBREAK ISOLATES
0.03

Panel D.