THE UNIVERSITY
of EDINBURGH

Population Genomic Analysis of Bacterial Pathogen Niche Adaptation

Rodrigo Bacigalupe

Thesis presented for the degree of Doctor of Philosophy

The University of Edinburgh

2017

I declare that I have composed this thesis, and that the research presented is entirely my own work, including the publications included, unless otherwise stated. This thesis has not been submitted for any other degree or professional qualification.

Rodrigo Bacigalupe August 2017

Acknowledgements

I am most grateful to my supervisor Ross Fitzgerald for his invaluable guidance, support and encouragement throughout my PhD. Thank you for keeping your office door open and providing excellent advice on my research and my career whenever I needed it. I would also like to thank my thesis committee, Mick Watson, Andrew Rambaut, Sam Lycett and Pam Wiener for their constructive suggestions and feedback on my thesis projects. I would also like to thank my former supervisors Álex Mira and Martha Trujillo for initiating me into the exciting field of microbial genomics and encouraging me to pursue a career in science.

Many thanks to all the LBEP members for the help and the fun during the last 4 years, including Emily Parr, Charlotte, Steve, Mariya and Laura. Special thanks to the bioinformaticians Emily Richardson, Paul, Bryan, Rebecca and Gonzalo for the fruitful technical discussions and comments on my analyses. I am also very grateful to Amy, Robyn and Chriselle for helping me in the lab.

In addition, I would like to thank Prof José Penadés, closest collaborator and mentor, for the motivation and opportunity to work on such incredible projects. I would also like to thank other collaborators, especially Geles Tormo, Diane Lindsay, Ewan Harrison, Lucy Weinert, Jukka Corander and members of the SHAC consortium for their help and input on my work. Thanks to Carmen Buchrieser for kindly sending me bacterial genomes and to Antonio Carvajal for his help with GenomePop2.

Many thanks to all my friends in the Roslin Institute and Edinburgh for making my time in Scotland so wonderful, particularly to Hsin-An, Darryl, Selene, Janice, CK, Cansu, Amanda, Irene, Leo and Maica.

Finally, I would like to thank my family and friends for their continuous support, encouragement and interest in my work. This thesis would not have been possible without them. Thank you Papá, Mamá, Paloma, Abuela, Teté, Yoana, Eugenio, Sara and Abel. And thank you Miriam for being always there for me.

Table of Contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of tables

Abbreviations

AIDS	Acquired immune deficiency syndrome
ANI	Average Nucleotide Identity
ANIb	Average Nucleotide Identity based on BLAST
BAPS	Bayesian Analysis of Population Structure
BCYE	Buffered charcoal yeast extract
BEAST	Bayesian evolutionary analysis by sampling trees
BLAST	Basic local alignment search tool
bp	Base pair
CA	Community acquired
CC	Clonal complex
CDC	Centres for Disease Control and Prevention
CDD	Conserved Domain Database
CDS	Coding DNA sequence
CFU	Colony forming unit
COG	Cluster of orthologous group
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CWA	Cell wall-associated
d	Day
dN	Non-synonymous mutation
DNA	Deoxyribonucleic acid
dS	Synonymous mutation
EARSS	European Antimicrobial Resistance Surveillance System
EID	Emerging infectious disease
EMRSA	Epidemic methicillin resistant *Staphylococcus aureus*
ENA	European Nucleotide Archive
ESS	ESAT-6 secretion system
Esx	Ess extracellular
FDR	False Discovery Rate
GATK	Genome Analysis Toolkit
GO	Gene Ontology
GTR	General time reversible
GWAS	Genome-wide association study

h	Hour
HA	Hospital associated/acquired
HGT	Horizontal gene transfer
HKY	Hasegawa Kishino Yano
HPS	Health Protection Scotland
ICE	Integrative and conjugative element
Indel	Insertions and deletions
kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
LD	Legionnaires' disease
LRT	Likelihood ratio test
MALDI-TOF	Matrix-assisted laser desorption/ionization time-of-flight
Mb	Megabase
MCMC	Markov chain Monte Carlo
MGE	Mobile genetic element
min	Minutes
mip	Macrophage infectivity potentiator
ml	Millilitres
MLEE	Multi-locus enzyme electrophoresis
MLST	Multi-locus sequence typing
MRCA	Most recent common ancestor
MRSA	Methicillin resistant *Staphylococcus aureus*
MSSA	Methicillin sensitive *Staphylococcus aureus*
NCBI	National Centre for Biotechnology Information
NSS	neighbour similarity score
PAUP	Phylogenetic analysis using parsimony
PBP2a	Penicillin binding protein 2a
PBS	phosphate-buffered saline
PCR	Polymerase chain reaction
PFGE	Pulsed field gel electrophoresis
PHE	Public Health England
PSA	Proportion of shared ancestry
RAxML	Randomized Accelerated Maximum Likelihood

RMS	Root Mean Square
RNA	Ribonucleic acid
SaPI	Staphylococcal pathogenicity island
SAPIbov	Bovine staphylococcal pathogenicity island
SCC	Staphylococcal cassette chromosome
SCIN	Staphylococcal complement inhibitor
SEP	Staphylococcal enterotoxin P
SEER	Sequence element enrichment analysis
Sg	Serogroup
SNP	Single nucleotide polymorphisms
SRA	Sequence read archive
ST	Sequence type
STAR	*Staphylococcus aureus* repeat
TSA	Tryptic soy agar
TSB	Tryptic soy broth
TSST	Toxic shock syndrome toxin
vWbp	Von Willebrand factor binding protein
w	Week
WGS	Whole genome sequencing

Abstract

Globally disseminated bacterial pathogens frequently cause epidemics that are of major importance in public health. Of particular significance is the capacity for some of these bacteria to switch into a new environment leading to the emergence of pathogenic clones. Understanding the evolution and epidemiology of such pathogens is essential for designing rational ways for prevention, diagnosis and treatment of the diseases they cause. Whole-genome sequencing of multiple isolates facilitating comparative genomics and phylogenomic analyses provides high-resolution insights, which are revolutionizing our understanding of infectious diseases. In this thesis, a range of population genomic analyses are employed to study the molecular mechanisms and the evolutionary dynamics of bacterial pathogen niche adaptation, specifically between humans, animals and the environment.

A large-scale population genomic approach was used to provide a global perspective of the host-switching events that have defined the evolution of *Staphylococcus aureus* in the context of its host-species. To investigate the genetic basis of host-adaptation, we performed genome-wide association analysis, revealing an array of accessory genes linked to *S. aureus* host-specificity. In addition, positive selection analysis identified biological pathways encoded in the core genome that are under diversifying selection in different host-species, suggesting a role in host-adaptation. These findings provide a high-resolution view of the evolutionary landscape of a model multi-host pathogen and its capacity to undergo changes in host ecology by genetic adaptation.

To further explore *S. aureus* host-adaptive evolution, we examined the population dynamics of this pathogen after a simulated host-switch event. *S. aureus* strains of human origin were used to infect the mammary glands of sheep, and bacteria were passaged in multiple animals to simulate onward transmission events. Comparative genomics of passaged isolates allowed us to characterize the genetic changes acquired during the early stages of evolution in a novel host-species. Co-infection experiments using progenitor and passaged strains indicated that accumulated mutations contributed to enhanced fitness, indicating adaptation. Within-host population genomic analysis revealed the existence of population bottlenecks associated with transmission and establishment of infection in new hosts. Computational simulations of evolving genomes under regular bottlenecks supported that the fitness gain of beneficial mutations is high enough to overcome genetic drift and sweep through the population. Overall, these data provide new information relating to the critical early events associated with adaptation to novel host-species.

Finally, population genomics was used to study the total diversity of *Legionella longbeachae* from patient and environmental sources and to investigate the epidemiology of a *L. longbeachae* outbreak in Scotland. We analysed the genomes of isolates from a cluster of legionellosis cases linked to commercial growing media in Scotland and of non-outbreak-associated strains from this and other countries. Extensive genetic diversity across the *L. longbeachae* species was identified, associated with intraspecies and interspecies gene flow, and a wide geographic distribution of closely related genotypes. Of note, a highly diverse pool of *L. longbeachae* genotypes within compost samples that precluded the genetic

establishment of an infection source was observed. These data represent a view of the genomic diversity of this pathogen that will inform strategies for investigating future outbreaks.

Overall, our findings demonstrate the application of population genomics to understand the molecular mechanisms and the evolutionary dynamics of bacterial adaptation to different ecological niches, and provide new insights relevant to other major bacterial pathogens with the capacity to spread between environments.

Lay Summary

Globally disseminated bacterial pathogens frequently cause outbreaks that are of major importance in public health. Understanding the evolution and epidemiology of such pathogens is essential for designing rational ways for prevention, diagnosis and treatment of the diseases they cause. Sequencing and analysing the entire genomes of multiple bacteria facilitates the understanding of the evolution of infectious diseases. In this thesis, several genomic analyses of pathogenic bacteria from different niches were performed to study unresolved aspects of their evolution. First, a global picture of the evolutionary history of *Staphylococcus aureus* was created, providing information of the frequent host-switching events that have occurred during the evolution of this species. In addition, we identified genes and signatures in the genome sequences that can explain adaptation to different host-species, which may represent novel therapeutic targets for controlling human and animal infections. To further investigate the genetic basis of host-adaptation, sheep were infected with *S. aureus* strains from humans, and bacterial isolates passaged in the animals were sequenced. These experiments allowed us to investigate genomic changes acquired during evolution in a new host-species. We identified multiple types of mutations and coinfection experiments of additional sheep with original and passaged strains indicated that such genetic changes were the result of adaptation. Our results were unexpected given the continuous bottlenecks introduced by lambs suckling and transmissions between animals, but computational simulations of bacteria in a similar scenario to the experimental settings revealed that beneficial mutations provide a fitness high enough to get fixed in the population. Finally, the origin of a cluster of diseases caused by *Legionella longbeachae* in Scotland in 2013 was investigated using

genomic approaches. In particular, we studied the evolution and relationships of bacterial isolates from patients and from environmental samples, aiming to identify the source of the infections. Our analysis revealed that a high diversity of bacteria exists in the environmental samples previous establishment of an infection, which complicates the epidemiological investigations if sampling is limited. In addition, genomic analysis showed an extraordinary exchange of genetic material between *L. longbeachae* and other species of the same genus. Overall, our study demonstrates the application of genomics and evolutionary analysis to understand the mechanisms underlying bacterial adaptation to different ecological niches and provide new insights relevant to other bacterial pathogens with the capacity to spread between environments.

1

Introduction

1.1. Bacterial pathogens and emerging infectious diseases

Globally disseminated bacterial pathogens that cause infectious diseases are responsible for millions of deaths every year, representing a major cause of human morbidity and mortality in the world (Binder *et al.*, 1999; WHO, 2017). Over the last few years, multiple infectious diseases are emerging and re-emerging, a trend expected to continue due to a combination of factors, including population growth, socio-economic changes and environmental and ecological conditions (Bloom *et al.*, 2017; Jones *et al.*, 2008). Examples of recent emerging infectious diseases (EID) include the 2010 cholera outbreak in Haiti (Tuite *et al.*, 2011), the foodborne epidemic of *Escherichia coli* O104:H4 in Germany in 2011 (Buchholz *et al.*, 2011), and the invasive non-typhoidal *Salmonella* disease in Africa (Feasey *et al.*, 2012). The emergence and spread of bacterial infectious diseases is unpredictable and has caused substantial impacts on global economies and societies (Bloom *et al.*, 2017). Despite the efforts made by the scientific community and medical specialists in aiming to reduce the impact of these pathogens, inappropriate use of antimicrobial drugs or the intensification of agriculture and farming are complicating this objective enormously (Tomley and Shirley, 2009). One of the main reasons behind the emergence of bacterial infections is the global spread of antibiotic-resistant strains such as epidemic methicillin-resistant *Staphylococcus aureus* (MRSA) (Moran *et al.*, 2005), multidrug-resistant *Streptococcus pneumoniae* (Nuermberger and Bishai, 2004), and vancomycin-resistant *Enterococcus* (O'Driscoll and Crank, 2015), which cause persistent infections difficult to treat. Multidrug-resistant pathogens are considered a serious global health crisis, and new antibiotics are urgently needed (CDC, 2013; Garner *et al.*, 2015).

In addition, many bacterial pathogens are also capable of infecting animals (Woolhouse *et al.*, 2001) and the majority of EIDs are caused by zoonotic pathogens (Palmer *et al.*, 1998; Taylor *et al.*, 2001). Generally, these pathogens can be transmitted by multiple hosts and are known as ecological generalists, as opposed to specialist pathogens, which are restricted to a single host-species (Woolhouse *et al.*, 2001). The capacity of bacteria to cross the host-species barrier represents an additional threat to global health and therefore investigating the factors that contribute to inter-species transmission and human disease emergence is important (Taylor *et al.*, 2001).

Understanding the evolution, natural reservoirs and virulence of bacterial pathogens is essential for epidemiological considerations, designing control and prevention measures, and for the diagnosis and treatment of the diseases they cause (Bentley and Parkhill, 2015). The ongoing development of high-throughput sequencing technologies and associated reduction of sequencing costs is enabling researchers to rapidly sequence the genomes of bacterial isolates (Li *et al.*, 2014). The great amount of data generated, combined with epidemiological and clinical information, are providing high-resolution insights which are revolutionizing our understanding of infectious diseases (Dye, 2014; Firth and Lipkin, 2013).

1.1.1. *Staphylococcus aureus*: A global re-emerging pathogen

Staphylococcus aureus is a Gram-positive bacterium that is part of the commensal flora of humans (Acton *et al.*, 2009; Frank *et al.*, 2010; Nilsson and Ripa, 2006). *S. aureus* is asymptomatically carried in the nares by around 25-30% of healthy people

and intermittently carried by another 60% (Kluytmans *et al.*, 1997; Wertheim *et al.*, 2005). However, *S. aureus* is also an opportunistic pathogen and carriage represents a risk factor of subsequent infection (Casewell and Hill, 1986). If the skin or mucosal membranes are damaged, *S. aureus* is able to invade the tissues and cause a wide range of diseases in humans (Lowy, 1998), which vary from mild skin infections (Frazee *et al.*, 2005) to very acute systemic infections, including bacteraemia, infective endocarditis, osteoarticular infections and pulmonary infections (Tong *et al.*, 2015). In addition, *S. aureus* produces a range of toxins that can cause food-poisoning and toxic shock syndrome after ingestion of contaminated products (Gordon and Lowy, 2008).

Staphylococcal infections commonly require antimicrobial treatment, and the increasing prevalence of resistance to almost all classes of antibiotics represents a major threat to public health (Howden *et al.*, 2014; Stefani and Goglio, 2010). Shortly after the introduction of penicillin in the 1940s, *S. aureus* strains resistant to this antibiotic were reported (Barber and Rozwadowska-Dowzenko, 1948; Kirby, 1944). Similarly, following the introduction of methicillin into clinical practice, methicillin-resistant *S. aureus* (MRSA) were reported in the UK in 1961 (Jevons, 1961), although they had emerged in the mid-1940s, selected by the use of penicillins (Harkins *et al.*, 2017). MRSA strains are particularly successful in developing antimicrobial resistance and have disseminated around the world (Nübel *et al.*, 2010). MRSA-infections were traditionally healthcare-associated (HA-MRSA), representing the leading cause of nosocomial infections (Diekema *et al.*, 1999). Although HA-MRSA is highly prevalent in hospitals worldwide, infection rates vary considerably between different regions, with the highest rates of over 70% in some Asiatic countries (Song *et al.*,

2011) and rates below 5% in some North European countries (Stefani *et al.*, 2012). In the UK, multidrug-resistant MRSA have been endemic in hospitals since the 1990s (Richardson and Reith, 1993) and HA-MRSA infections increased steadily every year until the late 2000s, when this trend shifted and started to decline (Ellington *et al.*, 2009). Nevertheless, HA-MRSA infections affect more than 150,000 patients annually in the European Union and represent an extra-cost of €380 million for the healthcare systems (Kock *et al.*, 2010). The risk factors for acquiring HA-MRSA include previous antibiotic treatment, prolonged hospitalization and intravenous catheters (Lowy, 1998; Raygada and Levine, 2009), and despite interventions for preventing MRSA infections in healthcare settings, some episodes will always likely occur (Török *et al.*, 2014).

Since the beginnings of the 1990s, new MRSA strains associated with the community (CA-MRSA) were described (Herold, 1998), and CA-MRSA infections rapidly increased in prevalence (Dukic *et al.*, 2013). CA-MRSA were mostly associated with children, young patients, militaries, athletes and otherwise healthy people without identifiable risk factors (Gorak *et al.*, 1999; Herold *et al.*, 1998). CA-MRSA strains usually present increased virulence compared to HA-MRSA (Voyich *et al.*, 2006), and often cause severe skin and soft-tissue infections (Boyle-Vavra and Daum, 2007; Frazee *et al.*, 2005). CA-MRSA have disseminated worldwide, representing an endemic problem in many countries (Dukic *et al.*, 2013), with several epidemics reported in the USA and Europe (Landrum *et al.*, 2012; Otter and French, 2010). Although hospital and community associated *S. aureus* strains were considered genetically and clinically distinct (Naimi *et al.*, 2003), the differentiation between them is becoming more ambiguous, with several recent reports describing increasing

transmissions between these two settings (Espadinha *et al.*, 2013; Stryjewski and Corey, 2014).

In addition to humans, *S. aureus* can also colonize and infect a diverse range of animal species, including wild animals, livestock and companion animals (Kloos, 1980; Peton and Le Loir, 2014; Smyth *et al.*, 2009). Some of the most common diseases caused by *S. aureus* in animals include osteomyelitis in poultry (McNamee and Smyth, 2000); exudative epidermidis in swine (Van Duijkeren *et al.*, 2007), skin infections and pododermatitis in rabbits (Corpa *et al.*, 2009), mastitis in sheep, goats and cattle (Holmes and Zadoks, 2011; Menzies and Ramanoon, 2001), all associated with economic loses in the livestock industry. MRSA carriage in companion animals such as cats and dogs has also been reported (Bierowiec *et al.*, 2016; Sasaki *et al.*, 2012), representing a major risk for development of infections by the animals themselves and as reservoirs for transmission into vulnerable people (Loeffler *et al.*, 2011). *S. aureus* has also been isolated from wild animals, including apes, monkeys, bats and marine mammals, which probably act as a reservoir of the pathogen (Monecke *et al.*, 2016; Schaumburg *et al.*, 2012).

During the last decade, the extensive use of antibiotics in farming has led to the widespread emergence of multidrug resistant *S. aureus* strains (Grave *et al.*, 2010), and numerous cases of MRSA of animal origin have been reported in association with human infection (Sakwinska *et al.*, 2011; Voss *et al.*, 2005). These strains are usually referred to as livestock-associated MRSA, and reveal the zoonotic potential of *S. aureus* (Voss *et al.*, 2005). As a generalist pathogen, the capacity of *S. aureus* to infect

multiple host-species and transmit between humans and animals represents a major threat to public health. Therefore, understanding the evolutionary dynamics and molecular basis of host-adaptation is important for implementing control measures to prevent *S. aureus* infections.

1.1.1.1 Genomic landscape of *S. aureus*

The typical *S. aureus* genome consists of a single chromosome of approximately 2.8 Mb in length (ranging from 2.7 Mb to 3.1 Mb), with low GC content (33%) and encoding approximately 2650 genes (Fitzgerald *et al.*, 2001; Holden *et al.*, 2004; Sass *et al.*, 2012). Comparative genomic studies revealed a conserved genome structure, with around 75% of genes shared across all genomes, which make up the core genome of this pathogen (Lindsay and Holden, 2004). In addition, the genome structure is highly conserved, with most strains presenting high levels of synteny between them (Lindsay and Holden, 2006).

The non-core genome or accessory genome of *S. aureus* is largely represented by mobile genetic elements (MGE) (Alibayov *et al.*, 2014; Lindsay and Holden, 2004, 2006), including plasmids, bacteriophages, staphylococcal pathogenicity islands (SaPI), staphylococcal cassette chromosomes (SCC), transposons and insertion sequences (Kuroda *et al.*, 2001). These elements are frequently exchanged between strains through horizontal gene transfer (HGT) and play an important role in staphylococcal evolution and ecological adaptation (Malachowa and Deleo, 2010).

S. aureus may carry one or more plasmids, circular DNA molecules that auto-replicate independently of the main chromosome (Lindsay, 2008). Plasmids are highly variable

in gene content and many encode resistance to antibiotics, metals, antiseptics, virulence factors or toxins (Bayles and Iandolo, 1989; Lindsay and Holden, 2006; Omoe *et al.*, 2003). Plasmids can be classified into three families, depending on their size, gene content and replication mechanism (Lozano *et al.*, 2012; Malachowa and Deleo, 2010; Shearer *et al.*, 2011): small multicopy plasmids (<5 kb) usually carrying one resistance determinant (Khan, 2005); larger low copy plasmids (15-30 kb) that can carry several resistance genes associated with transposable elements (Alibayov *et al.*, 2014); and large conjugative plasmids (up to 60 kb) containing multi-resistance genes (Berg *et al.*, 1998). Plasmids can be transferred vertically or horizontally by transduction or conjugation (Morikawa *et al.*, 2003), and once they have entered a new cell, plasmids can linearize and integrate into the chromosome (Malachowa and DeLeo, 2010).

Bacteriophages are important for the evolution and adaptation of *S. aureus* to different environments (Goerke *et al.*, 2009). Phages can be classified into three groups according to their size, integrase genes and insertion sites (Goerke *et al.*, 2009; Lindsay and Holden, 2004), but the majority of staphylococcal phages belong to the *Siphoviridae* family, of which one is usually at least present (Baba *et al.*, 2002; Brussow *et al.*, 2004; Canchaya *et al.*, 2003). These bacteriophages are around 40 kb and have conserved genes associated with their life cycle and maintenance in the host cell (Goerke *et al.*, 2009; Kwan *et al.*, 2005). However, they also encode virulence determinants, such as the Panton-Valentine leucocidin and staphylococcal enterotoxins such as SEP (Coleman *et al.*, 1989; Zou *et al.*, 2000), or genes encoding immune evasion factors specific for the human innate immune defence (van Wamel *et*

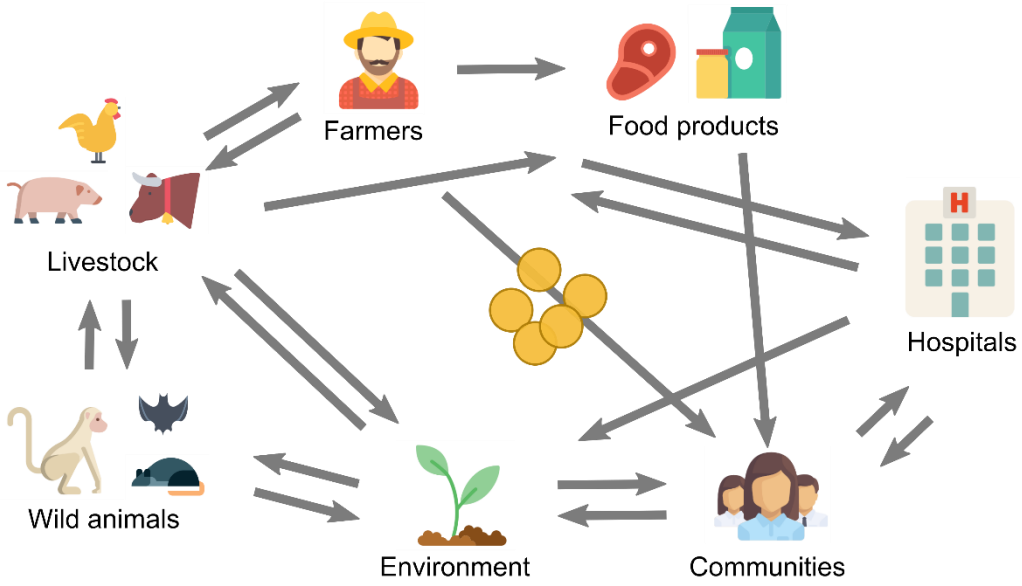
al., 2006). Phages are highly mosaic due to recombination (Kwan *et al.*, 2005), and *S. aureus* genome plasticity is largely the result of the gain and loss of these elements (Goerke *et al*., 2006; Goerke and Wolz, 2004).

Prophages can also participate in the mobilization of other MGEs, such as the helper phage 80α mediating the excision of SaPI1 (Fitzgerald *et al.*, 2001; Novick and Subedi, 2007). SaPIs represent a differentiated type of MGE that range from 15 kb to 17 kb in size, with highly conserved core genes for their own replication (Novick and Subedi, 2007; Ubeda *et al.*, 2007). Additionally, SaPIs encode a series of superantigens, including enterotoxins and the toxic shock syndrome toxins (TSST-1) (Fitzgerald *et al.*, 2001; Yarwood *et al.*, 2002), with the exception of the SaPIbov2 that encodes an adhesion protein important for bovine mastitis infections (Úbeda *et al.*, 2003).

Staphylococcal cassette chromosomes (SCC) are genetic elements ranging from 21 to 67 kb that insert into the *orfX* gene on the chromosome (Boundy *et al.*, 2013) and usually encode antibiotic resistance or virulence factors (Luong *et al.*, 2002). The most typical SCC element carries a methicillin-resistant gene *mecA* (SCCmec), encoding the penicillin binding protein that confers resistance to methicillin and all β-lactam antibiotics (Chambers and Deleo, 2009; Katayama *et al.*, 2000).

In addition, other MGEs exist, such as the integrative and conjugative elements (ICEs) (Smyth and Robinson, 2009) or the genomic islands νSaα and νSaβ, present in the majority of genomes and predicted to have arisen from HGT (Dobrindt *et al.*, 2004; Fitzgerald *et al.*, 2003).

a) *Staphylococcus aureus*



b) *Legionella longbeachae*

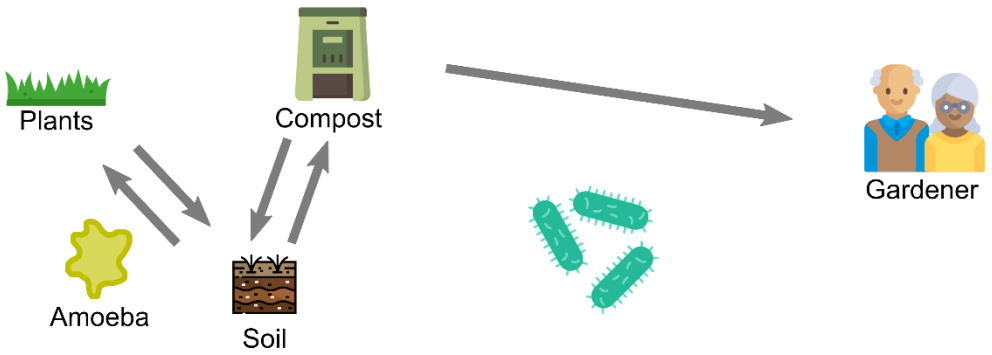


Figure 1.1. Bacterial pathogen transmission between niches. a) *Staphylococcus aureus* colonizes and infects a range of human and animal species, and can spread both into hospital and community associated settings and into the environment. b) *Legionella longbeachae* is present in the soil and growing media products, replicating within environmental protozoa and causing respiratory infections in humans.

1.1.2. *Legionella longbeachae*: A new emerging pathogen in Europe

Legionella longbeachae is a Gram-negative bacillus commonly found in soil and composted waste wood products (Whiley and Bentham, 2011). Similarly to other species of the *Legionella* spp. genus, *L. longbeachae* naturally replicates in a range of environmental protozoa, but when humans inhale contaminated aerosols, the bacteria is able to cause respiratory infections known as legionellosis (Fields *et al.*, 2002).

L. longbeachae was first isolated in 1980 from respiratory tract samples from patients with pneumonia in the USA (Bibb *et al.*, 1981; McKinney *et al.*, 1981). The symptoms of *L. longbeachae* infections are similar to other legionellosis, ranging from an influenza-like illness to severe pneumonia (Amodeo *et al.*, 2010; Cameron *et al.*, 2016), but infections can also be associated with gastrointestinal and neurological symptoms (Amodeo *et al.*, 2010; Mentula *et al.*, 2014). Moreover, *L. longbeachae* can cause extra-pulmonary infections, including endocarditis and sternal-wound infections (Grimstead *et al.*, 2015; Leggieri *et al.*, 2012; Mentula *et al.*, 2014). Although pre-hospital symptoms of *L. longbeachae* and *L. pneumophila* infections do not show any differences (Cameron *et al.*, 2016), *L. longbeachae* diseases are generally less severe (Buchrieser and Hilbi, 2013). In addition, two serogroups have been characterized within *L. longbeachae*, with most of the human infections being caused by Sg1 strains (Cameron *et al.*, 1991).

The risk factors for acquiring *L. longbeachae* legionellosis include smoking, immunosuppression and pre-existing medical conditions (Kenagy *et al.*, 2017; Whiley

and Bentham, 2011). Males over 50 years of age are also more predisposed to develop the infection (Amodeo *et al.*, 2010; Cameron *et al.*, 2016; O'Connor *et al.*, 2007).

Contrary to legionellosis caused by *L pneumophila*, typically linked to artificial water systems (Edelstein, 1982), *L. longbeachae* infections have been associated with exposure to soil-derived products, compost, and potting mixes (Koide *et al.*, 2001; Yu *et al.*, 2002). Microbiological and epidemiological investigations have identified that legionellosis by *L. longbeachae* usually affect keen-gardeners (Potts *et al.*, 2013; Pravinkumar *et al.*, 2010), and a recent survey of bagged potting composts in the UK revealed that 16% of the plant-growing media contained the pathogen (Currie *et al.*, 2013). Additionally, a study in Japan found that 8% of potting mixes tested positive for *L. longbeachae* presence (Koide *et al.*, 2001), and it has also been detected in low proportions in natural soils (Steele *et al.*, 1990) and potting mixes in other countries (Casali *et al.*, 2014; Velonakis *et al.*, 2010). It has been hypothesized that the heat and high humidity conditions reached during the composting process favour its growth (Steele *et al.*, 1990). However, the primary environmental reservoir and the transmission routes of this bacterium remain unknown (Whiley and Bentham, 2011).

The incidence of *L. longbeachae* is variable between counties. In Australia, New Zealand and South East Asia, this pathogen causes legionellosis in similar proportions to *L. pneumophila* (Cramp *et al.*, 2010; Graham *et al.* 2012; Li *et al.*, 2002), while in Europe and the United States, *L. longbeachae* only accounts for less than 5% of reported cases (Joseph *et al.*, 2010; Marston *et al.*, 1994). However, the number of infections in Europe is increasing, with a notable rise in the United Kingdom in recent

years (Cameron *et al.*, 2016; Lindsay *et al.*, 2012; Potts *et al.*, 2013). Here, the first reported case of *L. longbeachae* dates back to 1984, and during the following 3 decades, only a dozen additional cases were identified (Lindsay *et al.*, 2012). However, since 2008, *L. longbeachae* infections have been reported every year in Scotland, with several cases occurring simultaneously in 2013 and 2014 (Potts *et al.*, 2013; Cameron *et al.*, 2016). Public health authorities have associated this increase with a switch in the commercial availability of compost from peat-based growing media to pine sawdust and bark composts, bringing Europe into line with Australia (Whiley and Bentham, 2011). In addition, a series of factors such as the recent climatic conditions, the way growing media products are stored and handled, and the immune status of the patients could also explain the *L. longbeachae* outbreaks in Scotland (Lindsay *et al.*, 2012; Currie *et al.*, 2013; O'Connor *et al.*, 2007).

Of note, the increase in *L. longbeachae* incidence has not been observed in the rest of the UK, probably due to the improved sensitive testing protocols used by the Scottish Reference Laboratories (Potts *et al.*, 2013). Traditional techniques used for diagnosis of *L. longbeachae* and its differentiation from other *Legionella* spp. include culturing, serology, PCR and urinary antigen testing (Cameron *et al.*, 2016). However, community-acquired pneumonia by non-*L. pneumophila* species will continue to be under-diagnosed unless routine PCR tests on respiratory secretions samples from hospitalized patients are incorporated into the surveillance system (Murdoch *et al.*, 2013; Cameron *et al.*, 2016).

In epidemiological investigations, the rapid characterisation of related isolates is essential to identify the source of the infections and subsequent implementation of control measures that prevent future contagions (Reuter *et al.*, 2013). Whole-genome sequencing (WGS) facilitates diagnosis of bacterial pathogens, and its extraordinary discriminatory power (Claudio U. Köser *et al.*, 2012) has been applied to study several *Legionella* outbreaks (Graham *et al.*, 2014; Lévesque *et al.*, 2014; Mcadam *et al.*, 2014; Raphael *et al.*, 2016; Reuter *et al.*, 2013). In addition, WGS provides valuable information on the population structure of pathogens, and the availability of multiple genomes from isolates obtained in different geographic areas in different years can be useful for understanding the evolutionary dynamics of *Legionella* spp. (Rao *et al.*, 2013).

1.1.2.1 Genomic landscape of *L. longbeachae*

The genome of *L. longbeachae* strain NSW150 consists of a main chromosome of 4 Mb and a plasmid of 71 kb, which in total encode for more than 3500 proteins. Compared to a typical *L. pneumophila* genome, *L. longbeachae* is 500 kb larger, and around 65% of the genes are homologous between the two species (Cazalet *et al.*, 2010). Only five *L. longbeachae* strains have been sequenced to date (Cazalet *et al.*, 2010; Gomez-Valero *et al.*, 2011; Kozak *et al.*, 2010), and comparative genomic analysis with *L. pneumophila* genomes revealed the specific genetic features of *L. longbeachae* that reflect its niche adaptation (Cazalet *et al.*, 2010). Of note, *L. longbeachae* genomes contain several genes encoding proteins found in plant pathogenic bacteria, including enzymes capable of degrading vegetable material (Cazalet *et al.*, 2010). A major difference with *L. pneumophila* is that *L. longbeachae*

does not encode flagella, which may explain differences in mouse susceptibility to infection between these two species (Cazalet *et al.*, 2010, Kozak *et al.*, 2010). In addition, global gene expression profiles revealed differences in the life cycles of the two species, with *L. longbeachae* presenting a more subtle environmental-protozoa biphasic life cycle (Cazalet *et al.*, 2010).

A common feature of *Legionella* genus genomes is the presence of genes encoding many secreted virulence factors, known as effectors, which manipulate biological processes within host cells (Burstein *et al.*, 2016). These effectors are translocated into the cells via the Icm/Dot secretion system (Isberg *et al.*, 2009; Segal *et al.*, 1998), which is highly conserved among *Legionella* species (Feldman *et al.*, 2005). The effectors are variable between different species, and many of those uniquely encoded by *L. longbeachae* genomes have eukaryotic domains or resemble eukaryotic proteins that can mimic the activities of the host molecules and disrupt their functions (Cazalet *et al.*, 2010).

Comparative genomic analysis of *L. pneumophila* Sg1 genomes revealed that recombination and horizontal gene transfer between strains of the same and different species have played a major role in the evolution of this genus (Gomez-Valero *et al.*, 2011). Exchange of plasmids between strains has also been an important strategy for genome diversification and a series of plasmids have circulated between distinct *Legionella* species (Gomez-Valero *et al.*, 2011).

1.2. Population genomics of infectious diseases

In order to understand the evolutionary and epidemiological processes associated with the emergence of bacterial pathogens, we must consider the natural genetic variation existing within natural populations (Spratt and Maiden, 1999). The field of population genetics examines allele frequencies in populations and the evolutionary factors that affect their distribution over space and time (Lewontin, 1985). Population genetics originated in the first half of the 20th century and was driven by theoretical insights with limited empirical data. However, with the falling costs of next-generation sequencing technologies, it expanded from the analysis of a few specific genes to the study of entire genomes, giving rise to the emerging area of population genomics (Black *et al.*, 2001; Whitaker and Banfield, 2006).

Thus, population genomics refers to the analysis of whole genome sequences to study the evolutionary forces that influence genetic variation across populations (Gulcher and Stefansson, 1998). Understanding the effects of mutations, selection, gene flow and recombination on the genetic diversity of bacterial populations provides information of the pressures underlying bacterial adaptation (Perez-Losada *et al.*, 2006). In addition, population genomics examines the effects of forces influencing genomes at population level, including population bottlenecks and genetic drift, which are indicative of demographic patterns (Black *et al.*, 2001; Guttman and Stavrinides, 2010). Population genomics can provide fundamental insights into the evolution of pathogens, their population dynamics, transmission patterns, and genotype-phenotype associations, which are critical for the development of effective public health control strategies (Pérez-Losada *et al.*, 2006).

1.2.1. Typing and genetic diversity of bacteria

The ability to differentiate and classify bacterial isolates into sub-groups or types is an important step to understand their evolution and the epidemiology of the infectious diseases they cause. Combining this information with temporal, spatial and clinical data permits us to infer the geographical spread of pathogens and disease associations, which is needed to identify their sources and prevent further infections (Bentley and Parkhill, 2015).

Traditionally, typing methods of bacteria relied on the characterization of phenotypic properties, including serotype, bacteriophage type, antimicrobial susceptibility or biotype (Foxman *et al.*, 2005; Tenover *et al.*, 1997). Molecular typing techniques such as multilocus enzyme electrophoresis (MLEE) (Selander *et al.*, 1986) and pulse-field gel electrophoresis (PFGE) (Hennekinne *et al.*, 2003) were also widely used to investigate the genetic structure of natural populations of pathogens (Fitzgerald *et al.*, 1997; Hartl and Dykhuizen, 1984; Turabelidze *et al.*, 2000). Although these methods provided valuable insights for understanding the diversity of pathogen populations, they were limited in terms of the reproducibility achieved and in many cases they did not provide enough resolution to discriminate between bacterial lineages (Hunter, 1990).

With advances in DNA sequencing methods, new typing schemes based on the sequence variation were developed (Bentley and Parkhill, 2015). Among these, the most popular approach was multilocus sequence typing (MLST), which consists of sequencing a small number of loci, typically four to seven (Maiden *et al.*, 1998). MLST

represented an enormous improvement compared with previous techniques, as it captures the ancestral relationships between the isolates and therefore provides a better resolution (Vinatzer *et al.*, 2014). In addition, the information can be uploaded into databases accessible by the research community and MLST schemes for around 130 microbial taxa have been developed (pubmlst.org/databases.shtml). For *S. aureus*, MLST involves the sequencing of 7 house-keeping genes, and isolates are assigned to one of over 2200 sequence types (ST), which in turn can be grouped into clonal complexes (CC) (Enright *et al.*, 2000). MLST has proven to be effective for studying *S. aureus* evolution, virulence and antibiotic resistance (Turner and Feil, 2007). It was used to classify more than 3000 clinical and veterinary isolates of MRSA from different countries, which represented a vast diversity of *S. aureus* strains and provided an overview of pandemic, epidemic and sporadic strains (Monecke *et al.*, 2011). MLST has also been used to determine the level of clonality of *S. aureus*, the distribution of the disease-causing strains and the significance of homologous recombination in the diversification of this pathogen (Feil *et al.*, 2003).

Although this methodology has been essential for investigating multiple aspects of bacterial populations, the resolution provided by MLST schemes is still limited to fully comprehend the evolution, population structure and phylogeography of pathogens (Achtman, 2008). With the decreasing costs of WGS, sequencing of hundreds and thousands of bacterial isolates is now possible, which allows the study of bacterial populations at an extremely detailed resolution.

1.2.2. Population structure of bacterial pathogens

The population structure of bacterial pathogens is shaped by different evolutionary processes, including those that generate genetic variation and those that modulate the frequency of such variation (Andam *et al.*, 2017). Bacteria are primarily clonal, as they reproduce by cell division, with the descendant genomes being strictly identical to the parental genome (De Meeûs *et al.*, 2007). Mutations are the ultimate source of genetic variation, and if not purged, they may spread and fixate in the population (Barrick and Lenski, 2013). The majority of mutations are neutral or deleterious (Loewe and Hill, 2010), but long-term adaptation is driven by beneficial mutations, which improve the fitness of an organisms in a particular environment (Nielsen, 2005). Alleles conferring an advantage increase in frequency, and if several appear at the same time, they compete for fixation in a process known as clonal interference (Biek *et al.*, 2015). On the contrary, deleterious mutations influence the fate of the beneficial mutations, and reduce the adaptation rate (Charlesworth *et al.*, 1993).

Natural selection favours the appearance of fitter genotypes, with beneficial mutations sweeping through a population and leading to the loss of other genotypes (Levin, 1981). This process is referred to as periodic selection, and as a consequence, evolving lineages present sets of mutations associated in higher frequencies than expected by chance. This generates patterns of chromosomal polymorphism (linkage disequilibrium), which are a clear indication of clonal evolution (Smith *et al.*, 1993; Tibayrenc and Ayala, 2012). Some species of pathogens are highly clonal, such as *Mycobacterium tuberculosis*, *Bordetella pertussis* or *Yersinia pestis*, which are

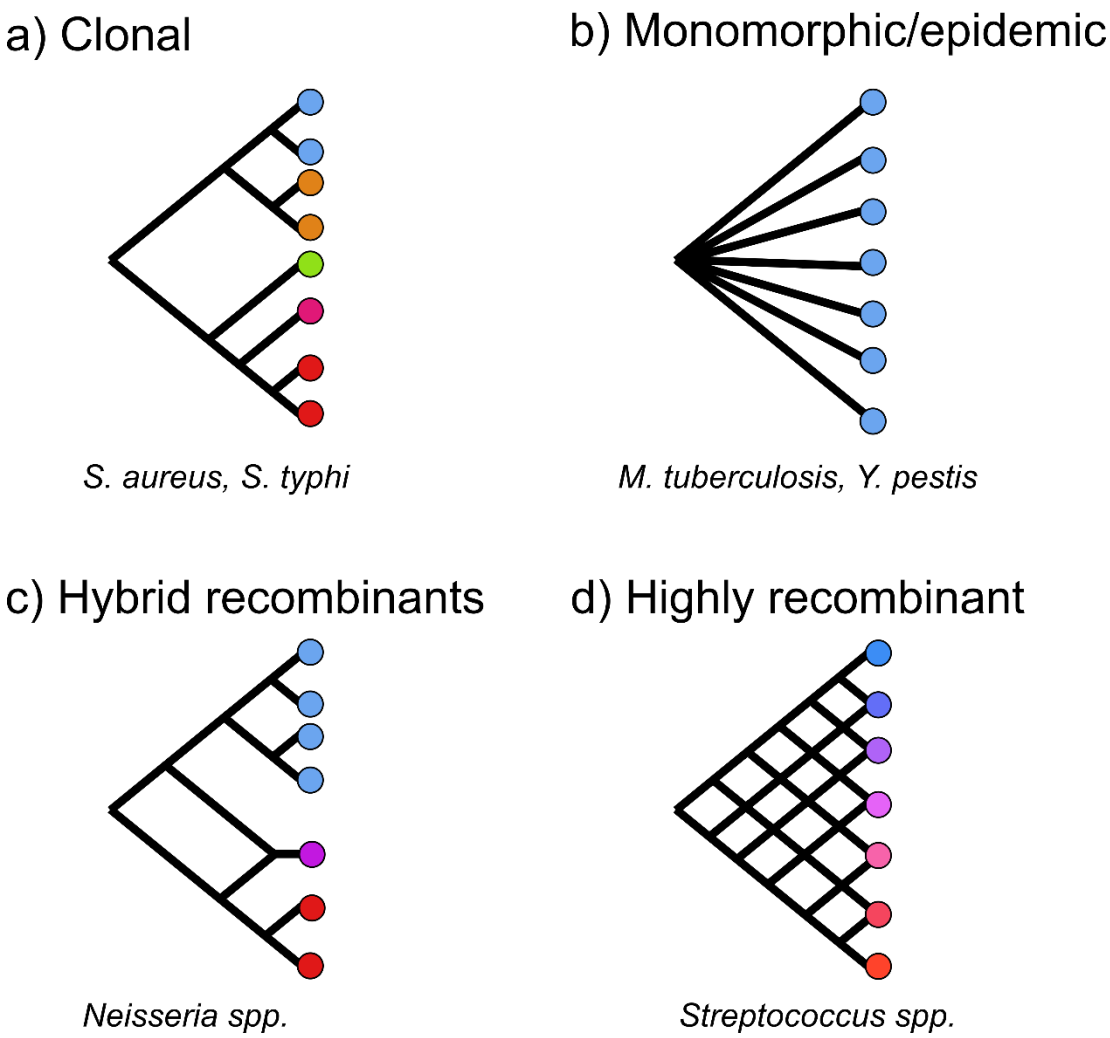


Figure 1.2. Population structures of bacterial pathogens. a) Clonal population structures contain distinguishable lineages. b) Monomorphic/epidemic populations are genetically uniform and usually emerge as independent lineages due to bottlenecks or selective sweeps. c) Recombination events between two different lineages can lead to hybrid clones. d) Highly recombinant species with strains exchanging genetic material with other members of the population result in reticulated genealogies.

genetically uniform and emerged as independent lineages from the accumulation of mutations from other species (Figure 1.2b) (Lan and Reeves, 2001; Spratt, 2004). Another common feature of clonality is the presence of distinguishable clones within the population and some bacterial pathogens such as *S. aureus* or *Salmonella typhi* are made of populations that consist of independent evolving lineages that represent quite stable clones (Figure 1.2a) (Baker *et al.*, 2010; Didelot *et al.*, 2011). These clones are groups of isolates that descend from a recent common ancestor and usually present similar biological features, such as specific patterns of virulence or host-specificity (Spratt, 2004). For example, in the case of *S. aureus*, studies have shown that many of the isolates belong to host-specific clones (sequence types or clonal complexes) (Guinane *et al.*, 2010; Lowder *et al.*,2009; Smyth *et al.*, 2009; Fitzgerald, 2012). Similarly, distinct enterotoxigenic *Escherichia coli* clades have been associated with specific virulence profiles (von Mentzer *et al.*, 2014).

1.2.3. Recombination and horizontal gene transfer

Another molecular process underlying the emergence of variation in bacteria is recombination. Recombination is usually associated with the horizontal transfer of DNA fragments between a donor and a recipient bacteria, which can take three forms: conjugation (plasmid exchange), transduction (mediated by bacteriophages) and transformation (absorption of DNA from the environment) (Prugnolle and de Meeûs, 2011). Horizontal gene transfer (HGT) is the principal mechanism of genetic exchange for most bacterial pathogens and is responsible for the transfer of adaptive mutations through a population in a similar way to sexual recombination (Narra and Ochman, 2006; Ochman *et al.*, 2000).

HGT and recombination can occur between distant taxa or closely related organisms, leading to the acquisition of new sets of genes or to the generation of novel combinations of alleles that facilitate adaptation of bacteria to new environments (Juhas, 2013; Treangen and Rocha, 2011). For example, *Streptococcus pneumoniae* exchanges genes with other streptococcal species sharing the same niche, such as *Streptococcus mitis*, *Streptococcus infantis* and *Streptococcus oralis*, and this represents the main evolutionary mechanism of this pathogen to rapidly adapt to selective pressures (Figure 1.2d) (Chaguza *et al.*, 2015; Donati *et al.*, 2010). In addition, high levels of homologous recombination have been associated with the emergence of nosocomial pathogens such as *Enterococcus faecium* (Arias and Murray, 2012), which also demonstrate a great capacity to gain genetic elements carrying antimicrobial resistance determinants (van Hal *et al.*, 2015).

Recombination has played an important role in the emergence of new pathogens (Takuno *et al.*, 2012). In bacteria with high levels of recombination, mosaic genotypes are frequently observed, resulting from the recombination of genomic fragments between two different species (Figure 1.2c). For example, some *Neisseria* spp. isolates cannot be clearly classified into *Neisseria meningitidis* or *N. lactamica*, as they contain gene fragments from both species (Corander *et al.*, 2012). In *S. aureus*, the hybrid clones ST239 and ST71 originated from large recombination events where strains from specific STs acquired large DNA fragments from others (635 kb and 329 kb respectively) (Holden *et al.*, 2010; Spoor *et al.*, 2015). Of note, hybrid clones may acquire new pathogenic traits such as immune evasion genes, antibiotic resistance determinants or host-adaptive factors (Spoor *et al.*, 2015). Although *S. aureus* is

considered to be largely clonal (Fei *et al.*, 2003), recent studies have shown that the recombination rate is much higher than previously estimated (Takuno *et al.*, 2012), with mobile genetic elements driving recombination hotspots in the core genome (Everitt *et al.*, 2014).

HGT is one of the major evolutionary processes shaping bacterial population structure and knowing the level of recombination for a given species is essential to understand its evolutionary dynamics (Tibayrenc and Ayala, 2012). Low levels of recombination will result in highly clonal populations, with well-defined lineages over time scales. On the contrary, highly recombinogenic pathogens are associated with fuzzy and diversifying species (Spratt and Maiden, 1999; Corander *et al.*, 2012). This is particularly important in epidemiological studies for identifying the source of bacterial infections and transmission chains. For example, genomic investigations of a legionellosis outbreak in Edinburgh by *L. pneumophila* revealed that infections were caused by multiple genetic subtypes of this pathogen, the majority of which had diversified from a single progenitor through mutation, recombination, and horizontal gene transfer, making the source attribution challenging (McAdam *et al.*, 2014).

1.3. Adaptive evolution to new niches: host adaptation

Population genomic analyses of pathogenic bacteria have provided substantial contributions to our understanding of how pathogens adapt to specific ecological niches and in particular to different host-species. Although some bacteria are specialized to colonizing only a single host-type, many pathogens can live in multiple host-species, providing opportunities for expansion into new host populations and thus

representing a major threat to global health (Jones *et al.*, 2008). Notably, the majority of emerging human infectious diseases are caused by pathogens that also infect animals (Taylor *et al.*, 2001) and most of them have been traced to an animal origin (Wiethoelter *et al.*, 2015; Woolhouse and Gowtage-Sequeria, 2005). The transmission of pathogens from other species into human populations is a natural consequence of the capacity of microbes to exploit new niches and adapt to new hosts (Karesh *et al.*, 2012). The ecological and demographical factors associated with the domestication of animals in the Neolithic period and the more recent agricultural industrialization and globalization have facilitated the transmission of pathogens between humans and animals. Among these factors, the large density of populations and high frequency of contacts between humans and animals have contributed significantly (Hudson *et al.*, 2008; Jones *et al.*, 2013; Taylor *et al.*, 2001; Karesh *et al.*, 2012).

Expansion into a new host-species involves multiple steps, starting with the invasion of the new host, followed by transmission within the new host-species, and finally the establishment of the bacteria in the new host population (Wolfe *et al.*, 2007). Many infections transmitted from animals to humans have limited subsequent person-to-person transmissions (spillover events), but pathogens with capacity to spread efficiently between people can give rise to local and global outbreaks (Bentley and Parkhill, 2015).

Numerous studies have made efforts to trace and quantify the dynamics of human diseases with zoonotic origins (Wolfe *et al.*, 2007; Woolhouse *et al.*, 2005; Jones *et al.*, 2013). In addition, WGS of pathogenic bacteria isolated from different host-

species have revealed multiple genetic changes associated with adaptation to specific host-types, including single nucleotide SNPs in *S. aureus* adaptation to rabbits (Viana *et al.*, 2015), gene inactivation in *Salmonella* adaptation to chickens (Foley *et al.*, 2013), pseudogenization in *Streptococcus agalactiae* for adaptation to the bovine udder environment (Almeida *et al.*, 2016), gene loss in various subspecies of *Salmonella enterica* for adaptation to different host-types (Langridge *et al*., 2015), and gene acquisition in pathogenic strains of *Leptospira* for adaptation to animals (Xu *et al.*, 2016). Gain and loss of genes is a commom mechanisms for bacteria to rapidly acquire novel functional traits and get rid of unnecesseary ones, driving their adaption to the environmental requirements determined by the niches associated with distinct host-types, dependent on specific host diets (Sheppard *et al.*, 2013), host-susceptibility to certain virulence factors (Langridge *et al.*, 2015), among others. Prophage mobilization plays an important role in the exchange of genes leading to host adaptation, for example in *Lawsonia intracellularis* adaptation to pigs (Vannucci *et al*., 2013) or *E. coli* adaption to cows (Lupolova *et al.*, 2016). Acquisition of new genes can also result from gene duplication, such as in *Streptococcus equi* adaptation to persistent infection within horses, achieved by copy number variation due to homologous recombination between insertion sequence (IS) elements (Harris *et al*., 2015). In addition, large-scale genomic rearrengments such as choromosmal inversions have also driven host-adaptation, for example in *Salmonella enterica*, with different serovars presenting variable rates of genomic inversion depending on their specific hosts (Matthews *et al*., 2010). Finally, rapid phenotypic switching enabling infections in novel hosts can be acquired through epigenetic events such as DNA methylation (Van der Woude, 2011). Nevertheless, given its clinical relevance and the

wide array of host-species it is able to colonize, *S. aureus* is an excellent model for examining host-adaptation.

1.3.1. Population genomics of *S. aureus* host-species association

S. aureus has a clonal population structure defined by lineages that have single or multiple host-tropisms (Feil *et al.*, 2003; Shepheard *et al.*, 2013; Weinert *et al.*, 2012). The majority of strains associated with animals belong to unique STs different from those infecting humans, and rarely cross species boundaries (Shepheard, 2013; Fitzgerald, 2012). For example, *S. aureus* infecting cows, sheep and goats, are typically represented by the CC97, CC130, CC133 and ST151 and ST771 (Sung *et al.*, 2008; Smyth *et al.*, 2009; Fitzgerald, 2012). In contrast, strains infecting poultry mainly belong to the ST1, ST5, ST398 and CC385 (Lowder *et al.*, 2009; Smyth *et al.*, 2009) and rabbits are primarily associated with the ST121 and ST96 (Guerrero *et al.*, 2015; Vancraeynest *et al.*, 2006). Several investigations have shown that certain host-species may present host-specific barriers to colonization by other host-restricted *S. aureus* strains (Lowder *et al.*, 2009; Viana *et al.*, 2015).

Conversely, some STs and CCs contain strains that have a broader host range and are able to colonize humans and other animals. Some of these include cases of zoonotic transmission from livestock to humans and from humans to livestock, which have been increasingly documented during the last decade (Price *et al.*, 2012; McCarthy *et al.*, 2011; Shepheard *et al.*, 2013). For example, the livestock-associated methicillin-resistant CC398, a lineage prevalently associated with pigs (Hasman *et al.*, 2010), was recognized among swine farmers in the early 2000s (Voss A, 2005) and a few years

later also in their close contacts (Price *et al.*, 2012; van Cleef *et al.*, 2011). Nevertheless epidemiological data indicated that these strains had low onward transfer and virulence rates (Price *et al.*, 2012). Similarly, Sakwinska and colleagues reported the emergence of bovine mastitis caused by strains belonging to a CC8 lineage that had been traditionally associated with humans (Sakwinska *et al.*, 2011). The majority of cases where a *S. aureus* strain is found to infect a different host-species probably represent single episodes of zoonotic transmission, transient spillover events that disappear without establishing additional transmissions within the new host-species (McCarthy *et al.*, 2011; Shepheard *et al.*, 2013).

However, in some cases *S. aureus* infecting a new host-species ends up establishing within the new host population, and several of these host-switches have been reported recently (Resch *et al.*, 2013; Spoor *et al.*, 2013). For example, Lowder and colleagues showed most of the ST5 *S. aureus* isolates from chickens have their origin in a single human-to-poultry host-jump that took place 40 years ago (Lowder *et al.*, 2009). More recently, Spoor *et al.* (2013) demonstrated that two human community acquired-MRSA CC97 clones originated from independent host-jumps from cattle; and the CC130 MRSA lineage associated with cows has also spread into humans recently (Harrison *et al.*, 2013). In addition, Weinert and colleagues reconstructed the frequency and timing of host-switches over a long-time period, demonstrating that multiple host-jumps back and forth between human and bovids (including cows, sheep and goats) have occurred, the first one about 5500 years ago (Weinert *et al.*, 2012).

1.3.2. Genetic basis of host-specificity

Population and comparative genomic investigations have provided increasing resolution for understanding the genetic basis of *S. aureus* host-specificity (Fitzgerald and Holden, 2016). Following a host-switch event, *S. aureus* may undergo a series of genomic changes that increase its fitness in the new niche, leading to successful adaptation to the new host-species (Engering *et al.*, 2013). Genetic mechanisms contributing to host-adaptation include lateral transfer of genes, genetic diversification, gene decay and core gene mutations (Figure 1.3) (Lowder *et al.*, 2009; Viana *et al.*, 2010; Uhlemann *et al.*, 2012; Spoor *et al.*, 2013; Viana *et al.*, 2015).

Lowder *et al.* (2009) demonstrated that the adaptive evolution of human *S. aureus* strains to poultry took place by acquisition of novel mobile genetic elements (MGE) from an avian niche-specific accessory gene pool, and by loss of function of proteins involved in human disease pathogenesis (Lowder *et al.*, 2009). In addition, recent investigations have demonstrated the role of recombination in the adaptation of *S. aureus* isolates within the CC5 clonal complex to poultry (Murray *et al.*, 2017).

Likewise, adaptation to ruminants is largely mediated via MGE (Guinane *et al.*, 2010; Herron-Olson *et al.*, 2007). For example, the CC133 clone, responsible for the majority of small ruminant infections, originated from a human-to-animal host jump by gene loss and diversification of genes encoding proteins responsible for host-pathogen interactions (Guinane *et al.* 2010). Similarly, Resch *et al.* (2013) found evidence that the bovine *S. aureus* CC8 evolved from a human progenitor by loss of a β-haemolysin converting prophage and acquisition of a new SCC element (Resch *et al.*, 2013). More

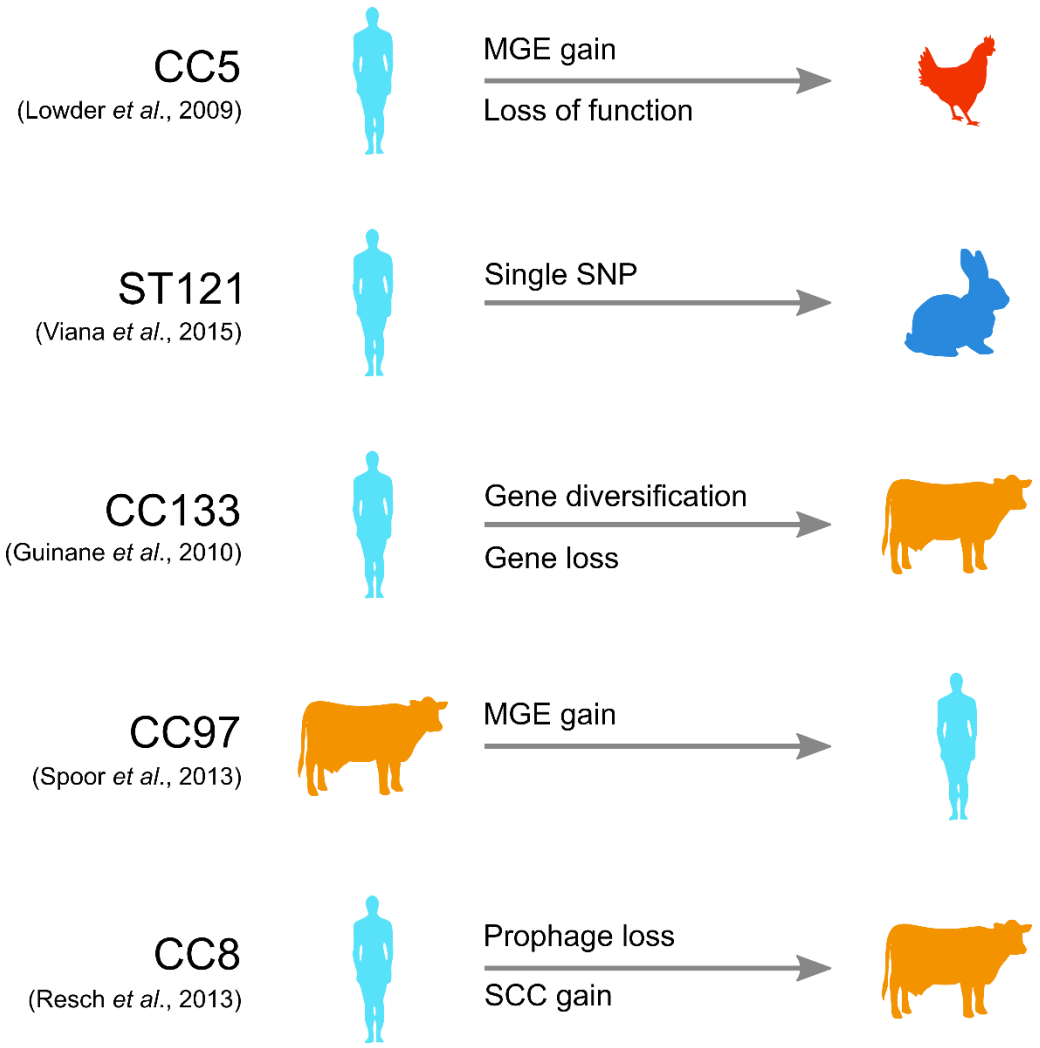


Figure 1.3 Genetic basis of *S. aureus* host-specificity. Adaptation to a new host-species is mediated by various genetic mechanisms, including lateral transfer of genes, genetic diversification, gene decay and core gene mutations.

recently, Spoor *et al.* showed that emergent human MRSA originated from bovine strains by acquisition of MGE encoding antibiotic resistance and human immune evasion genes (Spoor *et al.*, 2013), highlighting the importance of HGT in host-adaptation.

In addition, ruminant-associated strains have also been linked with carriage of antimicrobial resistant genes, highlighting the selective pressures introduced by antimicrobial use in farming (Fessler *et al.*, 2011). This is particularly relevant in the pig industry, in which antibiotics have been extensively used (Barton, 2014). Pig-associated strains of the ST398 have been described carrying novel antimicrobial resistance determinants on multi-resistance plasmids (Fessler *et al.*, 2011; Kadlec and Schwarz, 2009). Similarly, CC97 isolates from pigs have developed resistance to several antibiotics, while there is no evidence for the emergence of such resistance in isolates infecting dairy cows from the same clonal complex (Spoor *et al.*, 2013).

Despite the importance of MGE in host-adaptation, mutations in the core genome may also play a relevant role in adaptation to new niches. For example, Viana *et al.* (2015) described that *S. aureus* adaptation to rabbits was dependent upon a single nonsynonymous nucleotide mutation in the *dltB* gene. Following a human-to-rabbit host-jump, only this single mutation was sufficient to convert a human-specific *S. aureus* strain into one that could infect rabbits (Viana *et al.*, 2015).

Bacterial genes involved in host adaptation are under distinct selective pressures in different animals (Petersen, 2007). Positive diversifying selection, unlike HGT, may

be associated with a slow evolutionary process and its identification at the molecular level is carried out by estimating the ratio between nonsynonymous changes and synonymous substitutions (Goldman and Yang, 1994; Lefebure *et al.*, 2007; Soyer *et al.*, 2009; Xu *et al.*, 2011). In *S. aureus*, positive selection has been investigated in the context of adaptation to small ruminants (Guinane *et al.*, 2010) and to different host-species in general (Lowder *et al.*, 2009). These studies revealed that around 5% of the *S. aureus* genome is under positive selection and no functional categories exhibited higher levels of positive selection than others (Lowder *et al.*, 2009).

Identifying the variety of molecular mechanisms underlying host-adaptation is essential to infer the potential for lineages to cross the species-barrier and infect new host-species (Fitzgerald and Holden, 2016). In addition, bacterial determinants required for host-specificity may represent novel therapeutic targets for controlling human and animal infections (Fitzgerald, 2012).

1.4. Genomic epidemiology

The application of population genomics to epidemiological investigations has given rise to the field of genomic epidemiology. Traditionally, public health microbiology used molecular diagnostic methods for the identification of infectious pathogens, for understanding their distribution and for monitoring their emergence in human populations (Foxman and Riley, 2001; Fawley and Wilcox, 2005; Mothershed and Whitney, 2006; Versalovic and Lupski, 2002). However, these methods were limited in their capacity to discriminate localized outbreaks from cases of endemic disease, or to reconstruct the global or national spread of individual clones (Pérez-Losada *et al.*,

2013). In addition, although MLST has been successfully used to differentiate closely related species (Hanage *et al.*, 2005; Sheppard *et al.*, 2008), it does not provide enough discriminatory power for most epidemic investigations (Odds and Jacobsen, 2008).

With the advent of high-throughput sequencing technologies, WGS of pathogenic bacteria provides the resolution required to discriminate at nucleotide-level, allowing for clear differentiation of strains (Li *et al.*, 2014). In addition, WGS of multiple isolates also facilitates an understanding of outbreak dynamics in great detail, allowing inference of transmission chains (Harris *et al.*, 2013; Köser *et al.*, 2012) and tracking of infection sources (Snitkin *et al.*, 2012). Such analyses may reveal unsuspected transmission events (Price *et al.*, 2014) or reveal reasons for failure of infection control strategies (Howden *et al.*, 2013). Moreover, the integration of epidemiological methods allows us to generate predictions that could not be made otherwise. To define the extent of an outbreak, monitor the effectiveness of control measures and support surveillance (Didelot *et al.*, 2012).

1.4.1. Pathogen diagnosis and outbreaks detection

Rapid identification of bacterial samples is crucial for the correct management of an infectious disease and deciding on the therapies to treat the pathogen (Figure 1.4a) (Didelot *et al.*, 2012; Hasman *et al.*, 2014). Traditional and current clinical diagnosis methods are mostly based on culturing of isolates and subsequent phenotypic and biochemical characterization, a process that may take up to several days (Mercante and Winchell, 2015). With the incorporation of culture-free methods, such as PCR of the 16S rRNA gene (Reller *et al.*, 2007), and more recently matrix-assisted laser

desorption/ionization time-of-flight (MALDI-TOF) (Croxatto *et al.*, 2012), rapid clinical diagnosis can be performed directly on samples. However, these methods do not provide additional information beyond species characterization and require curated reference databases (Singhal *et al.*, 2015).

WGS allows us to characterize pathogens based on their sequence and phylogenetic approaches resolve the relationship of the species at levels not possible with lower resolution methods (Aarestrup *et al.*, 2012; Didelot *et al.*, 2012; Köser *et al.*, 2012). In addition, antimicrobial resistance determinants and virulence factors can also be predicted from genome sequences (Tang and Gardy, 2014), which is particularly important to guiding medical specialists on the antibiotics used for treatment of pathogens in health-care facilities (Reuter *et al.*, 2013). Nevertheless, routine application of WGS for species identification still needs further development to become time-efficient, cost-effective and with minor technical requirements. Additionally, despite the high precision achieved for species differentiation, deep sampling and sequencing of multiple isolates will also reveal the ambiguity of highly recombinogenic species, which may, in turn, challenge the species definition (Hanage *et al.*, 2005).

Determining whether a cluster of infections corresponds to an outbreak is crucial for the application of infection control practices to prevent future cases (Grad and Lipsitch, 2014; Robinson *et al.*, 2013). An outbreak usually refers to a sudden high incidence of the number of cases above normal levels, to the emergence of a previously unrecognized pathogen, or to the entry of a new pathogen into community.

Epidemiological genomics has been successfully applied to determine the existence and nature of several outbreaks associated with hospital settings (Snitkin *et al.*, 2012; Harris *et al.*, 2013; Reuter *et al.*, 2013).

Additionally, to define the extent of the infections, it is essential to determine the population structure among the isolates. Although most outbreaks are caused by the transmission of a clonal lineage of a pathogen, in some cases, the co-existence of several strains can lead to mixed outbreaks (Layton *et al.*, 1997). For example, a series of patients infected with *Neisseria meningitidis* in a hospital were epidemiologically unlinked, indicating diverse infections rather than clonality expected from an outbreak (Reuter *et al.*, 2013). Similarly, WGS was used to delineate *Mycobacterium tuberculosis* outbreaks, allowing identification of linked patients as well as epidemiologically unlinked patients (Walker *et al.*, 2013).

Thus, genomic data integrated in outbreak detections provides much clearer resolution, permitting better decision-making interventions. Additionally, when epidemiological data are missing, the high-resolution achieved by genomic data helps to generate hypothesis regarding the source of the infections.

1.4.2. Inferring the source and transmission chains of pathogens

Identifying the source of an outbreak and determining the pathogen transmission routes is crucial to stop the disease from spreading (Figure 1.4b) (Gardy *et al.*, 2011; Harris *et al.*, 2013; Reingold, 1998).

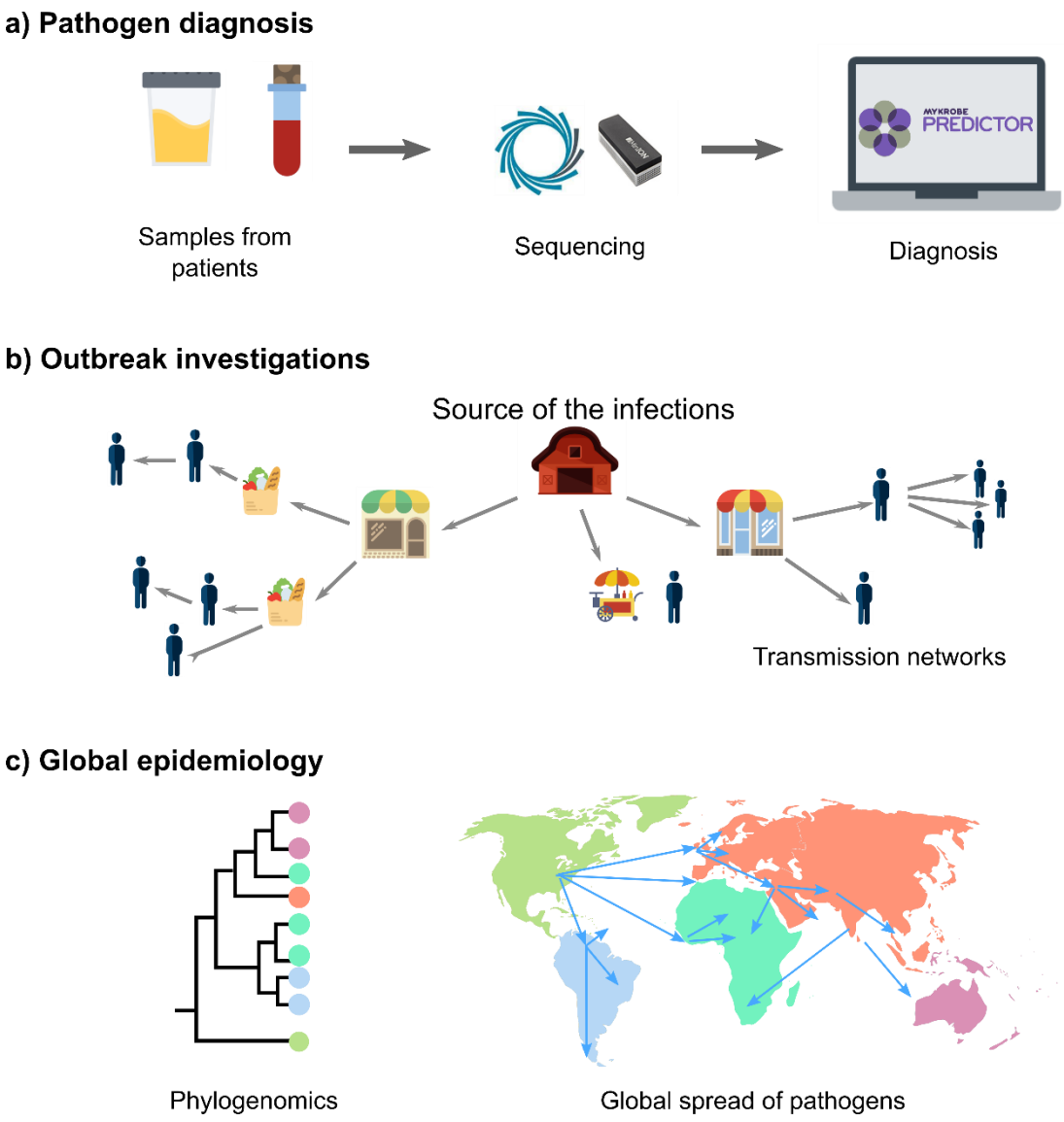


Figure 1.4. Genomic epidemiology of bacterial pathogens. a) Rapid pathogen diagnosis of bacterial samples is crucial for the correct management of an infectious disease and deciding on the therapies to treat the pathogen. b) Outbreak investigations: Identifying the source and the transmission routes of the pathogen is necessary to stop the disease from spreading. c) Global epidemiology permits the study of the worldwide spread of successful bacterial pathogens and their global diversity.

Over the last few years, numerous investigations have used genomic epidemiology for tracing nosocomial and community-associated outbreaks, aiming to identify their sources and inferring the transmission chains of pathogens (Croucher and Didelot, 2015; Reuter *et al.*, 2013). For example, using WGS, health care-associated outbreaks of *S. aureus* (Miller *et al.*, 2014; Nübel *et al.*, 2013), *Klebsiella pneumoniae* (Snitkin *et al.*, 2012), *Enterobacter cloacae* (Stoesser *et al.*, 2014) and *Clostridium difficile* (Jia *et al.*, 2016) have been linked to different reservoirs within the hospitals where they were reported. Genomics has been also applied for resolving community-associated epidemics, such as the *L. pneumophila* outbreaks in Scotland and Canada (McAdam *et al.*, 2014; Knox *et al.*, 2017), but in these cases, although the relationships between the isolates suggested a common source for the infections, the absence of a confirmed environmental source impeded the identification of reservoirs of the pathogens (Knox *et al.*, 2016). Additionally, outbreaks within health-care associated settings may have an external source, such as the *Salmonella enteritidis* outbreak in a UK hospital associated to community (Quick *et al.*, 2015). Finding this link was only possible by integrating the outbreak sequences with whole-genome data from surveillance sequencing (Quick *et al.*, 2015), highlighting the importance of routine WGS. Considering the long-term value of sequence data, researchers from the UK and Ireland recently created a genomic framework for prospective MRSA surveillance in these countries (Reuter *et al.*, 2016).

Identifying when and where an outbreak was originated depends on how representative the sampled data are (Grad and Lipsitch, 2014). If public health authorities were able to sample every individual infected, finding the source of an outbreak may be relatively

easy, but the number of reported cases are usually limited or biased. This can be due to pathogens causing asymptomatic infections or presenting environmental vectors and reservoirs, or to the existence of gaps in the sampling process, which lead to underestimation of the number of infection imports into the population (Jombart *et al.*, 2014; Ypma *et al.*, 2013) and complicate the reconstruction of the transmission chains (Eyre *et al.*, 2013). Nevertheless, using phylogenetic and phylodynamic approaches that can analyse the geographical and temporal information associated with the isolates, one can estimate the most recent common ancestor (MRCA) of the pathogens sampled and infer an approximate date and location of the source of the outbreak investigated (Stoesser *et al.*, 2015). In addition, phylodynamic methods permit reconstruction of the transmission chains and also estimation of the epidemiological parameters useful for understanding disease dynamics (Rasmussen *et al.*, 2014; Volz *et al.*, 2009).

In addition, WGS has been typically applied to a single isolate per host, but such low level of sampling may not provide enough resolution to infer direct transmissions (Eyre *et al.*, 2013). Sequencing of multiple colonies isolated from individuals can reveal the breath of the cloud of bacterial diversity within hosts (Harris *et al.*, 2013; Paterson *et al.*, 2015), which is important for accurately inferring the rates and routes of transmission (Figure 1.5) (Worby *et al.*, 2014). For example, WGS of deeply sampled *C. difficile* infections in a hospital revealed not only direct transmissions between patients but also cryptic transmissions via carriers (Eyre *et al.*, 2013). Similarly, deep sampling of *M. tuberculosis* isolates from several patients permitted differentiation of individuals who had been recently infected, indicating that this

technique could identify super-spreaders and predict the existence of undiagnosed cases (Walker *et al.*, 2013).

Finally, most genomic epidemiological investigations are performed retrospectively (Joensen *et al.*, 2014), but WGS can be used to track ongoing outbreaks and reveal insights regarding the source and transmission dynamics of pathogens in real-time (Quick *et al.*, 2015). With new rapid sequencing technologies such as the Oxford Nanopore MinION, same-day diagnostic, surveillance and outbreak detection are now feasible (Joensen *et al.*, 2014; Votintseva *et al.*, 2017).

1.4.3. Global epidemiology

Population genomics can be applied to trace the worldwide spread of successful bacterial pathogens and study their global diversity (Figure 1.4c) (Croucher and Didelot, 2015). Using similar strategies to global surveillance programs for monitoring the antimicrobial resistance around the world (O'Brien, 1995), sharing WGS data on a global scale can facilitate infectious disease detection and surveillance (Deng *et al.*, 2016), ultimately improving global public health (Hendriksen *et al.*, 2011).

It is important to understand how bacterial strains causing infections in a specific geographic location are related to those recovered worldwide (Spratt and Maiden, 1999). For example, comparing the *Vibrio cholerae* genomes responsible for the 2010 Haitian cholera outbreak to global collections of *V. cholerae* isolated over several years, it was clear that the outbreak genomes were a monophyletic group related to a strain from South Asia (Hendriksen *et al.*, 2011). These data led to the conclusion that the outbreak was caused by isolates imported by Nepalese peacekeeping troops (Rene

S. Hendriksen *et al*., 2011; Katz *et al*., 2013). Similarly, WGS has been decisive for confirming the source of multi-country food-borne outbreaks, such as the *Salmonella* Newport gastroenteritis in Europe associated with watermelons from Brazil (Byrne *et al.*, 2014) or the *Salmonella* Enteritidis European outbreak associated with egg distribution networks (Dallman *et al.*, 2016). Conversely, WGS on global epidemics was decisive to rule out transmissions of antibiotic resistant *S. enterica* Typhimurium between animals and humans (Mather *et al.*, 2012), indicating that the source of infections was probably imported food, environmental reservoirs, or result of transmission between persons (Mather *et al.*, 2013).

In addition, WGS studies at a global scale can provide information on long-term microbial evolution, including gene content variation and impact of selection pressures on the genomes (Croucher *et al.*, 2013). Finally, integrating genomes isolated from different hosts with other temporal and geographic information is essential to track the pandemic spread of particular strains, yielding a long-term overview of epidemiological patterns.

1.5. Population genomics of within-host evolution

Advances in WGS technologies have facilitated studies on the evolution of bacterial populations derived from the same individual over the course of a single infection, revealing previously unexpected levels of within-host genomic diversity (Bryant *et al.*, 2013; Lieberman *et al.*, 2011; Smith *et al.*, 2006). The spatial heterogeneity provided by the complex niches within a host may drive the diversification of the pathogen population, highlighting the importance of sequencing multiple isolates, as the genome

of a single isolate often does not represent the entire pathogen population in the infected individual (Figure 1.5) (Markussen *et al.*, 2014).

1.5.1. Adaptation to the host-environment

Within-host diversity is largely shaped by genetic drift, which depends on the size of the bottlenecks during transmissions, the creation of bacterial subpopulations in different body locations and the fluctuations in the pathogens population size (Didelot *et al.*, 2016; Golubchik *et al.*, 2013). In addition, intra-host diversity is affected by purifying and diversifying selection (Didelot *et al.*, 2016; Worby *et al.*, 2014). During long-term colonisation and persistence within the host, bacterial populations face numerous challenges, as they need to evade the host immune system, compete with the native microbiota and establish a successful niche for acquisition of nutrients (Didelot *et al.*, 2016).

Within-host adaptive evolution is in part mediated by mutations in global regulatory networks (Damkiaer *et al.*, 2013; Howden *et al.*, 2011; Lieberman *et al.*, 2011; Marvig *et al.*, 2015). For example, monitoring of long-term evolution of *P. aeruginosa* in the airways of several patients with cystic fibrosis patients revealed mutations in the global regulatory genes *mucA*, *rpoN* and *lasR* (Yang *et al.*, 2011). These mutations resulted in the remodelling of regulatory networks facilitating the generation of new phenotypes (Damkiaer *et al.*, 2013). Similarly, in a different longitudinal study of *P. aeruginosa* in various patients with cystic fibrosis over a 38 year period, several mutations were also found in genes related to the remodelling of regulatory functions (Marvig *et al.*, 2015). By changing the expression of several genes, dramatic

phenotypic changes can be produced, allowing pathogens to quickly optimize their fitness to the fluctuating conditions of the within-host environment (Damkiaer *et al.*, 2013).

Global transcriptional remodelling has been linked to increased bacterial virulence (pathoadaptive mutations) (Lieberman *et al.*, 2011; Young *et al.*, 2012). In an individual carrying *S. aureus* in the nose who developed a severe bloodstream infection, WGS of isolates revealed that the progression from carriage to disease was associated with mutations that caused truncation of a transcriptional regulator implicated in pathogenicity (Young *et al.*, 2012).

Attenuation of virulence represents a common evolutionary strategy for within-host adaptation (Price *et al.*, 2013; Smith *et al.*, 2006; Zdziarski *et al.*, 2010). WGS comparison of *Burkholderia pseudomallei* isolates obtained twelve years apart from the same patient showed the pathogen underwent several adaptive changes, including the inactivation of virulence factors (Price *et al.*, 2013). In the case of a chronic *Salmonella* Enteritidis infection in an immunocompromised patient, evolution resulted in genome degradation targeting genes encoding virulence determinants (Klemm *et al.*, 2016). A potential reason for virulence attenuation in strains co-evolving within their hosts is that reduced severity may permit the transmission of pathogens, leading to their long-term survival (Didelot *et al.*, 2016).

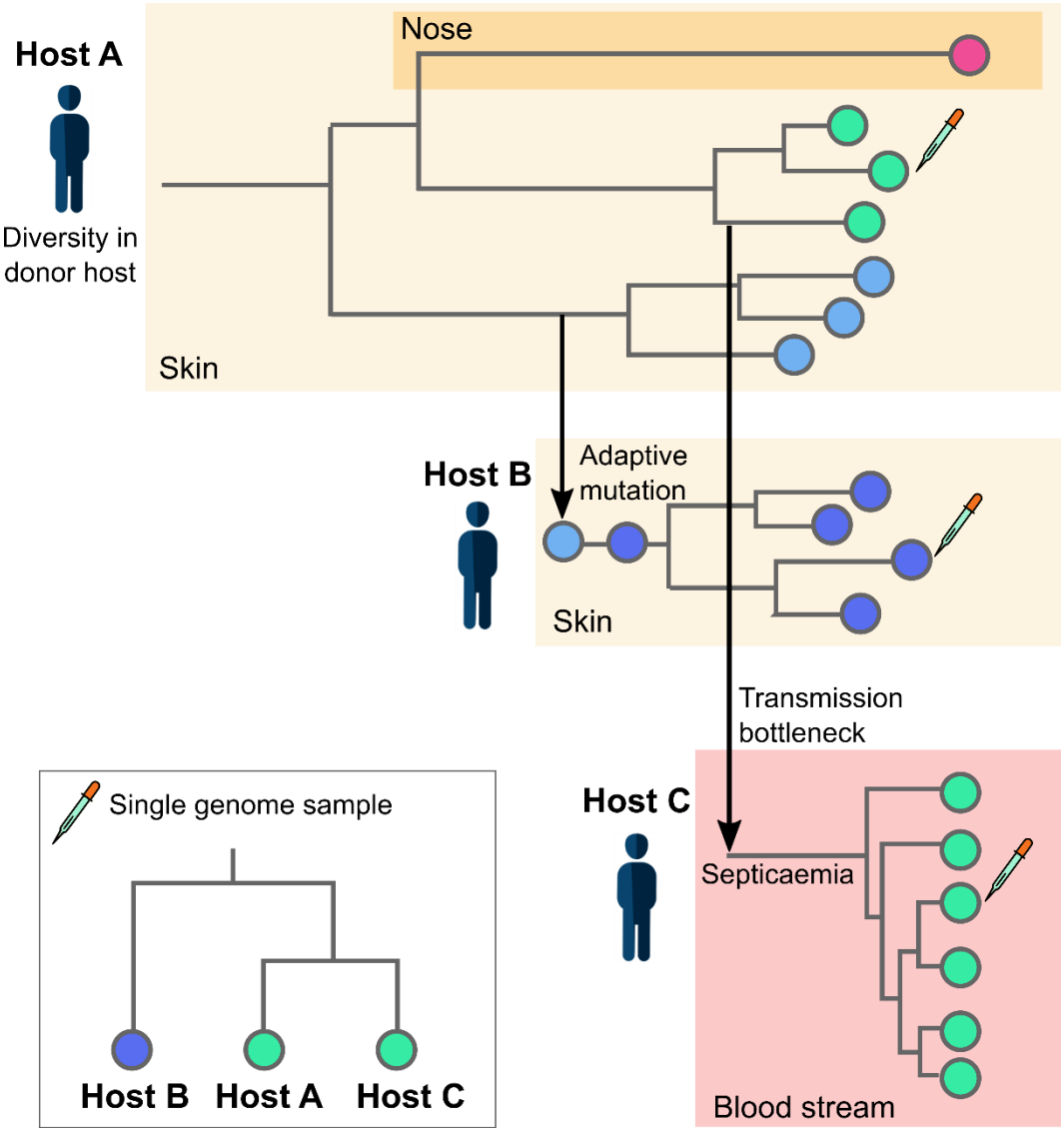


Figure 1.5. Dynamics of within-host adaptive evolution. Host A, colonised by a diverse population of a pathogen, transmits two different bacterial sub-lineages to hosts B and C. In host B, an adaptive mutation quickly becomes fixed in the population. In host C, the transmission leads to a bloodstream infection. Sequencing only a single genome per host would have not revealed the cloud of diversity in each host nor the actual transmission chain of the pathogen. *Figure adapted from Didelot et al., 2016.*

Distinct environments within the host may drive the diversification of the pathogen population. For example, *P. aeruginosa* infecting a patient with cystic fibrosis over 32 years, diverged into distinct coexisting sublineages spatially separated within the lung (Markussen *et al.*, 2014). Bacterial diversification also occurs when the same strain colonizes different persons, suggesting individual host environments drive distinct adaptive evolution (Zdziarski *et al.*, 2010). Nevertheless, convergent evolution has been reported in same studies and can be used to identify pathoadaptive mutations (Lieberman *et al.*, 2011; Marvig *et al.*, 2014). Several of the mutations found by Marvig and colleagues in their longitudinal study of *P. aeruginosa* were in the same genes associated with the cell envelope and regulatory functions (Marvig *et al.*, 2015). Similarly, *Burkholderia dolosa* isolates from several subjects with cystic fibrosis revealed parallel evolution affecting genes encoding antibiotic resistance and bacterial membrane proteins (Lieberman *et al.*, 2011).

In addition, within-host adaptation may be mediated by hypermutators (Feliziani *et al.*, 2014; Markussen *et al.*, 2014; Marvig *et al.*, 2013). Genomic analysis of *Salmonella* Enteritidis isolates from a 16 year chronic infection in an immunocompromised patient revealed a mutation in the mismatch repair gene *mutS*, which accelerated the mutation rate of the pathogen (Klemm *et al.*, 2016). Similarly, Lieberman *et al.* (2014) and Marvig *et al.* (2015) found samples with hypermutator genotypes, suggesting that these strains drive diversification that enables them to explore alternative evolutionary pathways to the host environment (Lieberman *et al.*, 2011; Marvig *et al.*, 2015).

The antimicrobial therapies used to treat infections mean that in many cases genetic evolution has been related to the development of antibiotic resistance (Didelot *et al.*, 2013; Eldholm *et al.*, 2014; Mathers *et al.*, 2015). Within a single patient, and over a period of 3.5 years, multidrug-resistant *M. tuberculosis* isolates evolved from a susceptible ancestor by individual mutations and a gradual increase in fitness in the presence of antibiotics (Eldholm *et al.*, 2014). Some SNPs conferring resistance appeared shortly after a new drug therapy was initiated, reflecting the rapid expansion and displacement of some clones over others during infection (Eldholm *et al.*, 2014). This is particularly relevant when pathogens acquire resistance to the last available antibiotic to treat their infections, as the *S. aureus* strains acquiring resistance to vancomycin (Howden *et al.*, 2011). Howden and colleagues compared susceptible and resistance strains before and after antibiotic treatment failure, revealing a common pattern of mutations among the resistant isolates (Howden *et al.*, 2011). Single nucleotide substitutions within the *walKR* two-component regulatory locus were acquired in less than 6 w, and not only caused resistance to the antibiotic but also attenuated virulence (Howden *et al.*, 2011).

The application of WGS on multiple isolates from individual hosts over different timescales demonstrates the potential of population genomics for studying the microbial evolution within-hosts. This has enabled a better understanding of the effects of various selective forces in the adaptation of pathogens to colonisation and persistence, including those imposed by the host environment and others with external origin, such as antibiotics.

1.5.2. Within-host experimental evolution

In experimental evolution studies, populations of organisms are grown in defined environments in order to study the evolutionary processes occurring over hundreds or thousands of generations (Elena and Lenski, 2003; Kawecki *et al.*, 2012). Experiments to reproduce evolution in laboratory conditions were initiated nearly three decades ago (Hindré *et al.*, 2012; Kawecki *et al.*, 2012) and have revealed many insights into the dynamics and genetic basis of bacterial adaptation (Elena and Lenski, 2003). With the advent of high-throughput sequencing, experimental evolution coupled with WGS facilitated the identification of the genetic changes of adapted populations to a novel environment (Barrick *et al.*, 2009; Charusanti *et al.*, 2010; Toprak *et al.*, 2012), and the "evolve and resequence" approach has also enabled monitoring of molecular evolution in real-time (Long *et al.*, 2015). To date, the great majority of experimental evolution studies have been performed in laboratory conditions (Hoang *et al.*, 2016; Azevedo *et al.*, 2016; Ensminger, 2013).

Within-host experimental evolution represents an excellent framework for better understanding the evolutionary dynamics underlying the adaptation of pathogens to their hosts, allowing testing of hypothesis and predictions, while monitoring for experimental factors that are difficult to control using natural populations (Giraud *et al.*, 2017). This approach has been used to investigate the adaptation of *E. coli* to the gut in mice (Barroso-Batista *et al.*, 2014; Lescat *et al.*, 2017), revealing strong patterns of convergence in the mutations and genes affected (Lescat *et al.*, 2017), and high levels of clonal interference, with different adaptive mutations not reaching fixation (Barroso-Batista *et al.*, 2014). Using similar approaches, the adaptation of the plant

pathogens *Pseudomonas syringae* and *Ralstonia solanacearum* to their native hosts and to alternative plants has been investigated (Guidot *et al.*, 2014; Meaden and Koskella, 2017). With the increased ability to link phenotypic adaptations to specific genomic signatures, experimental evolution to study the adaptation of bacterial pathogens to their hosts and to investigate the development of pathogenesis offers enormous potential to understand bacterial infections and disease progression (Ensminger, 2013).

1.6. Novel bioinformatics methods in population genomics

The continuous development of next-generation sequencing technologies is associated with reducing costs of WGS, which is leading to increasing amounts of bacterial genomes being sequenced. Subsequently, new bioinformatics tools able to handle very large amounts of genomic data are required for population genomic studies.

1.6.1. Large-scale population genomic analysis

Analysing vast numbers of genomes requires algorithms that can infer evolutionary, epidemiological, phylogeographic and phylodynamic patterns in an efficient and rapid manner. In addition, methods that integrate information from multiple sources could provide more precise insights in infectious diseases (Baele *et al.*, 2017). Novel phylodynamic approaches include the reconstruction of transmission chains and the origin of outbreaks (De Maio *et al.*, 2015) or understanding the association of phenotypic traits correlated through evolution (Cybis *et al.*, 2015). In addition, new tools for detecting recombination can handle hundreds of genomes in a matter of hours (Didelot and Wilson, 2015; Mostowy *et al.*, 2017). Additionally, pangenomic pipelines

can analyse datasets with thousands of samples using standard computers (Page *et al.*, 2015). Other pipelines have been developed to integrate bioinformatics tools with epidemiological information to generate complete public health microbiology reports, facilitating laboratories outbreak investigations. Furthermore, other recently developed methods include visualization platforms that deliver dynamic data exploration (Argimón *et al.*, 2016) or interactive applications allowing fast exploration of datasets and outputs from multiple large-scale population genomic analyses (Hadfield *et al.*, 2017). In summary, tools aiding the large-scale analysis, visualization and interpretation of complex data, and integration of multiple results are revolutionizing population genomics.

1.6.2. Genome-wide association studies

The ability to link bacterial genetic variants with specific phenotypes can provide new insights into the underlying basis of microbial traits, which could explain many biological mechanisms of infectious diseases (Collins and Didelot, 2017). Genome-wide association studies (GWAS) have been extensively applied in human genomics, linking hundreds of genetic variants with complex traits and diseases (Stranger *et al.*, 2011; Visscher *et al.*, 2012). Although the application of GWAS on microbial genomes was proposed more than 10 years ago (Falush and Bowden, 2006), the first studies on bacteria only appeared recently (Sheppard *et al.*, 2013; Farhat *et al.*, 2013). In the last four years, several bacterial GWAS investigations have been published, finding associations between genetic features (SNPs, k-mers and genes) with clinically relevant phenotypes, including antibiotic resistance (Farhat *et al.*, 2013; Alam *et al.*, 2014; Chewapreecha *et al.*, 2014; Wozniak *et al.*, 2014), virulence (Laabei *et al.*,

2014), pathogenicity (Howell *et al.*, 2014), and host-associations (Sheppard *et al.*, 2013; Weinert *et al.*, 2015). While some of these studies have successfully used statistical approaches developed for human genomics (Alam *et al.*, 2014; Chen and Shapiro, 2015), new tools specifically designed for microbial GWAS that account for specific challenges of bacterial populations, including high levels of clonality or recombination, have been released recently (Lees *et al.*, 2016; Brynildsrud *et al.*, 2016; Collins and Didelot, 2017). GWAS in bacterial pathogens is a promising field expected to grow in the incoming years (Power *et al.*, 2017) and has the potential to improve enormously the way we understand infectious diseases (Collins and Didelot, 2017).

1.6.3. Simulations of evolution of bacterial genomes

Simulation of genetic sequences have been widely used to infer population parameters (Keightley, 1998), for hypothesis testing (Hoban *et al.*, 2012) and for studying the evolution of diseases in humans (Sellers *et al.*, 1998; Peng *et al.*, 2007). In bacterial studies, simulations have also been used for method testing (Falush *et al.*, 2006; Hedge and Wilson, 2014). A range of tools to simulate the evolution of sequences exist, and these can be classified into coalescent-based methods, which simulate populations backward in time (Delport, 2006; Excoffier and Foll, 2011; Ewing and Hermisson, 2010); and forward-time methods, which simulate entire populations from past to present (Balloux, 2001; Carvajal-Rodriguez, 2008). Although the former methods are more computationally efficient, they do not keep track of ancestral information and present difficulties for incorporating complex evolutionary models, which limits their use in population genomic studies. Nevertheless, some methods can simulate complex multispecies coalescent histories, including events such as recombination, population

subdivision and migration, but no selection (Arenas and Posada, 2014). In addition, there has been a lack of general and efficient tools specifically designed for simulating genome-wide bacterial evolution (Brown *et al.*, 2016), and although a few methods have been recently published (Worby and Read, 2015; Brown *et al.*, 2016; De Maio and Wilson, 2017), these are based on coalescent algorithms and do not integrate selection options. Thus, the development of new simulators that account for multiple evolutionary and population events, specifically designed for bacteria, in combination with population genomic methods, could improve our understanding of pathogen evolution and global epidemics.

1.7. Summary

With the development of high-throughput sequencing technologies, WGS of multiple isolates is facilitating bacterial population genomic investigations at very-large scales, which is revolutionizing our understanding of infectious diseases. Recent studies have revealed novel insights into the population structure of bacterial pathogens, their evolutionary dynamics, causes of the emergence of new clones and the genetic determinants of niche adaptation. In addition, the high-resolution delivered by WGS is impacting outbreak investigations, allowing the accurate characterization of bacterial pathogens and inference of the source and patterns of spread with high precision. Finally, longitudinal sampling of individual hosts allows studying the within-host evolution of bacteria during the colonization and infection of hosts. These approaches along with new bioinformatics methods will result in new strategies to design control measures for prevention, diagnosis and treatment of the infections caused by pathogens.

1.8. Aims of this study

The general purpose of this work is to use population genomics and genomic epidemiology approaches to investigate the molecular mechanisms and the evolutionary dynamics of bacterial adaptation to different ecological niches, specifically humans and animals. The individual aims are:

- To investigate the evolutionary history of the *S. aureus* species in the context of its host-associations and identify signatures of host-adaptation.

- To investigate the host-adaptive evolution of *S. aureus* during experimental infections in the face of regular transmission events.

- To examine the aetiology of a cluster of *Legionella longbeachae* infections in Scotland in the context of the global species diversity.

2

Evolutionary history of a multi-host pathogen and signatures of host-adaptation.

2.1. Introduction

Staphylococcus aureus is able to colonize and cause disease in a wide range of host-species. Contrary to single host-type pathogens, the capacity of generalists to switch host-species can provide opportunities for expansion into new host populations. The domestication of animals in the Neolithic period and the more recent intensification of livestock farming provided increased opportunities for the spread of bacterial pathogens between humans and animals (Morand *et al.*, 2014). Of note, the majority of emerging human infectious diseases have been traced to an animal origin (Woolhouse and Gowtage-Sequeria, 2005).

S. aureus has a clonal population structure defined by lineages that have single or multiple host-tropisms (Feil *et al.*, 2003; Shepheard *et al.*, 2013; Weinert *et al.*, 2012). Inter-host-species transmission can be of critical public health importance, as exemplified by the emergence of *S. aureus* livestock-associated strains. For example, the livestock-associated methicillin-resistant CC398, which is associated with pigs and other livestock, can cause zoonotic infections of pig-farmers and their close contacts (Price *et al.*, 2012; Van Cleef *et al.*, 2011). These infections from one host-species to another were transient, spillover events that disappeared without leading to additional transmissions within the new host-species (McCarthy *et al.*, 2012). However, inter-species transfers leading to onward transmissions and establishment within the new host-population are increasingly been recognised (Resch *et al.*, 2013; Spoor *et al.*, 2013). Such jumps are of particular interest due to the capacity of *S. aureus* to acquire resistance to antimicrobials and antiseptics, resulting in infections that are refractory to treatment and that persist in hospitals and care centres. With the emergence of new

clones, several studies have investigated the evolution of animal strains and their potential zoonotic capacity (Harrison *et al.*, 2013; Monecke *et al.*, 2016; Weinert *et al.*, 2015).

Understanding *S. aureus* host switching dynamics, including the directionality and frequency of host-jumps during its evolutionary history, would allow for the design of more efficient protective measures. Previous work employed multi-locus sequence typing (MLST) to relate the population structure of *S. aureus* with its host-association, providing evidence for the occurrence of ancient and recent host-jump events from humans leading to the emergence of *S. aureus* clones in livestock populations (Shepheard *et al.*, 2013; Weinert *et al.*, 2012). More recently, whole genome sequencing has been employed to investigate the evolution of clones of human and animal origin, providing new insights into the emergence, transmission and acquisition of antibiotic resistance in hospital, community, and agricultural settings (Holden *et al.*, 2013; McAdam *et al.*, 2012; Price *et al.*, 2012; Spoor *et al.*, 2013).

Studying the genetic basis of *S. aureus* host-specificity is important to understand the molecular basis of pathogenesis and the potential for *S. aureus* to cross the species barrier to infect new host-species. Comparative genomic analyses have revealed a role for specific mobile genetic elements and core gene mutations in the host-adaptation of *S. aureus* (Lowder *et al.*, 2009; Viana *et al.*, 2010; Viana *et al.*, 2015). In addition, other studies identified an array of genetic signatures linked to host adaptation, including allelic diversification and gene decay (Spoor *et al.*, 2013). Bacterial

determinants required for host-specificity could represent novel targets for controlling human and animal infections.

However, previous investigations have focused on specific clonal complexes, particular host-species or involved limited numbers of genomes. A large-scale, genome-based analysis of the *S. aureus* evolutionary landscape in the context of its host ecology is lacking, and the scale and molecular processes underpinning *S. aureus* adaptation to different host-species remains poorly understood.

2.2. Aims

- To reconstruct a high-resolution phylogeny of the clonal diversity of *S. aureus* associated with a wide range of host-species.
- To investigate the evolutionary dynamics of *S. aureus* in the context of its host-associations, including the number of jumps between humans and animals and its ancestral host state.
- To identify genomic signatures responsible for host-association towards understanding novel mechanisms of host-adaptation.

2.3. Material and Methods

2.3.1. Strains selection[1]

For selection of isolates, the literature was reviewed (date: November 2013) and all available *S. aureus* strains associated with animals and humans for which genomes had been determined were identified. In order to represent the breadth of clonal

[1] Strains selection was performed in collaboration with Emily Richardson, Ewan Harrison and Lucy Weinert.

complexes, host-species diversity, multiple geographical locations and a wide temporal scale (1930 to 2014), we sequenced 172 isolates for the current study, and included publicly available sequences as follows; 74 reference genomes, 302 from the EARRS project (Aanensen *et al.*, 2016) and 252 from other published studies of collaborators (Supplementary Table 1). Given a predominant European origin of the animal isolates, due to the contemporary interest in animal *S. aureus* in Europe, we chose to enrich the number of human isolates with the EARSS collection (Aanensen *et al.*, 2016). Overall, we included 800 isolates representative of 43 different host-species and 77 clonal complexes (CCs), isolated in 50 different countries across 5 continents (Supplementary Table 1).

2.3.2. Mapping reads, variant calling and phylogenetic reconstruction

Bacterial DNA was extracted and sequenced using Illumina HiSeq2000 with 100-cycle paired-end runs at the Wellcome Trust Sanger Institute or Illumina HiSeq2000 at Edinburgh Genomics. The nucleotide sequence data were submitted to the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) with the accession numbers listed in the Supplementary Table 1. Completed genomes downloaded from the NCBI database were converted into pseudo-fastq files using Wgsim (https://github.com/lh3/wgsim). We excluded isolates that might have been contaminated or had poor quality sequence data, including those with large number of contigs, a large number of 'N's in the assemblies or an unusually large genome size (>2.9 Mb). Sequence types were determined from the assemblies using mlst (http://www.mlst.net/) and MLST_check (https://github.com/sanger-pathogens/mlst_check), with the MLST database for *S.*

aureus (http://pubmlst.org/saureus/). Sequence reads were mapped[2] to a relevant reference genome (European Nucleotide Archive (ENA) ST425 (strain LGA251, accession number FR821779), using SMALT (http://www.sanger.ac.uk/science/tools/smalt-0) following the default settings to identify single nucleotide polymorphisms (SNPs). We used GATK (McKenna *et al.*, 2010), samtools (Li *et al.*, 2009) and VCFtools (Danecek *et al.*, 2011) to produce a matrix of 137,556 high-quality SNPs. Consensus sequences of every genome were obtained from the vcf files using vcf_to_consensus_sequence.py and were further concatenated into a multiple genome sequence alignment. A maximum likelihood tree for the whole dataset and another tree for only the *S. aureus* species isolates were constructed using RAxML with default settings and 1000 bootstrap replicates (Stamatakis, 2006).

2.3.3. *De novo* assembly and genome annotation

Sequencing reads were used to create multiple assemblies using VelvetOptimiser v2.2.5 (Zerbino, 2010) and Velvet v1.2 (Zerbino and Birney, 2008) or SPAdes (Bankevich *et al.*, 2012). The assemblies were improved by scaffolding the best N50 and contigs using SSPACE (Boetzer *et al.*, 2011) and sequence gaps filled using GapFiller (Nadalin *et al.*, e 2011). Prokka (Seemann, 2014) was used to annotate the genomes.

2.3.4. Estimating ancestral host-state and host jumps

To estimate the number of host jumps occurred during the evolution of *S. aureus* we first inferred past host association states on the phylogenetic tree using Adaptml v1.0

[2] Mapping to the reference genome was performed by Ewan Harrison.

(Hunt *et al.*, 2008). This software uses an evolutionary Markov model to classify isolates into ecologically similar habitats based on their genetic and ecological similarity. We used the 783 isolates ML phylogenetic tree and isolates were assigned to one of the following host categories: birds, cows, goats, sheep, carnivores, horses, humans, rabbits, monkeys, sheep, rodents, pigs, bats or unknown. AdaptML merges ecotypes into habitats that represent ecologically meaningful groups of genotypes and assigns habitats to the ancestral nodes of the phylogenetic tree, allowing us to infer the number of habitat transitions/host jumps that took place during the evolution of *S. aureus*. The robustness of the analysis was assessed by running AdaptML with different sets of parameters: initial number of habitats from 10 to 20, converge threshold from 0.001 to 1 and collapse threshold from 0.05 to 0.5. AdaptML reached high convergence and consistently inferred the same number of habitats from the provided phylogeny.

2.3.5. Genome-wide association analysis

A range of methods for performing GWAS in bacteria are available (reviewed in Power, Parkhill, and de Oliveira, 2016). Although these programs are based on different analytical approaches, all of them work by finding statistically significant associations between a given genomic feature and a phenotype of interest. We used ROADTRIPS (Thornton and McPeek, 2010), a tool that was shown to perform well in a previous *S. aureus* GWAS study (Alam *et al.*, 2014). ROADTRIPS tests for case-control association and accounts for samples with related individuals and partially or completely unknown population structure. As cofounding effects can lead to spurious associations and reduce the power of GWAS, population structure was determined

with BAPS (Bayesian Analysis of Population Structure) (Corander *et al.*, 2008) and the phylogenetic relationships. BAPS clusters at the levels 1 and 2 containing isolates from different clades of the phylogenetic tree were split into different subpopulations. The core genotype matrix of biallelic SNPs was parsed to ROADTRIPS and four independent runs were performed, one for each host state tested (human, bovids, birds and pigs), where host-species of interest was defined as cases and the rest of hosts were defined as controls. To avoid false positives due to multiple testing, a Bonferroni-correction was applied and SNPs were considered to be statistically significant if their p-value was below the significant threshold $\alpha = 0.05/81{,}680$ SNPs = 6.12×10e-7. Given the plasticity of bacterial genomes, SNPs-based GWAS can only identify a fraction of the genetic determinants responsible for phenotypic variation. To overcome this limitation we also performed a GWAS testing using SEER (beta version), a method that identifies sequence elements (k-mers) significantly enriched in a phenotype of interest (Lees *et al.*, 2016; Weinert *et al.*, 2015). In this analysis[3], short sequences (k-mers) for the three major groups of hosts (humans, bovids and birds) were obtained from their corresponding genomes. Next, k-mers enriched in different hosts were identified by running two pairs of comparisons: humans versus ruminants and humans versus birds. Similar to the Roadtrips analysis, population structure was taken into account from the BAPS clusters using a basic association test with a strict p-value threshold of $10e^{-8}$. Enriched k-mers were then mapped against the *S. aureus* pangenome using BLASTn in order to identify gene sets associated with each host.

[3] The k-mer enrichment step of the SEER analysis was run by Jukka Corander.

2.3.6. Positive selection analysis

To identify genes under positive selection in different host groups, we first identified lineages (STs or CCs) correlated with particular hosts. As the power of the selection analysis is determined by the number of isolates included, only clades with more than 10 isolates associated with a host were considered. Based on these criteria, 15 CCs from four groups of hosts were analysed: 9 for humans, 3 for ruminants (CC133, primarily associated with sheep and goats and the cows related CC151 and CC97), 2 for birds and one for pigs (Supplementary Table 2). Although the CC398 clade also contained several human isolates, these mostly represent spill-over events rather than an established association so the CC398-human group was not included for the analysis. Given the variable number of isolates of each CC-host group, in order to standardize the analysis while preventing the underestimation of genes under positive selection, 10 isolates linked with a host were analysed at a time. Replicates or triplicates of different subsets of genomes using sampling with replacement was carried out if the number of isolates for that lineage was large enough. Next, we identified orthologous genes in each of these groups using the algorithm OrthoMCL (Li *et al.*, 2003) integrated in get_homologues (identity >70%, similarity >75%, f50, e-value = 1e-5) (Contreras-Moreira and Vinuesa, 2013). Genes were considered orthologous if they were present in at least 70% of the genomes. Since alignment of coding DNA sequences may insert gaps in codons and produce frame-shifts, we aligned genes at the protein level using MUSCLE 3.8.31 (Edgar, 2004) and translated these sequences back to DNA using pal2nal v14 (Suyama *et al.*, 2006). Genes identified as inparalogous that turned out to be duplications were kept for further analyses, otherwise discarded. For every alignment, recombination was detected using

the NSS, Max Chi and Phi tests included in PhiPack (Bruen *et al.*, 2006) and recombinant genes removed from further analyses. For the gene clusters containing 10 isolates, phylogenetic trees were extracted from the 783 isolates ML tree. For clusters with less than 10 genomes, subtrees were produced from the general tree using the tree prune function in ete268. The DNA alignments and trees were used for PAML analysis (Yang, 2007). We employed the site evolution models of Codeml (M1a, M2a, M7, M8 and M8a) to perform codon-by-codon analysis of dN/dS ratios (nonsynonymous to synonymous substitution, ω) of genes and a likelihood ratio test (LRT) was used to determine significant differences between nested models M1a-M2a, M7-M8, M8a-M8, where one accounts for positive selection (alternative hypothesis) and the other specifies a neutral model (null hypothesis). Statistic tests were assessed to a chi-square distribution with 2 and 1 degrees of freedom (Yang, 2007). Bayes Empirical Bayes (Yang, Wong, and Nielsen, 2005) was used to calculate the posterior probabilities of amino acid sites under positive selection of proteins that had significant LRTs. As independent replicates from similar CC/Host groups resulted in slightly different genes positively selected, we used get_homologues to merge the core genomes and genes selected for each group using same parameters as above. Genes under positive selection were considered when they were in common for different replicates with a p-value of 0.05 or were identified in different replicates with a stringent p-value (0.05/number of genes per core genome).

2.3.7. Annotation of functional categories and enrichment analysis

To explore functional categories under positive selection we performed classification of Clusters of Orthologous Groups (COGs), annotated Gene Ontology terms (GO) and

analysed metabolic pathways (KEGG). To assign COG terms, we performed BLASTp of single representatives of the orthologous clusters against the prot2003-2014 database, retrieving the top 5 hits to include alternative annotations. We mapped the gene IDS obtained to the cog2003-2014.csv database from which the COGs were inferred. Frequencies of COGs for positively selected genes in each CC-host were compared with the average COG frequencies in the respective core genomes. GO annotations were obtained by mapping the genes to the go_20151121-seqdb, uniprot_sprot and uniprot_trembl databases using BLASTp. From these, the UniprotKB were mapped to the gene_association_goa database and filtered by bacteria domain to obtain the GO categories. To visualize and identify overrepresented GO categories of positively selected genes in different hosts, we used BiNGO (Maere *et al.*, 2005), a plugin available in Cytoscape (Shannon *et al.*, 2003). We identified overrepresented categories using the hypergeometric test with the Benjamini and Hochberg False Discovery Rate (FDR) multiple testing correction at a significance level of 5%. We chose the 'Biological Process' category and the prokaryotic ontology file (gosubset_prok.obo). However, as most groups did not show significant overrepresentation, we visualized all the GO categories of genes under positive selection and used REVIGO (Supek *et al.*, 2011) with the p-values from BiNGO in order to obtain summaries of non-redundant GO terms classified into functional categories. We performed metabolic pathway reconstruction and enrichment by mapping the protein sequences of genes to the KEGG database (ko.ep.fasta integrated in Kobas v2.0) with BLASTp. We used Kobas 2.0 (Xie *et al.*, 2011) to identify enriched KO annotations, which were uploaded to the interactive pathways explorer iPath2 server (Yamada, L*et al.*, 2011) to visualize the pathway maps.

2.4. Results

2.4.1. Diverse patterns of host-association explain the evolution of clonal lineages of *S. aureus*

To construct an accurate phylogeny of *S. aureus*, we sampled isolates from the breadth of known clonal complexes, host-species diversity, multiple geographical locations and a wide temporal scale. Overall, we selected 800 isolates representative of 43 different host-species and 77 clonal complexes (CCs), isolated in 50 different countries across 5 continents (Supplementary Table 1). We mapped the genomic reads of all the isolates against the reference genome *S. aureus* LGA251 (accession number NC_017349) and built a core genome alignment. A total of 115,149 SNPs were detected, which were used to reconstruct a maximum-likelihood phylogeny of the *S. aureus* species complex (Figure 2.1a).

The ML phylogenetic tree shows the existence of highly divergent clades belonging to two recently described novel species, *Staphylococcus argenteus* and *Staphylococcus schweizeri*, which are part of the extended *S. aureus*-related complex (Steven Y C Tong *et al*., 2015). *S. argenteus*, an emergent cause of human clinical infections, is more closely related to bats isolates than to other human *S. aureus* STs, consistent with a possible non-human evolutionary origin for *S. argenteus*. *S. schweitzeri* was isolated from monkeys and bats and has only been recovered once from human hosts (Schaumburgh *et al.*, 2012; Tong *et al.*, 2015).

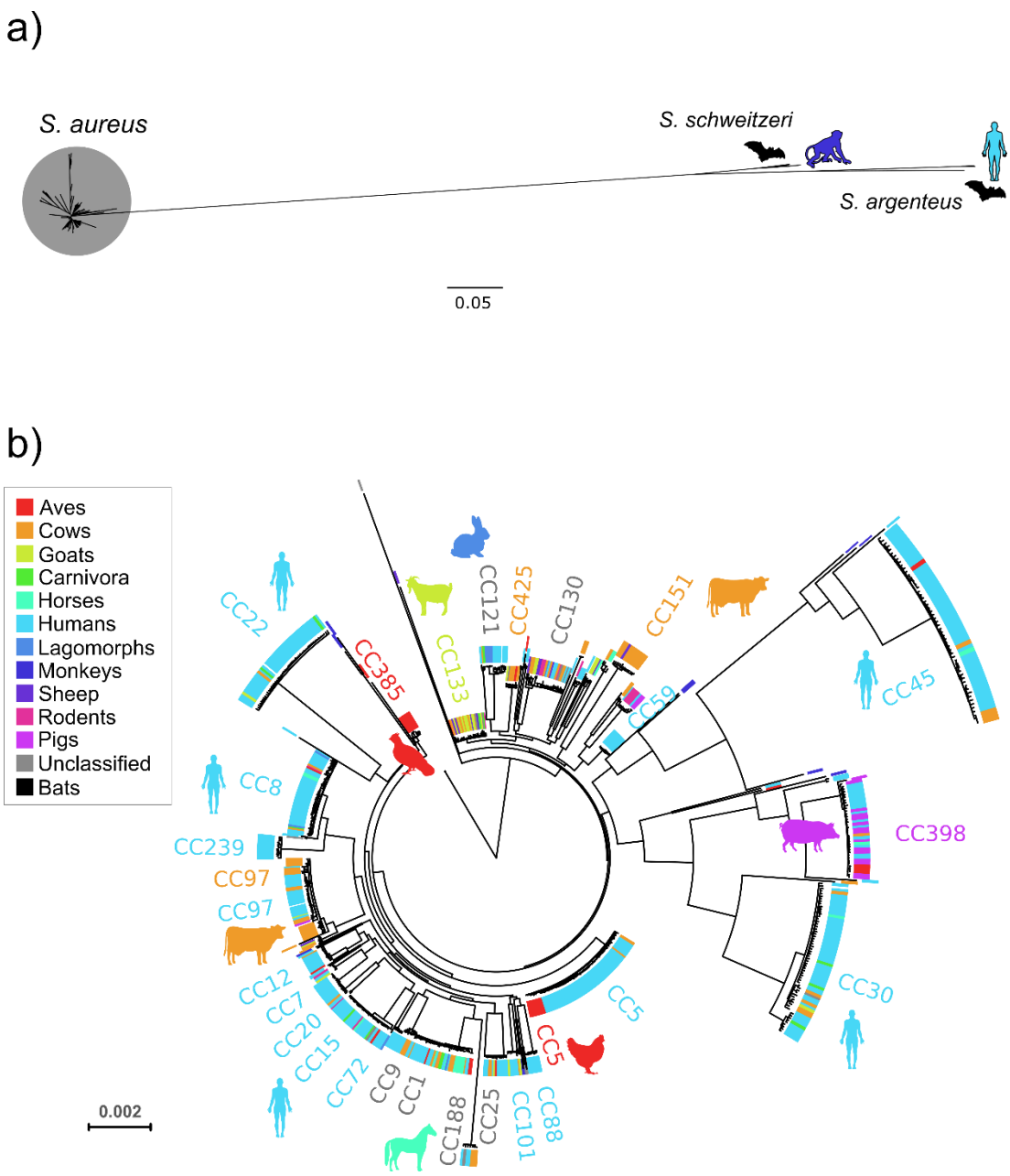


Figure 2.1. Evolutionary history of *S. aureus* and host associations. a) Maximum likelihood phylogenetic tree of 800 isolates included in the study. *S. schweitzeri* and *S. argenteus* are the closest species to *S. aureus*, with monkeys, bats and humans as main hosts. b) ML tree of 783 *S. aureus* isolates with major clonal complexes (CC) and respective host-associations displayed.

Removal of strains from the divergent clades resulted in 783 isolates with a core genome of 86,129 SNPs. The maximum likelihood phylogenetic tree (Figure2.1b) revealed a population structure largely consistent with previous work showing two major groups in *S. aureus* (Feil *et al.*, 2003; Everitt *et al.*, 2014). Group one included the well characterised CC8, CC239, CC15 and CC5, and group two contained the CC45, CC30 and the isolate JKD6159, the most divergent *S. aureus* strain described so far, belonging to the ST93 (Chua *et al.*, 2011). In addition, the phylogeny segregated according to the clonal complexes and sequence types defined by MLST typing, providing information about their intra-clonal relationships. Nevertheless, as we did not remove recombinant fragments from the genomic alignment, the relationships between lineages in our tree and previous phylogenetic reconstructions accounting for recombination are slightly different (Everitt *et al.*, 2014).

To investigate the host-species diversity across different clonal lineages, we mapped the host-types associated with every isolate on the tree (Figure2.1b). Three main patterns of host-association were revealed, of which the most common was individual lineages associated with a single host group. These included the major clades of human origin with expansion of several epidemic hospital (HA) and community-associated (CA) clones, including CC8, CC22, CC30 and CC45 (Aanensen *et al.*, 2016). Similarly, other clones contained isolates only found in animals, including the CC151, CC385 and CC133, respectively associated with cows, wild birds and small-ruminants (sheep and goats). In contrast, clades CC130, CC1 and CC188 presented isolates colonising a wide range of hosts, indicating these lineages are generalists. These isolates may harbour certain genetic features for infecting a wide number of hosts or lack host-specificity factors that narrow their capacity to infect single host-species. A

third group of clades were primarily associated with two host-types. For example, CC5 was mainly linked with humans and poultry, CC97 with humans and cows, and CC398 with humans and pigs. In the first two cases, host segregation exists within the clades and represent two well documented host jumps; CC5 from humans to chickens around 50 years ago (Lowder *et al.*, 2009), and CC97 from cows to humans 90 years ago (Spoor *et al.*, 2013). In contrast, the CC398 presented a mixed host distribution, suggesting several host switches during the evolution of this clade occurred, many of which represent spillover events from pigs to farmers (Voss *et al.*, 2005).

2.4.2. Humans represent the major hub for the emergence of epidemic *S. aureus* clones

We used AdaptML to predict ancestral host associations on the high-resolution ML tree and infer the frequency of host switching events that have occurred during the evolutionary history of *S. aureus*. AdaptML performs a habitat learning process of the isolates by using their ecotypes, i.e. host-species associations, and the evolutionary relationships between them. In order to optimize the analysis, we initiated the program with various sets of values in the input parameters and checked convergence in the number of habitats predicted by the model. Using a range of initial number of habitats, converge thresholds and collapse thresholds as input, AdaptML reached high convergence and consistently inferred two or three habitats (Figure 2.2).

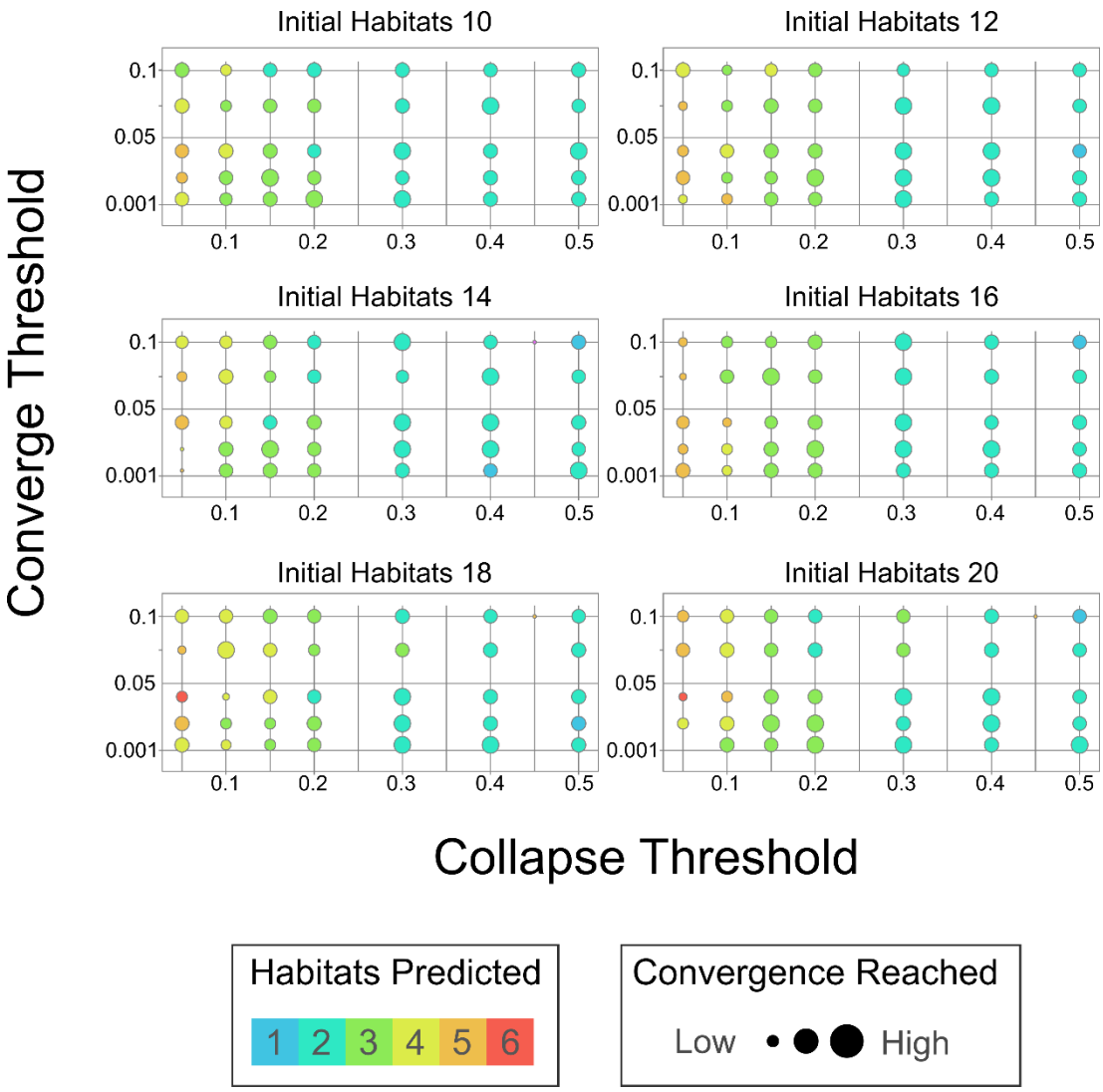


Figure 2.2. Number of habitats predicted by AdaptML. Optimization of AdaptML habitats prediction analysis using different combinations of input parameters. The analysis reached the highest convergence with two or three habitats being predicted. Different initial number of habitats and converge threshold values hardly had an effect on the number of habitats predicted, which were mostly affected by the collapse threshold parameter.

While the initial number of habitats had little impact on the final number of habitats predicted, the collapse threshold initial value segregated the ecological distributions, with higher values leading to fewer habitats, and with the default value of 0.1 predicting three habitats. The habitats inferred represent meaningful clusters of genotypes with an evolutionary and ecological significance. We projected the habitats onto the host-type dimension in order to reveal the distribution of ecotypes and understand the composition of the niches predicted. The distribution of hosts represent emission probabilities of genotypes being found in those habitats given their source of isolation (Hunt *et al.*, 2008). Assuming the existence of three habitats, each one was associated with major host-groups (Figure 2.3): (1) human habitat, with over 92% of genotypes isolated from persons; (2) bovid-associated habitat, including 60% of isolates from cows, sheep and goats and (3) broad-host habitat, which comprised most birds, pigs, horses and isolates from other host-species. In contrast, in the two habitats scenario, one group was strongly linked with humans (emission probability of 86%) and the second group with animals (probability of 78%), as previously inferred by (Shepheard *et al.*, 2013). All the habitats predicted contained genotypes isolated from other host-species, which is biologically meaningful given the zoonotic nature of *S. aureus*, as these may represent transient host transmission events.

AdaptML predictions are remarkable considering the algorithm has no previous knowledge of which host groups belong to humans or animals, nor which groups of animals belong to the same taxonomical category. While the two habitats results are comparable to previous observations (Shepheard *et al.*, 2013), our three habitats

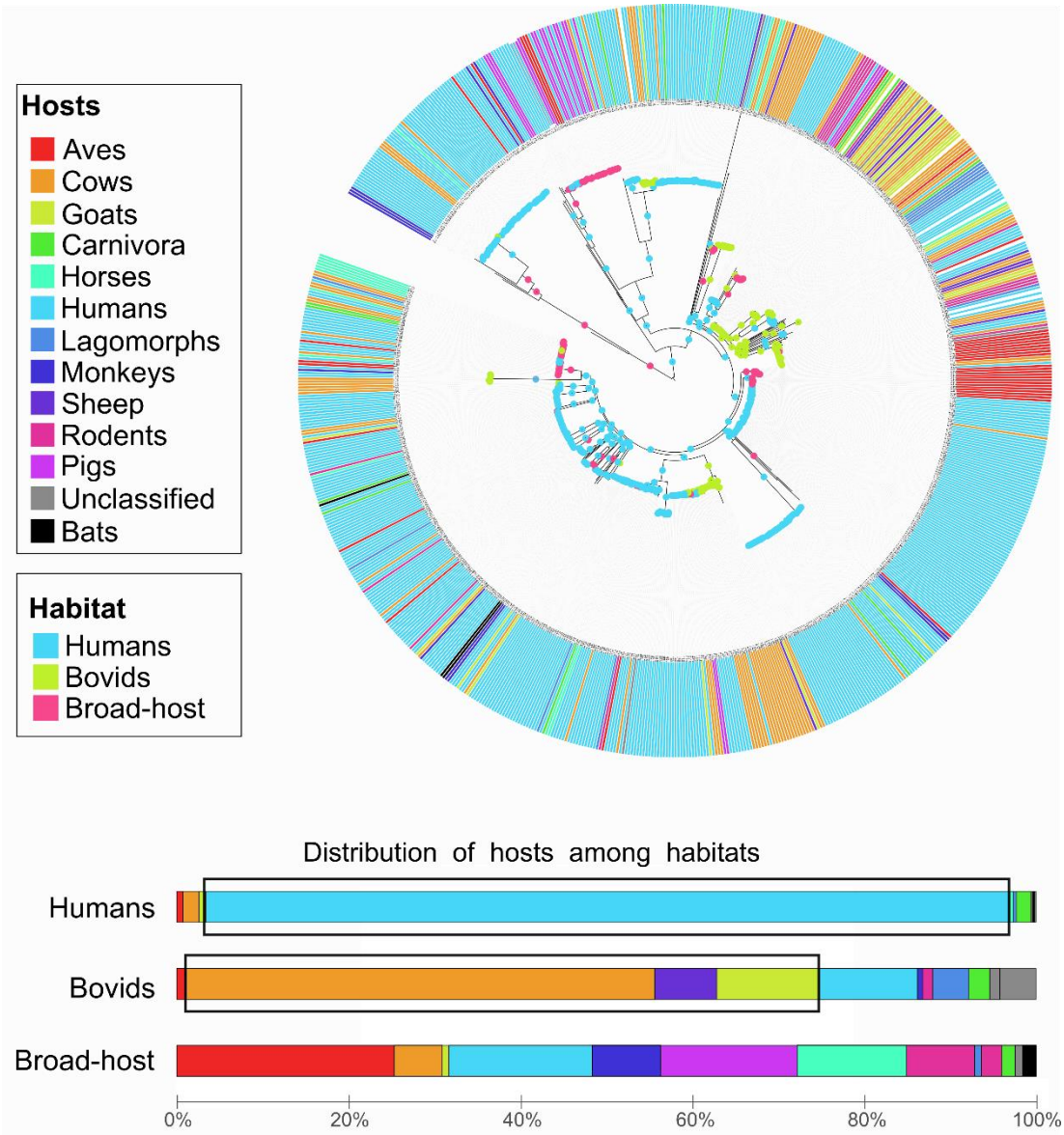


Figure 2.3. AdaptML with three habitats predicted. The ML tree of 783 isolates with a range of human and animal hosts. Nodes colours on branches correspond to habitat associations inferred by AdaptML. External bars represent the host-species from which genotypes were isolated. The distribution of hosts among habitats show one is mostly associated with humans, another one is mostly linked to bovids or ruminants (including cows, sheep and goats) and a third broad-host habitat.

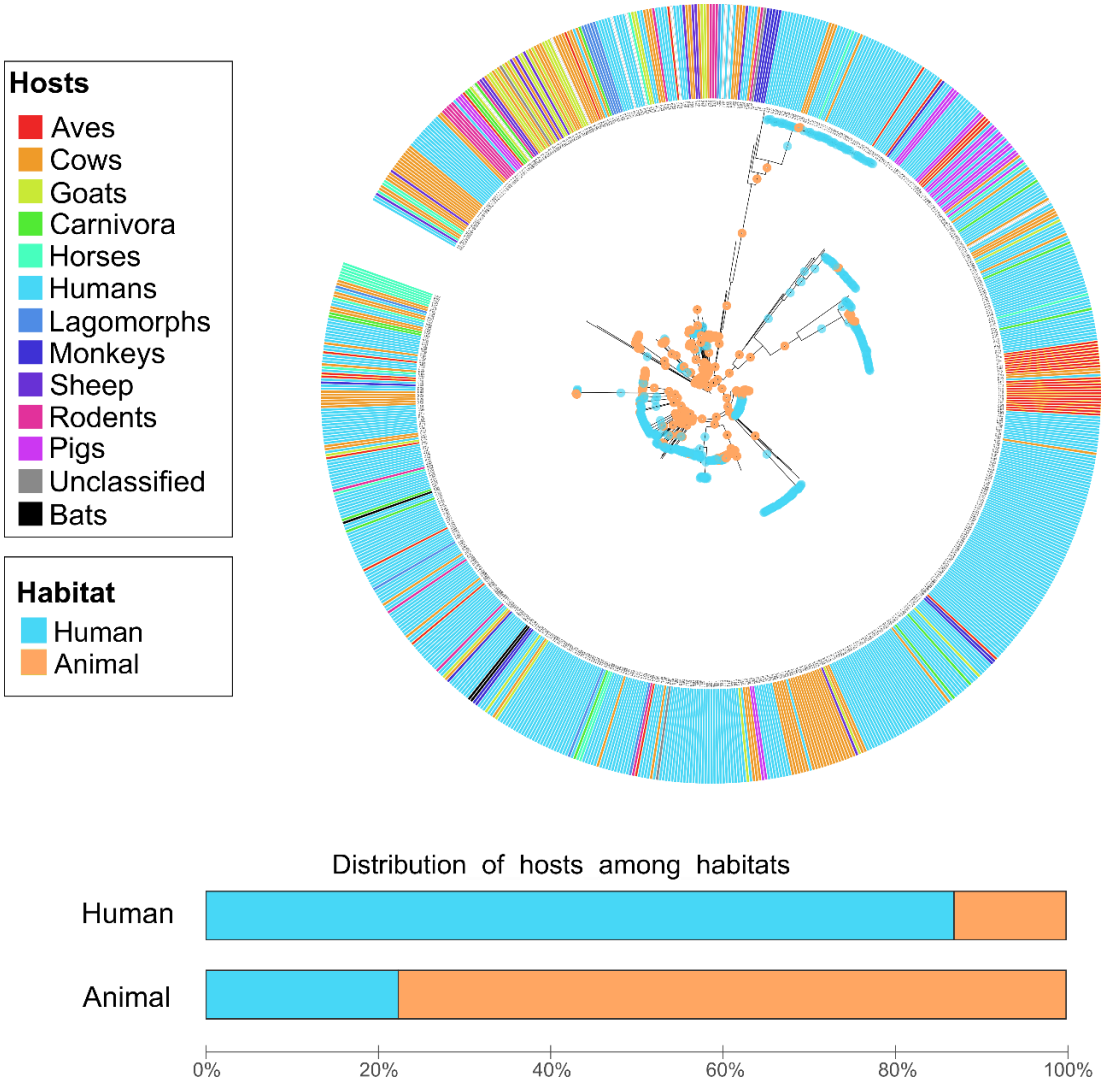


Figure 2.4. AdaptML with two habitats predicted. The ML tree of 783 isolates, nodes colours on branches correspond to habitat associations inferred by AdaptML. External bars represent the host-species from which genotypes were isolated. Projection of habitats onto host state reveals one habitat associated with humans and another one with animals.

analysis can be considered more robust because AdaptML was able to identify the bovid habitat, containing cows, sheep and goats, all belonging to the same taxonomical group. In addition, the distribution in the three habitats scenario represents more accurately the current knowledge of *S. aureus* host-species association and the dynamics of historical host-jumps studied (Lowder *et al.*, 2009; Spoor *et al.*, 2013; Weinert *et al.*, 2012; Price *et al.*, 2012).

Since AdaptML predicts current and past habitat states and maps them on every node of the phylogenetic tree, we counted the transitions between habitats along the branches on the tree to estimate the number of host jumps. In this way, we predicted overall ancestral jumps rather than recent host-switch events. For the two habitats prediction, an animal origin was inferred for *S. aureus* phylogeny and most ancestral nodes were assigned to the 'animals habitat'. Thus, the estimated number of transitions from animals to humans was 23, much higher than the 10 humans to animals jumps. Considering the more reliable analysis that predicts three habitats, the ancestral state for *S. aureus* was placed in humans, and during the evolution, 8 ancestral jumps from animals to humans would have occurred, compared to the 22 jumps from humans into animals. This is consistent with previous studies (Weinert *et al.*, 2012) and suggests that humans represent the predominant donor species, with host-jumps identified from humans into the other host groups examined. The most common recipient for *S. aureus* jumps from humans was ruminants, which, in turn, represented a major donor for host-switching events back into humans. Thus, these data indicate that humans represent the likely donor species for the emergence of contemporary endemic *S. aureus* clones responsible for diseases in livestock and poultry.

2.4.3. GWAS analysis reveals accessory genes associated with host-species

Several previous studies have identified core genome variants and mobile genetic elements (MGEs) of *S. aureus* with host-specific functional activity (Koop *et al.*, 2017; Viana *et al.*, 2015). A major aim of this study was to characterize genomic features that contribute to host-species adaptation and that may have been acquired by independent lineages during parallel evolution.

In order to identify genetic polymorphisms associated with specific host-types, we performed GWAS tests on the core genome of all the *S. aureus* isolates using ROADTRIPS (Thornton and McPeek, 2010). We tested SNP associations for the four main groups of hosts, including humans, ruminants (cows, sheep and goats), birds (poultry and wild birds) and pigs. ROADTRIPS accounts for known and unknown population structure and works by comparing groups of genotypes with a particular host phenotype (cases) to isolates from other groups (controls). To prevent the identification of predictive SNPs related with ancestry rather than host association, we defined population structure using a combination of two strategies; BAPS in a hierarchical manner (Chewapreecha *et al.*, 2014) and clusters obtained from the phylogenetic tree (Alam *et al.*, 2014). BAPS classified the 783 isolates into several clusters, and those clusters representing distinct clades of the tree were divided into subgroups. As a result, the isolates were stratified in 48 families of 1 to 106 individuals each. Next, ROADTRIPS was run for each host-group of interest, with variable numbers of cases and controls.

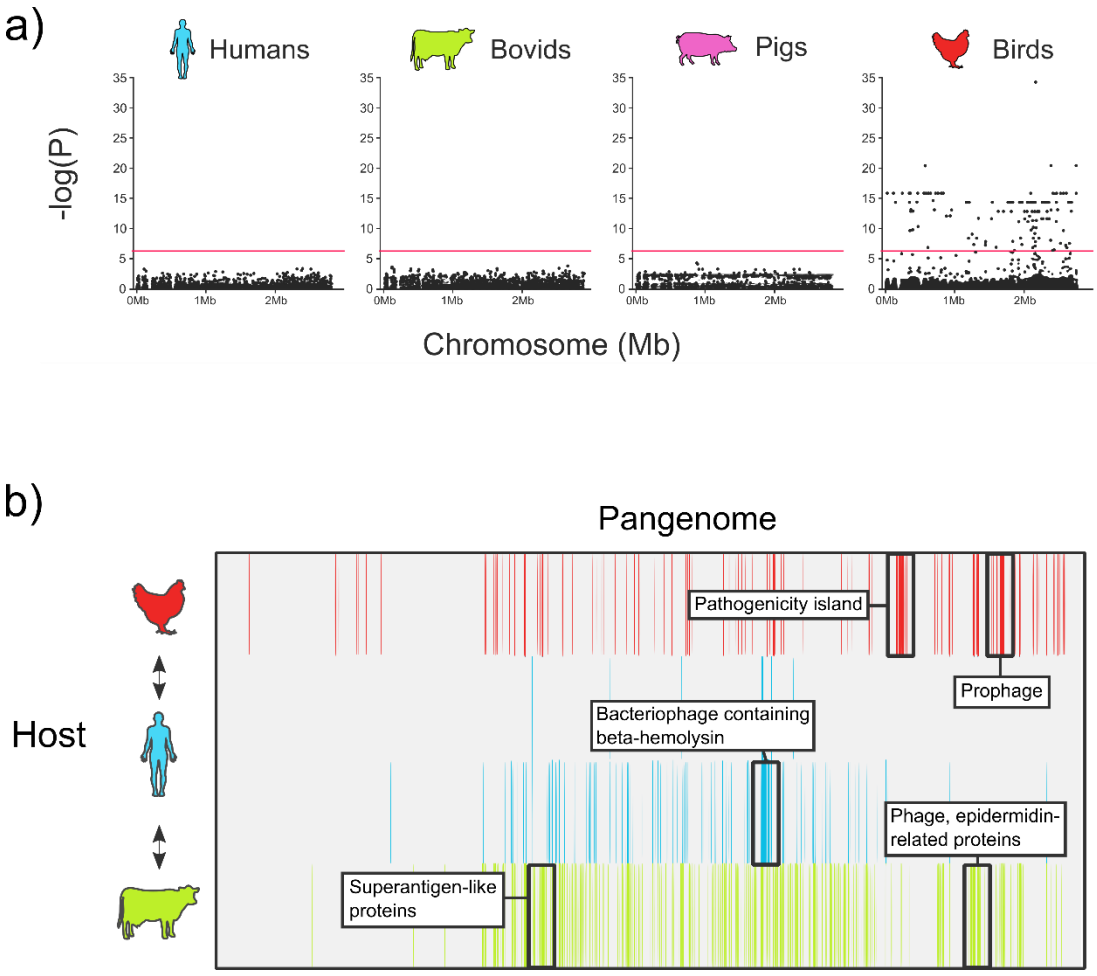


Figure 2.5. Genome-wide association analysis. a) Manhattan plots of GWAS for humans, ruminants, pigs and birds on the core genome of 81,680 biallelic SNPs. P values (-log10) are represented against chromosomal location. GWAS failed to identify SNPs associated with particular host-groups. b) K-mers enriched in birds, humans and ruminants using the panGWAS approach were mapped against the pangenome sequence. Pairwise comparisons are represented: birds versus humans (top), humans versus birds (second row), humans versus bovids (third row) and bovids versus humans (bottom). Regions with contiguous enriched k-mers represented genome fragments associated with host-adaptation, which included pathogenicity islands and prophages.

After controlling for population structure and correcting for multiple testing, none of the 81,680 biallelic SNPs in the original matrix were significantly associated with any of the four hosts (Figure 2.5a). Only birds presented SNPs with significant p-values but further examination revealed that these were a consequence of population structure, which was not possible to correct due to the small number of type 1 alleles in controls. These results suggest that convergence in the genomic variants of the core genome is not an evolutionary path used by *S. aureus* for adapting to specific hosts. However, the limited number of genomes included in the analysis likely limited the power of detection.

In order to investigate host adaptive genomic features in the accessory genomes of human and animal isolates, we employed a pangenome-wide association approach to search for genes enriched in specific host-species (Lees *et al.*, 2016). We ran SEER to identify k-mers enriched in humans versus ruminants and in humans versus birds, revealing multiple short sequences associated with each of these host groups. Mapping these k-mers against the pangenome revealed multiple hits broadly distributed (Figure 2.5b) and as the pangenome was constructed maintaining synteny, k-mers grouping together represented gene sets. These primarily corresponded with mobile genetic elements, including phages and pathogenicity islands; and clusters of proteins related to virulence, suggesting a potential role for those genes in the host-species ecology of *S. aureus* (Supplementary Table 3). In the avian-human comparison, the lack of k-mers enriched in humans and the multiple bird hotspots is in line with our understanding that *S. aureus* jumped from humans to chickens acquiring novel mobile genetic elements from an avian specific gene pool (Lowder *et al.*, 2009). Conversely,

multiple k-mers were detected for both host-types in the human-ruminant comparison, consistent with MGE independently acquired and lost during multiple host jumps between these two host-species (Guinane *et al.*, 2010; Spoor *et al.*, 2013; Weinert *et al.*, 2012). Several of the MGEs have previously been identified among host-specific clones and demonstrated to encode proteins with host-specific activity. For example, the β-haemolysin converting phage ϕSa3 encodes modulators of the human innate immune response, and pathogenicity islands encoding superantigens or von Willebrand factor-binding proteins with ruminant-specific activity were identified among cow and sheep strains of *S. aureus* (Deringer *et al.*, 1997; Fitzgerald, 2012; van Wamel *et al.*, 2006).

Taken together, these analyses suggest that *S. aureus* host-specificity is mediated through specific pools of genes in the accessory genome rather than common genetic polymorphisms in the core genome.

2.4.4. Long-term refinement of host adaptation involves diversification of distinct biological pathways

In addition to accessory genes, adaptive mutations in the core genome may be selected for in response to environmental changes such as antibiotic exposure or a switch in host-species (Howden *et al.*, 2014; Viana *et al.*, 2015). In order to examine the impact of host-species on diversification of the *S. aureus* core genome, we identified groups of related isolates (e.g. within CCs or STs) associated with a specific host-species for genome-wide analysis of positive selection. 15 groups of isolates were selected representing 9 of human origin, 3 from ruminants, 2 from birds and one of pig origin.

Positive selection was identified across all host-associated groups examined, with an average of 68 genes (33 to 129) representing approximately 2.7% (1.3% to 5.14%) of a clade-specific core genome (Table 2.1). Birds and pigs related clades presented the lowest numbers of genes positively selected, with 1.62% and 1.72% respectively, while this value increased up to 3.91% for ruminants. Although there was variation in the proportion of the genome exhibiting positive selection for groups associated with the same host-species, within-host variance of positively selected genes was smaller than the observed variation between hosts (Figure 2.6).

The variation in the proportion of genes under selection for groups associated with the same host-species was not proportional to the number of years elapsed since the ancestral host jump that gave rise to those clonal complexes. For example, the CC97 and CC133 lineages, with inferred dates for the most common ancestor 1,832 and 3,133 ya respectively, presented 4.74% and 5.22% of genes under positive selection, while the CC151, with a much earlier origin (5,429 ya), only contained 2.11% of genes under positive selection (Spoor *et al.*, 2013; Weinert *et al.*, 2012). These discrepancies are due to the clonal structure of *S. aureus* and sampling bias, since the branch length of the CC151 was much longer than the branches of the CC97 and CC133, but the node leading to the isolates used in the analysis was more recent, and therefore fewer genes under selection were detected.

Table 2.1. **Genes under positive selection.** Groups selected for the positive selection analysis, with core genome sizes and number of genes under selection.

	Core genome	Positive selection	Proportion
CC12_Humans	2444	36	1.47%
CC15_Humans	2406	64	2.66%
CC22_Humans	2596	83	3.20%
CC45_Humans	2488	61	2.45%
CC59_Humans	2395	42	1.75%
CC5_Humans	2456	86	3.50%
ST239_Humans	2571	67	2.61%
ST30_Humans	2558	85	3.32%
ST8_Humans	2545	87	3.42%
CC151_Cows	2554	51	2.00%
CC97_Cows	2408	111	4.61%
CC133_Caprids	2512	129	5.14%
CC385_Wildbirds	2419	47	1.94%
CC5_Poultry	2532	33	1.30%
CC398_Pigs	2445	42	1.72%

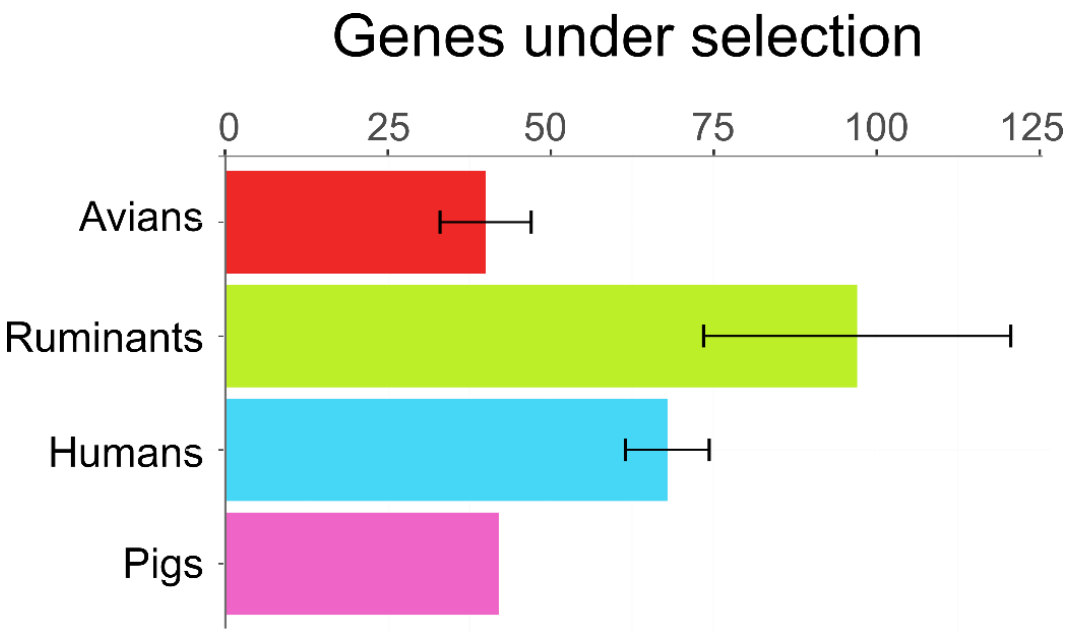


Figure 2.6. Genes under positive selection in different hosts. Within-host-type variance of positively selected genes was smaller than the observed variation between host-species, but observations were not statistically significant (ANOVA p-value = 0.0682).

A limited number of genes were positively selected across multiple host-species, including several encoding membrane proteins, lipoproteins and a protein involved in biofilm formation. These genes may have a role in host-species independent immune evasion as opposed to host adaptation. Some genes were identified in distinct lineages that were associated with the same host-species (mostly human), suggesting strong selective pressure leading to convergent evolution. The majority of these genes encoded proteins predicted to be located in the bacterial cell envelope or involved in pathogenesis. In ruminant strains, positively selected genes encoded proteins involved in DNA replication, metabolism and pathogenesis, while in bird strains genes under positive selection primarily encoded hypothetical proteins. However, for the most part, our analysis detected distinct sets of genes under positive selection in different lineages, suggesting that signatures of host-adaptation are dependent on the genetic background of the strain, and that host-adaptation can be achieved via multiple trajectories through modification of distinct pathways.

We predicted functional categories of genes under positive selection by analysis of Clusters of Orthologous Groups (COG) and Gene Ontology (GO). This revealed several functional groups that were enriched for positively selected genes independently of the host-species (Figure 2.7), including genes linked to the host-pathogen interface, immune evasion and maintenance of MGEs (Kalia and Bessen, 2004; Petersen *et al.*, 2007). However, the majority of the functional categories were host-species dependent, consistent with distinct mechanisms underpinning adaptation to different host-species (Figure 2.8). For example, genes under positive selection in

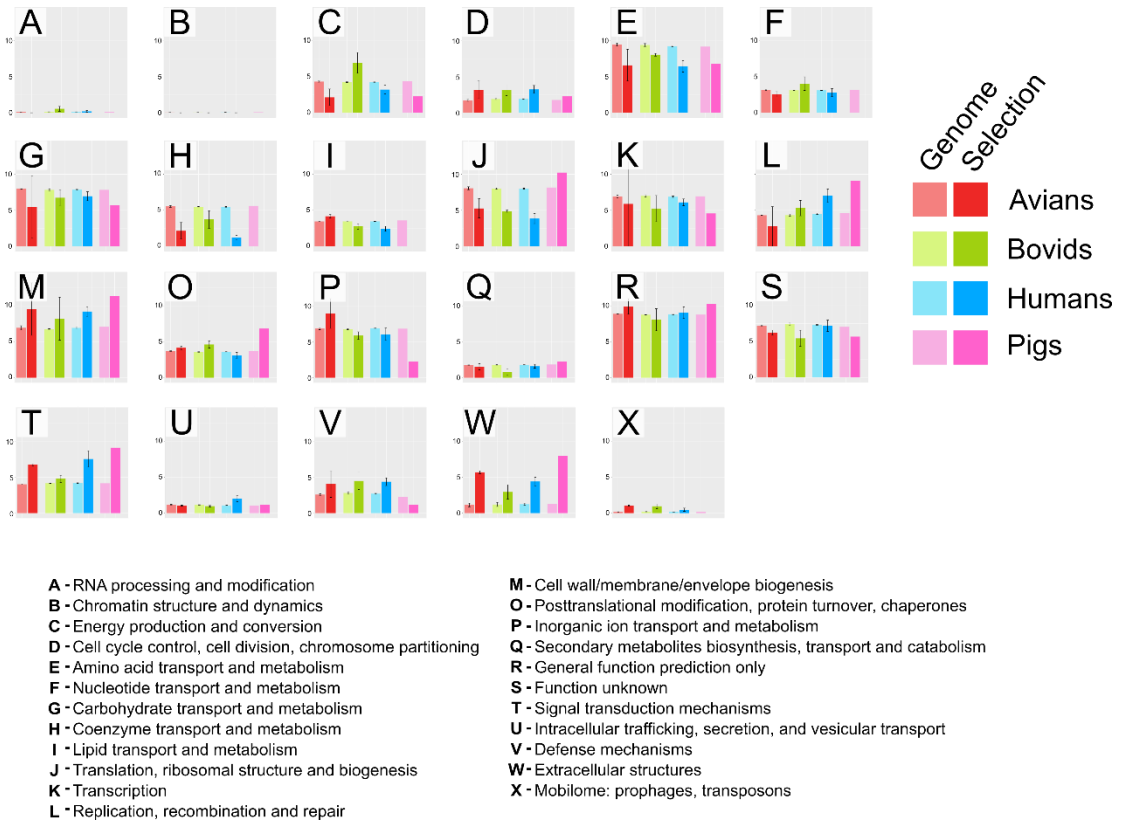


Figure 2.7. Differences in COG functional categories for genes under positive selection in different hosts (selection) compared to all the genes for the respective host groups (genome). Bars represent the average proportions of genes in each functional category compared to the host group.

human associated *S. aureus* tended to be linked to metabolism and biosynthesis of glycerolipids, cellular amino acids and metabolism of nucleotides. Glycerolipid biosynthesis genes influence levels of cardiolipin, critical for *S. aureus* prolonged survival under conditions of high salinity, during growth on skin and mucus membranes and survival on dry surfaces during indirect transmission (Tsai *et al.*, 2011) and glycerophospholipids, previously demonstrated to be involved in resistance to antimicrobial peptides (Peschel *et al.*, 2001). Identification of positive selection acting on genes involved in amino-acid biosynthesis could be a metabolic adaptation driven by the limited availability of key nutrients in distinct anatomical niches in humans.

Genes associated with transport of carbohydrates demonstrate signatures of positive selection. In ruminant strains, this is particular true for genes involved in the transport of disaccharides and oligosaccharides, and we note that lactose (a dissacharide) is the primary carbohydrate available in the cow or sheep udder. In humans, glycosamino-glycans such as hyaluronic acid are the main source of carbohydrates available on human skin, and in this case carbohydrate transport genes impacted by positive selection are more commonly associated with monosaccharides. In addition, the requirement for acquisition of iron and its sequestration in distinct niches in different host-species is reflected by the specific iron-acquisition systems that exhibit footprints of diversifying selection. These included biosynthesis of the siderophore aerobactin and citrate transporters, which recruit iron by binding to it, and systems involved in transport of iron.

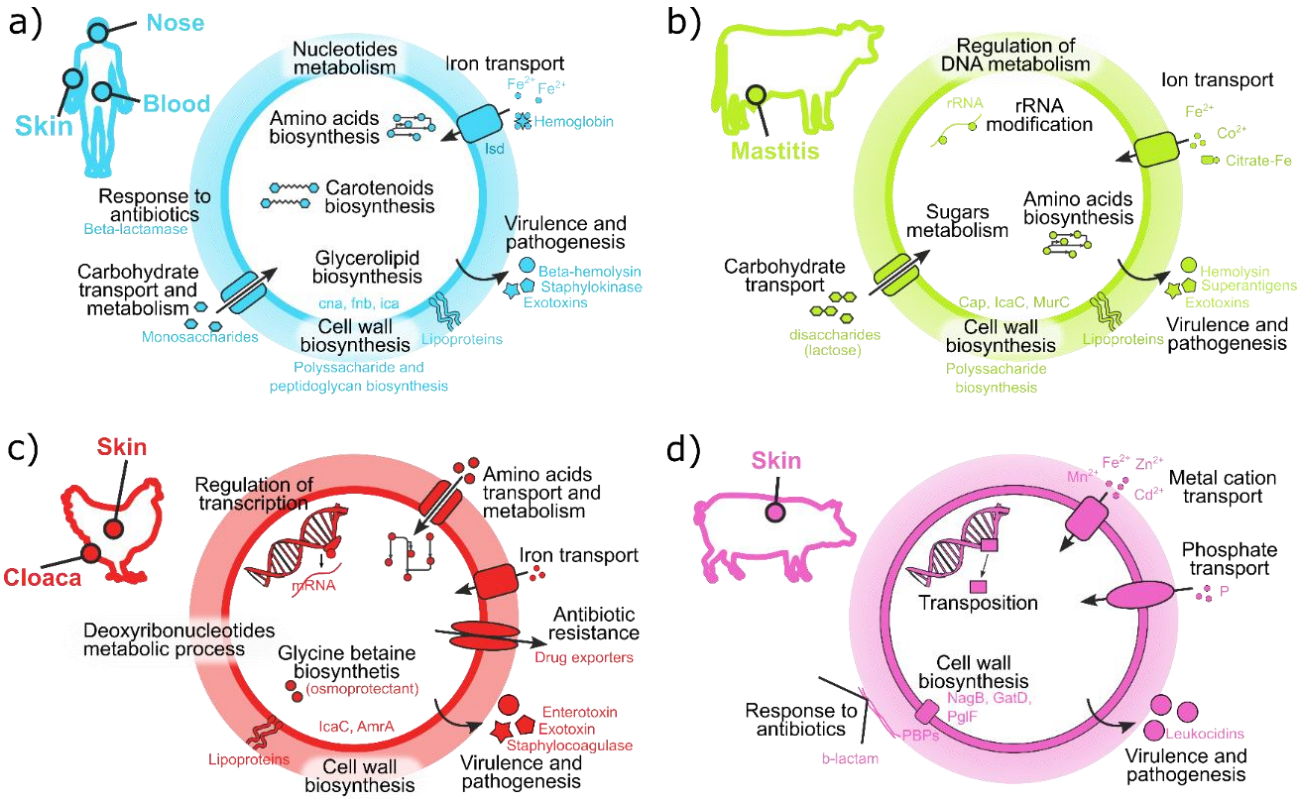


Figure 2.8. Schematic representation of selected biological pathways under positive selection in different host-species. The main anatomical isolation sites on each host group are indicated by filled circles. Functional categories virulence and pathogenesis, resistance to antibiotics, transport of ions and cell wall biosynthesis were under positive selection in all 4 host-species groups. In humans (a) and ruminants (b) the categories amino acids biosynthesis and transport/metabolism of carbohydrates were positively selected. The categories amino-acid transport/metabolism and biosynthesis of osmoprotectants were under positive selection in birds (c) and transposable elements in pigs (d).

S. aureus from humans, ruminants and birds also contain genes associated with pathogenesis that are under diversifying selection. For ruminant derived isolates these included genes responsible for the production of exopolysaccharides which is involved in biofilm formation, and strongly linked to bacterial survival and resistance to antibiotics during bovine mastitis (Melchior *et al.*, 2006).

In birds, *S. aureus* is mostly present under the feathers on the skin and in the nasal cavities (Devriese *et al.*, 1972; Rosenthal *et al.*, 2009). Notably, some of the biological pathways under diversifying selection in avian *S. aureus* are associated with the transport and metabolism of amino acids or products of their degradation such as ketones or phenol catechol. This may reflect the fact that feather keratins differ in composition and amino-acid usage in comparison to mammalian keratins (Harrap and Woods, 1964). Accordingly, there are differences in nutrient availability that may require attenuation of amino acid biosynthesis pathways depending on nutrient availability at colonisation sites in birds. In addition, Staphylococci are also part of the normal microbiota of the mucosal surface of the cloacae in birds (Bowman and Jacobson, 1980; Rosenthal *et al.*, 2009). Allantoin, mannitol, catechol and organic acids are secreted by birds via the urinary tract (Quebbemann and Rennick, 1968) and genes associated with the metabolism of these products all indicated positive selective pressure in avian *S. aureus* strains only.

S. aureus from pigs had fewer genes exhibiting signatures of positive selection, which may reflect the fact that only a single group of isolates (ST398) was examined and that the host-jump from humans into pigs occurred relatively recently (Figure 2.1). For

pigs, the most affected functional category was related to DNA recombination and transposition and antibiotic resistance, consistent with the large volumes of antibiotics used in the pig farming industry (Visschers *et al.*, 2014).

Finally, we reconstructed the enriched metabolic pathways of genes under positive selection in different hosts. Human strains presented the widest range of pathways affected by positive selection, mostly involving metabolism of carbohydrates, amino acids and energy. Pathways under diversifying selection in ruminants partly overlapped with those in human strains, but nucleotide metabolism was specially affected in ruminants. Positive selection in bird isolates affected fewer pathways and were more sparsely distributed, reflecting in part the lower number of genes under positive selection identified (Figure 2.9).

In summary, *S. aureus* host-adaptive evolution is mediated by selection acting on a variety of biological pathways, enhancing survival in specific microenvironments via innate immune evasion and nutrient acquisition

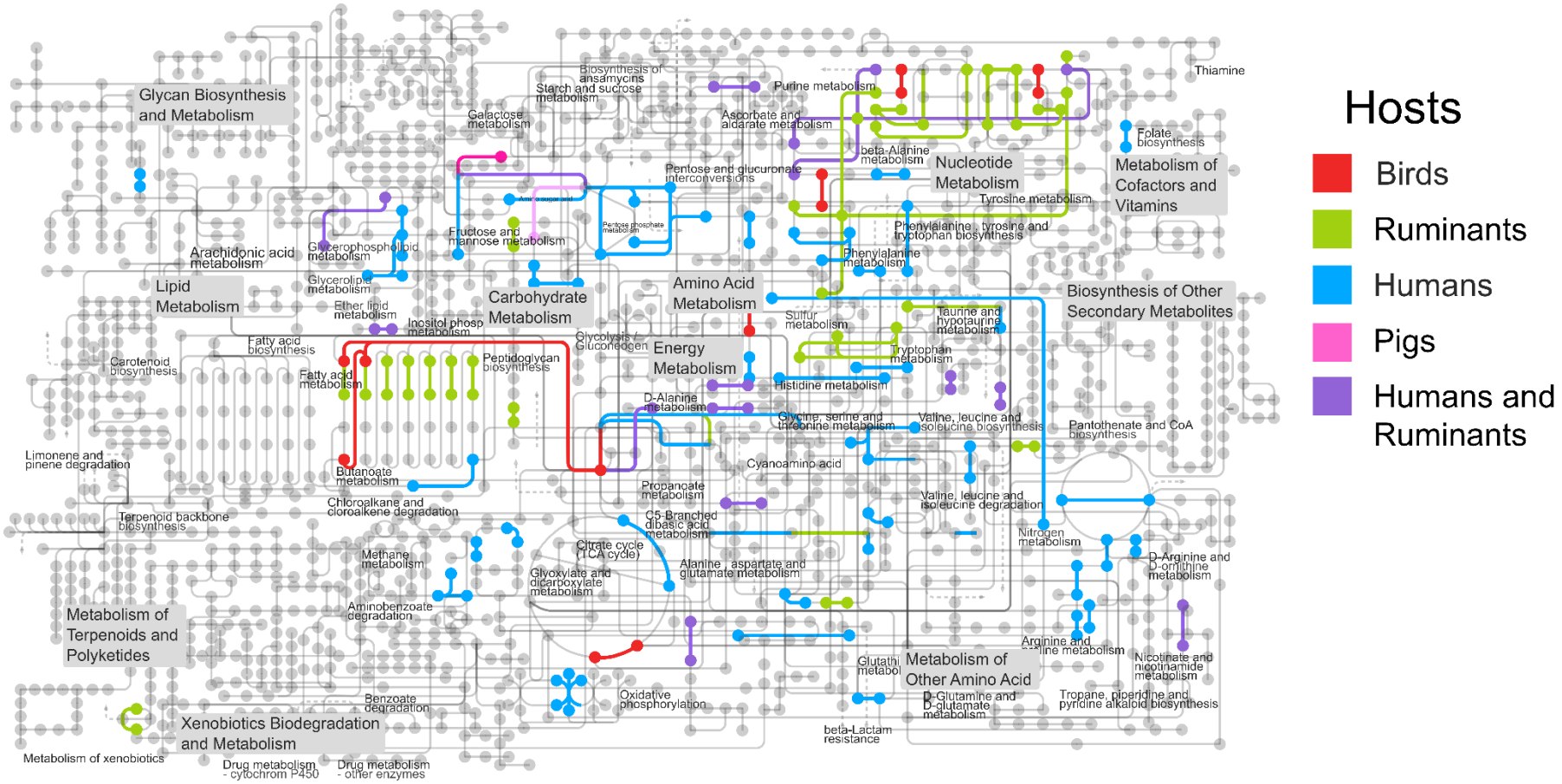


Figure 2.9. Metabolomes under positive selection in different host-species mapped onto KEGG global metabolic pathways. Nodes in the figure represent metabolic compounds and edges are enzymatic reactions. Pathways under positive selection uniquely in specific hosts are coloured as indicated in the legend. Humans and ruminants presented some common routes affected by positive selection (purple).

2.5. Discussion

In this study, we investigated the evolutionary history of *S. aureus* in the context of its host-association. Using whole-genome sequences of hundreds of isolates representative of the clonal diversity associated with a wide range of host-species, we examined the dynamics of *S. aureus* host-switching with unprecedented resolution. In addition, we provided new insights into the genetic basis of *S. aureus* adaptation to different host-species.

Many new pathogens emerge following zoonotic or anthroponotic events, providing the opportunity for spread within a new host population. *S. aureus* is considered a generalist bacterial species, capable of colonizing a wide range of hosts. However, our phylogenomic analysis revealed that the majority of isolates associated with animals clustered into animal sub-lineages, and most of the human epidemic hospital-associated and community-associated clones were part of human-specific lineages, supporting that distinct sub-lineages are associated with particular hosts or host-groups (Sung *et al.*, 2008). This stratification of the clonal host-specificity is important for public health authorities in order to assess the relevance of future emerging epidemic clones. Knowing if a novel host-type emerging clone descends from a host-restricted or generalist lineage can provide information on the level of adaptation to the new host-species, which is important to design measures to limit the expansion of those clones. Our current understanding of *S. aureus* host-association is limited due to sampling bias or temporal changes in the *S. aureus* population structure in specific hosts, as many clades previously associated with specific host-groups have been

reported in new species in recent years (Espinosa-Gongora *et al.*, 2014; Spoor *et al.*, 2013).

Given the above, *S. aureus* represents an excellent model to explore the dynamics of a bacterial pathogen at the human-animal interface. Using a statistical approach based on a Markov model, we classified isolates into ecologically similar habitats, which allowed us to predict ancestral host-association states. We demonstrated that the segregated host-specialism of *S. aureus* arose via multiple cross-species transmission events, leading to the emergence of successful endemic and epidemic clones circulating in distinct host-species populations. We identified humans as the major hub for the spread of *S. aureus* to livestock, reflecting the role of humans in domestication of animals, and subsequent opportunities for cross-species transmission events. These results are consistent with previous research using a similar approach based on MLST data (Shepheard *et al.*, 2013). Importantly, we also identified cows as the major reservoir for the emergence of *S. aureus* clones that are epidemic in human populations. However, given the differences in the AdaptML results using two and three habitats, other approaches such as Bayesian phylogenetics could be more suitable for predicting host jumps and ancestral states. In addition, our study is limited by the sampling bias towards human isolates from Europe, and deeper sampling of other geographical locations and adding isolates from more wild animal species would strength future population genomic studies on the transmission dynamics of *S. aureus*.

To investigate the molecular basis for host-adaptation, we used GWAS approaches on the core SNPs and on the accessory genome k-mers, leading to the detection of

genomic signatures associated with host-groups. The identification of MGEs associated with specific host-species provides compelling evidence for the key role of horizontal gene acquisition in the adaptation of *S. aureus* to their hosts. While several MGEs have previously been identified to be associated with host-specific clones (Guinane *et al.*, 2010; Lowder *et al.*, 2009; Spoor *et al.*, 2013), our pangenome-wide analysis reveals new combinations of putative host-adaptive genes providing new avenues for investigating mechanisms of bacterial host-adaptation. Nevertheless, we did not search for potential genomic differences between isolates from host-restricted and generalist lineages, and such investigations could provide traits associated with the capacity for a multi-host ecology.

Although a single natural nucleotide mutation was previously identified to be responsible for a human to rabbit host-jump (Viana *et al.*, 2015), we were not able to identify any SNP underlying association with humans, ruminants, pigs and birds at the entire *S. aureus* species level, indicating that generally such small changes in the genome do not mediate host-tropism. Unlike antibiotic resistance and bacterial toxicity, which are traits encoded in single genes or consequence of mutations in a small number of genes (Alam *et al.*, 2014; Laabei *et al.*, 2014), host-species association is likely a more complex trait involving larger scale genomic changes such as gene decay, gene duplication and MGE exchange. Importantly, genotype-phenotype association studies often require very large datasets to be powered to identify genetic variants underlying complex phenotypic variation (Power *et al.*, 2017). Given the bias in our dataset towards human isolates and small number of samples from most species, we had to perform the analysis on higher taxonomical groups, for example birds and

ruminants, as opposed to chickens and cows, potentially reducing the power of our analysis (Figure 2.10). Finally, linkage disequilibrium and extensive bacterial population stratification significantly reduce our ability to identify SNPs associated with an avian host tropism.

Finally, we identified evidence of adaptive evolution in the core genome including diversification of gene function. Although some functional categories of genes were under positive selective pressure in strains from all host-species including those associated with immune evasion, the host-pathogen interface and maintenance of DNA, the majority were host-species dependent suggesting host-specific selective pressure driving the diversification of biological pathways that are involved in survival or transmission. Furthermore, in some cases, distinct pathways were under positive selective pressure in different clones associated with the same host-species, implying that multiple distinct pathways may mediate host-adaptation depending on the genetic background of the strain. In particular, pathways linked to carbohydrate transport and metabolism, amino acid metabolism and iron acquisition exhibited signatures of host-adaptation.

Overall, these findings inform a model of *S. aureus* host-adaptation in which acquisition of a specific set of MGEs confer the capacity for survival in the new host, largely through targeting of the innate immune response. Subsequently, positive selection acts on the core genome via point mutation causing allelic variation that results in metabolic remodelling in response to distinct nutrient availabilities.

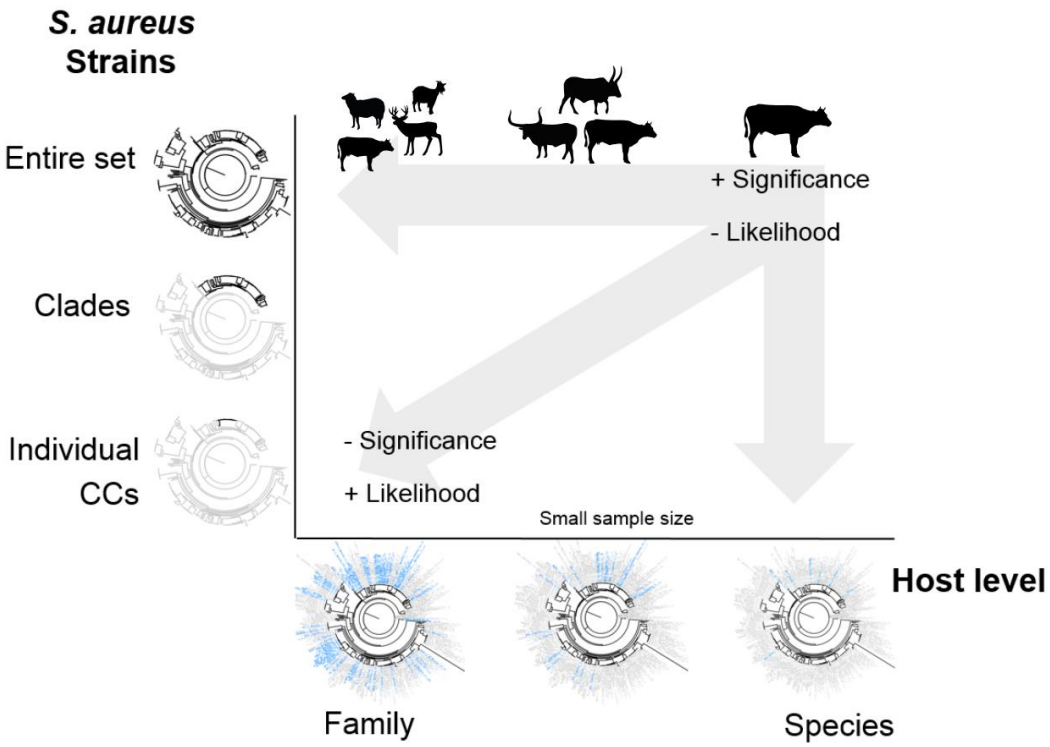


Figure 2.10. Selection of *S. aureus* taxonomic group and host-type level affect the power of GWAS statistical analysis. Genetic variants associated with a particular host-type and distributed across the entire dataset would present high significance of host-association (top-right side of the graph). These include, for example, variants found only in cow-associated isolates from multiple clonal complexes. On the contrary, variants identified only in a sub-lineage of the tree and associated with host-types of higher taxonomical level have a lower likelihood to explain host-association (bottom-left side of the graph). For instance, SNPs present in a specific clonal complex and linked with family of ruminants. Such clade-specific variants are a consequence of the population structure, and while being more frequently identified, they mostly represent spurious associations.

2.6. Conclusions

Taken together, our data provide broad insights into the evolutionary landscape of the *S. aureus* species, highlighting the evolutionary dynamics of *S. aureus* in the context of its host-associations. We generated a high-resolution view of the capacity for a model multi-host pathogen to undergo radical changes in host-species ecology by genetic adaptation.

Further investigation into the functional basis of these genetic changes will reveal key host-pathogen interactions that could be targeted for novel therapies. Further, the identification of the most common routes for *S. aureus* livestock-human host-species switches informs the design of more effective farm security measures to limit the emergence of new clones. These findings will be relevant to other major bacterial pathogens with the capacity to spread between livestock and humans.

3

Host-adaptive evolution during experimental infection in the face of regular bottlenecks.

3.1. Introduction

Staphylococcus aureus is a generalist pathogen, capable of colonizing and causing infections in humans and a variety of animal species (Weese, 2010). The majority of strains associated with animals belong to unique clades different from those infecting humans, and rarely cross species boundaries (Shepheard *et al.*, 2013). However, domestication of animals, agricultural industrialization and globalization have contributed for the transmission of bacteria between humans and animals, leading to the emergence of livestock-endemic clones in recent decades, causing important economic loses in the agriculture and food industry (Fitzgerald, 2012; Lowder *et al.*, 2009). For example, *S. aureus* intramammary infections in cows (Smith *et al.*, 2005), sheep and goats (Bergonier *et al.*, 2003) are associated with significant economic losses in milk production worldwide (Halasa *et al.*, 2009). In addition, the ability for *S. aureus* to switch host-species represents a major concern for global public health.

Our current understanding of *S. aureus* host-switch events is that following a transmission, *S. aureus* may undergo a series of genomic changes that increase its fitness in the new niche that leads to successful adaptation to the new host-species (Engering *et al.*, 2013). Several studies have revealed multiple genetic mechanisms for host-adaptation, including genetic diversification, gene decay and lateral transfer of genetic elements (Lowder *et al.*, 2009; Uhlemann *et al.*, 2012; Viana *et al.*, 2010). For example, Guinane *et al.* (2010) demonstrated that the CC133 clone, responsible for the majority of small ruminant infections, originated from a human-to-animal host jump, with the genome evolving by gene loss and diversification of proteins responsible for host-pathogen interactions (Guinane *et al.* 2010). In the same way, the *S. aureus*

bovine-associated CC8 clone originated from humans but adapted via loss of a β-haemolysin converting prophage and the acquisition of a new staphylococcal cassette chromosome (Resch *et al.*, 2013). In addition, gain or loss of mobile genetic elements was the primary explanation for a bovine to human host jump by a CC97 strain that led to emergence of a pandemic clone (Spoor *et al.*, 2013).

Although successfully describing host-associated genomic signatures by comparative genomics and phylogenetic analysis, to fully comprehend the process of adaptation to a new host-species, we need to investigate the short and medium-term evolutionary scales to understand the evolutionary changes that occur during the early stages of a host-switch. In addition, the capacity of *S. aureus* to colonize and persist in a new host-species depends on its interactions with the host. With the reduction of WGS costs, population genomics have been extensively applied for studying the evolutionary dynamics, transmission, and population structure of bacterial pathogens within individual host (Didelot *et al.*, 2013; Golubchik *et al.*, 2013; Lieberman *et al.*, 2013; Young *et al.*, 2012). In addition, the dynamics of evolutionary adaptation has been studied recently in bacteria (Barroso-Batista *et al.*, 2014; Søndberg and Jelsbak, 2016) and yeasts (Hong *et al.*, 2014; Lang *et al.*, 2013; Voordeckers *et al.*, 2015). However, these studies only examined naturally occurring hosts or very short times, and with the exception of research on plant pathogens (Guidot *et al.*, 2014; Meaden and Koskella, 2017), the dynamics of adaptation to a new host-species have not been investigated yet.

Finally, *in vitro* experimental evolution studies (Barrick and Lenski, 2013 for a review) in combination with computing simulations of evolution (Hindré *et al*, 2012; Hoban *et al.*, 2012) have contributed to clarify theoretical frameworks of genetic variation and evolutionary dynamics.

In the present study, we integrate *in vivo* experimental evolution, within-host population genomics and *in silico* evolution simulations for studying bacterial adaptation to a new species after a host-switch.

3.2. Aims

- To investigate the molecular mechanisms of *S. aureus* adaptive evolution to a new host-species.
- To examine the nature of the processes affecting the genome during short-term and long-term adaptation.
- To understand the within-host evolutionary dynamics during infections and transmissions.

3.3. Materials and Methods

3.3.1. Bacterial strains, sheep infections and *in vitro* passages[4]

We simulated *S. aureus* host-jumps by inoculating the mammary gland of sheep with two human-associated strains, NCTC8325 and N315. For the inoculations, 40 cfu from overnight clonal cultures were suspended in 1ml of PBS and injected into the teat

ducts. Sheep were kept isolated from their lambs for 4 h post-inoculation to prevent loss of bacteria due to suckling, but then were released into the farm. A few days after the inoculation, some ewes developed subclinical intramammary infections while others cleared the bacteria. We milked infected sheep every day and after 47 to 75 d, we simulated transmission by passing bacteria to a new sheep. For these passages, 50 to 100 cfus were picked from primary isolation plates and used to infect new animals as described before. This process was repeated for 165 to 400 d, depending on the strain and success of the infections, representing 4 to 7 passages. From the last isolation plates obtained from the milk samples, each of three colonies was picked for whole genome sequencing (Supplementary Table 4). *In vitro* passages of strains NCTC8325 and N315 were performed by preparing cultures in TSB and transferring 5 μl of each culture into 5 ml of fresh medium every 12 h. Passages were performed for 120 to 126 d and from the last tubes bacteria were plated and single colonies picked for whole genome sequencing (Supplementary Table 4).

3.3.2. Coinfection experiments and detection of infective isolates[4]

Coinfections of additional sheep with similar amounts of original (*wt*) and passaged strains, or with the original strain (*wt*) and a strain with a synonymous SNP in the vWbp pseudogene (*ss*) were performed as indicated above. Infected sheep were milked every day and after 40 d, bacterial cultures were prepared from the milk samples and individual colonies were identified. PCRs of genomic regions containing new variants

[4] Sheep infections, *in vitro* growing of strains and PCR analysis were performed by Jose Penades and Angeles Tormo.

(see below) and with the *ss* mutation were performed on 10 to 20 isolates for identifying the proportions of different clonal populations.

3.3.3. Genomic sequencing, assembly and annotation of genomes

Genomic DNA from overnight cultures of *S. aureus* isolates was extracted using the PurEluteTM bacterial genomic kit (EdgeBiosystems, MD) with modification as previously described (Lowder *et al.*, 2009). Illumina libraries were prepared with the Nextera XT kit for both MiSeq and HiSeq sequencing at Edinburgh Genomics. Two paired-end sequencing runs were produced for every isolate, obtaining reads of 100 (HiSeq) or 200 bp (MiSeq). Quality control was performed using FastQC, adapters in reads were counted using the count_barcodes_by_kmer.pl scripts and top adapters and low quality reads were trimmed off using Trimmomatic (Bolger *et al.*, 2014). *De novo* genomic assemblies were performed using SPAdes 3.8 (k values of 21, 33, 55, 77, 99, 127) (Bankevich *et al.*, 2012), resulting in an median of 67 contigs per genome (32-237 contigs) with an average N50 value of 199 kb (28 kb to 817 kb). We then annotated the genomes using Prokka1.11 (Seemann, 2014) with the -usegenus *Staphylococcus* option.

3.3.4. Identification of genomic variants: SNPs, deletions and insertions

Sequencing reads of the isolates were mapped to their respective reference genomes (NCTC8325, NC_007795 (Gillaspy *et al.*, 2006) and N315, NC_002745 (Kuroda *et al.*, 2001) using BWA with default parameters (Li and Durbin, 2010). SNPs and small indels were identified using the Genome Analysis Toolkit (McKenna *et al.*, 2010) and Picardtools (https://broadinstitute.github.io/picard/). In GATK we used the indel

realignment and base recalibration options and variants were recalibrated before filtering, discarding those with RMS Mapping Quality below 40 and PHRED quality below 30. Medium-size indels (few tens or hundreds bp) were identified using Pindel (Ye *et al.*, 2009) and large deletions (> 1 kb) with the coverageBed utility of Bedtools (Quinlan and Hall, 2010), by splitting the reference genome into windows of 1 kb that were then scanned in search of those presenting at least 500 bp with zero coverage. We searched for potential acquisition of mobile genetic elements (MGEs) and long insertions by assembling unmapped reads and running BLAST searches of contigs longer than 1 kb against the NCBI database. A pangenomic matrix was also built using roary with default options (Page *et al.*, 2015) and gene content was visualized using phandango (Hadfield *et al.*, 2017).

3.3.5. Identification of genes under positive selection

To detect genes under positive detection, we first identified orthologous genes using the OrthoMCL algorithm (Li *et al.*, 2003) implemented in get_homologs (Contreras-Moreira and Vinuesa, 2013) with the following parameters: identity >85%, coverage >80% and -t 0 to report all clusters. Identification of genes under positive selection was performed using the pipeline POTION (Hongo *et al.*, 2015) and paralogous genes within groups were removed and only single-copy genes selected for the analysis. For each of these, alignments were performed using PRANK (Löytynoja and Goldman, 2005), and recombination was detected using the phi test (Bruen *et al.*, 2006). Genes that did not show significant recombination were used to build phylogenetic trees using dnaml (Felsenstein, 1989). We employed the site evolution models of Codeml (M1, M2, M7, M8) to perform codon-by-codon analysis of dN/dS ratios (nonsynonymous

to synonymous substitution, ω) of genes. Values below 1 indicate purifying selection and over 1 positive selection. Since the overall value of ω for a gene might be below 1 while some regions can still be under positive selection, we tested "site models" of positive selection, which allow ω ratio to vary among sites (Yang, 2007). Finally, a likelihood ratio test (LRT) was used to determine significant differences between nested models M1-M2 and M7-M8.

3.3.6. Functional annotation of genes

Since the reference strains NCTC8325 and N315 presented 54% and 25% of their CDSs annotated as "hypothetical proteins", we re-annotated the genes identified in our analysis using InterProScan (Jones *et al.*, 2014), and BLAST searches against the Conserved Domain Database (Marchler-Bauer *et al.*, 2003) or Pfam protein families database (Finn *et al.*, 2016). In addition, we assigned clusters of orthologous groups to mutated genes and genes under positive selection using the eggnog mapper (Huerta-Cepas *et al.*, 2017).

3.3.7. Phylogenetic and population genetic analysis

To infer the isolate genealogies and branch lengths we manually constructed core genome SNPs alignments from the variants characterized and used them to build minimum evolution phylogenetic trees using the maximum composite likelihood method in MEGA (Tamura *et al.*, 2007). From the variants identified we estimated the substitution rates as described in equation 1: number of mutations (m) divided by the genome size (N) times the generations (t/g). Considering a replication time of 30 min for *S. aureus*, 1 year is approximately equivalent to 18,000 generations.

$$\mu = \frac{m}{N * \left(\frac{t}{g}\right)} \quad (1)$$

To understand within-host dynamics we calculated the total number of SNPs per isolate, as well as fixed and variable SNPs present in each of the three isolates from individual sheep. Given the very-short evolutionary timescales, we applied a simple linear regression on SNP counts versus number of days to estimate differences between the genetic distances for transmissions and within single host population dynamics. Pairwise genetic distances were calculated as the number of SNPs between two isolates from the same host.

3.3.8. Bacterial genome evolution simulations

We simulated the evolution of genome populations using Genomepop2, a forward time simulation program (Carvajal-Rodriguez, 2008). A limitation of this algorithms is the exponential increase of computational intensiveness with genomic length. Thus, we only simulated 1% of the typical *S. aureus* genome, i.e. 28 kb, specifying a mutation rate of 0.0001 SNPs per genome per generation (around 3.5 x 10-9 mutations per base per generation) and with recombination set to 0. Assuming *S. aureus* has a generation time of 30 m and that experiments were performed for 360 d, we simulated 17281 generations, with a maximum population size of 1.25 x 10^6 individuals, which is equivalent to 5000 cfu/ml in a maximum volume of 250 ml. In order to examine the

effect of different types of bottlenecks, we simulated a constant population size, tight bottlenecks produced by transmissions, wide bottlenecks produced by suckling lambs (feeding bottlenecks) and a combination of both. In addition, we simulated evolving populations under two selection models, lack of selection (all neutral mutations) and a hypothetical distribution with selection coefficients as plotted in Figure 3.1. This distribution was based on previous research (Bank *et al.*, 2014; Eyre-Walker and Keightley, 2007) and since we were simulating adaptation to a new environment, most mutations were set as lethal, deleterious (gamma distribution) and neutral, and very few mutations slightly beneficial. In total, we simulated 8 scenarios and ran 100 replicates for each one. We sampled 1000 genomes every 1000 generations and estimated a number of population parameters, including the genetic diversity (as mean pairwise number of SNPs between samples), fixed number of SNPs, variable number of SNPs, and frequency and variation of genotypes.

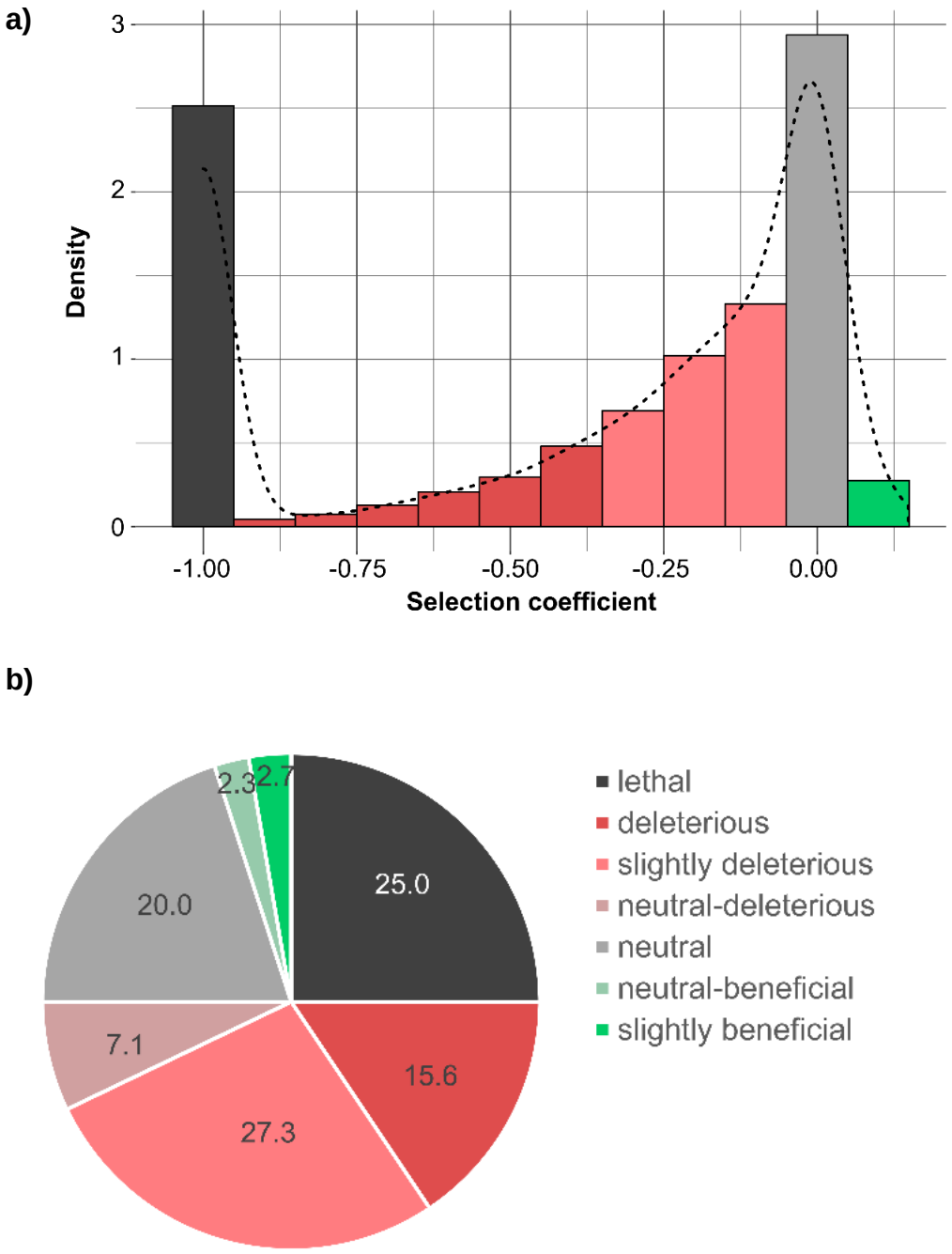


Figure 3.1. Distribution of selection coefficients in simulations. a) Distribution of selection coefficients. Deleterious mutations follow a gamma distribution. b) Pie chart showing the proportions of different selection coefficients.

3.4. Results

3.4.1. Simulating *S. aureus* host jumps and successive transmissions in the new host

To examine the evolution of *S. aureus* in a new host-species, we simulated a series of host jumps from humans into sheep. For this, we injected human-associated *S. aureus* strains (NCTC8325 and N315) into the teat ducts of ewes and the animals were released in a farm environment, where they were in direct contact with other sheep and uninfected lambs, which often suckled from their mothers. Successful host jumps are different from spill-over events in that bacteria do not only infect and replicate in the novel host-species but are also able to transmit to new host individuals (Woolhouse and Gowtage-Sequeria, 2005). Thus, we also simulated transmission passages from sheep-to-sheep by inoculating additional animals with bacteria isolated from previous ewes. This process was repeated up to 6 or 7 times, and some isolates were used to infect two or more sheep, leading to tree-form transmission chains with clearly defined lineages and sub-lineages (Figure 3.2). The maximum infection time was 400 d, which considering *S. aureus* replicates every 25-30 min, is equivalent to around 18,000-24,000 generations (Figure 3.2).

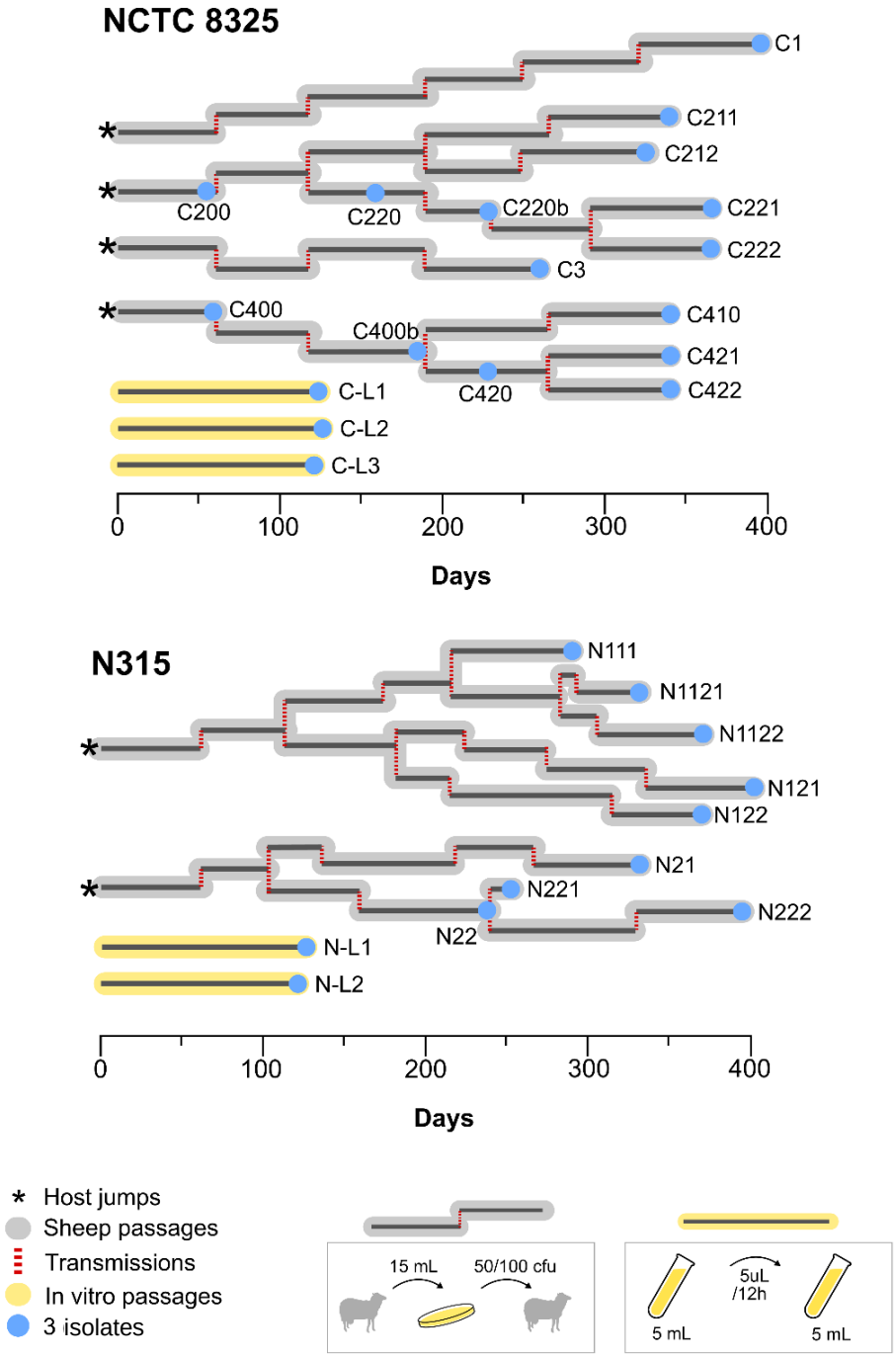


Figure 3.2. Experimental design of sheep infections. From the parental strains, host jumps were reconstructed (*) and serial passages were performed every 3 to 5 w (red lines). Each thick line represents a lineage. We selected three clones from some intermediate and last isolation plates of every lineage for genomic DNA sequencing (blue circles).

3.4.2. Passages result in increased *S. aureus* fitness in the new host

We hypothesized that strains that had been replicating in sheep for a longer period of time might have acquired adaptations to the new host that would enhance infective capacity. In order to examine this, we calculated the number of sheep infected versus the sheep inoculated for each passage and strain. The infection rates remained relatively constant for all N315 strains, with 47% of the inoculated sheep developing infections in passages 6 and 7, compared to the 40% in passages 1 and 2. For NCTC8325, infectivity rates slightly increased from 40% in passages 1 and 2 to 63% in the last passages (Figure 3.3). However, this was not statistically significant (p-value 0.22, Mann-Whitney test), suggesting passaging had limited effect on infectivity. However, the data were affected by extensive variation between animals, reflecting differences in physiological and anatomical host factors (Prasad and Newbould, 1968); and extremely variable infection rates in the final passages (ranging from 0 to 1), due to the small number of sheep inoculated with a given strain compared to the first passages (8 and 5 ewes were inoculated with the NCTC8325 and N315 strains respectively) (Supplementary Table 5).

Another approach to measure adaptation to a new environment is competition experiments with ancestral strains (Visser and Lenski, 2002). To determine if passaged strains were fitter than their respective parental strains, we performed coinfections of additional sheep with similar loads of both original and passaged strains (two derived from the N315 and two from the NCTC8325).

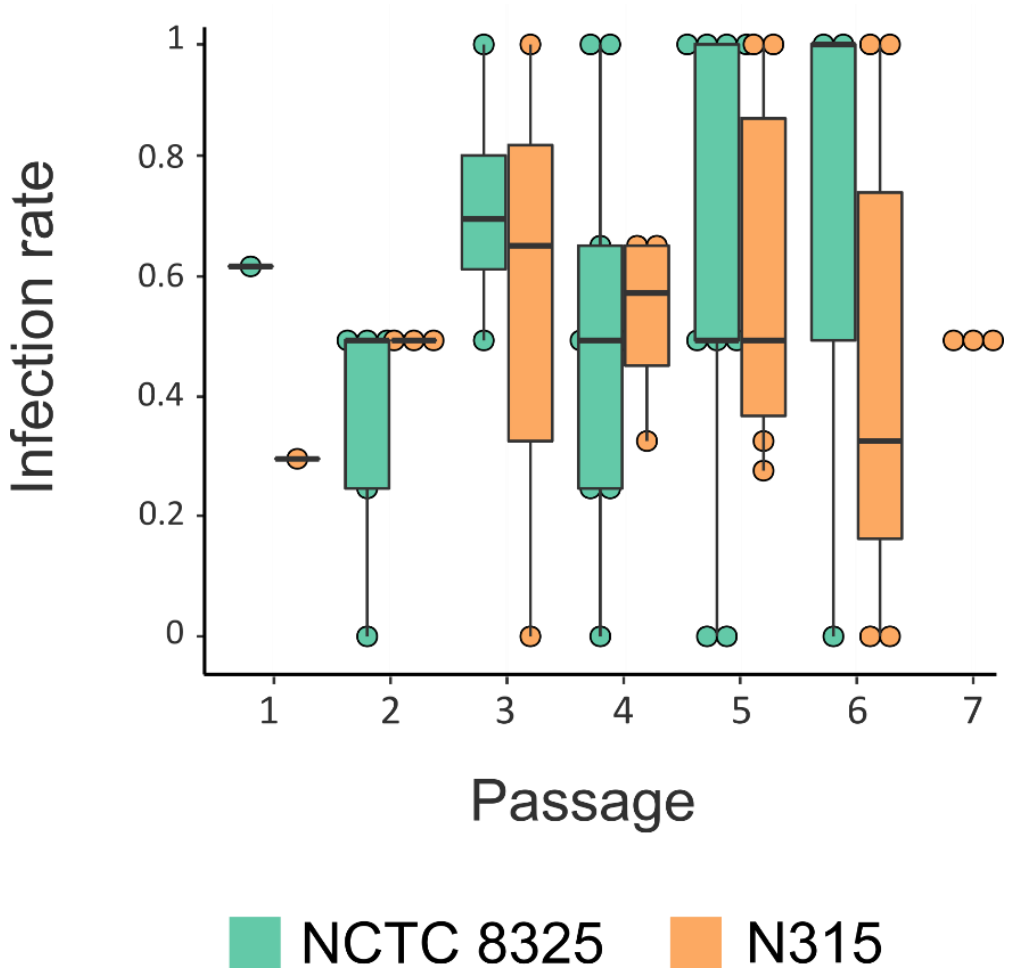


Figure 3.3. Boxplots of infection rates estimated for every passage. The relative amount of infected sheep out of the inoculated sheep does not significantly vary in accordance with number of passages. Boxplots show means and interquartile ranges, while points indicate the infection rates for every passage using the NCTC 8325 strain (green) and the N315 strain (orange).

40 d after the inoculations, sheep were milked and bacteria isolated from those that had developed intramammary gland infections. In all cases, we only recovered isolates representing one of the co-infecting strains, suggesting that the initial equilibrium between the two populations was unstable and ended up shifting towards one of them. Notably, passaged strains were recovered more frequently (Fisher exact test, p= 0.0394, p=0.027 Barnard's test), suggesting they were more likely to survive than the original ones, consistent with increased fitness associated with a competitive advantage (Figure 3.4a). We observed differences between the four passaged strains, which may suggest variable rates in adaptive evolution. For example, out of the 8 ewes co-infected with the N315 strain and its descendant N1122, 4 (50%) had the original strain and 4 (50%) the passaged one. In contrast, sheep co-infected with NCTC8325 and its successor C221 presented the passaged line in 90% of cases.

The survival of only a single strain in all coinfections suggests within-host dynamics is influenced by strong oscillations of bacterial populations, consistent with the nature of the udder environment. The volume of milk in the glands is considerably reduced every time a lamb suckles, contributing to narrow population bottlenecks that can be followed by clonal expansions of subpopulations, similarly to previous reports (McVicker *et al.*, 2014). To examine this phenomenon further, we infected 20 additional sheep with 2 isogenic strains, the NCTC8325 (*wt*) and its clone containing a single synonymous mutation in the vWbp pseudogene (*ss*), presumably identical in fitness. Excluding five ewes that were not infected or cleared the infection after 1 w, all others presented only a single strain (Figure 3.4b).

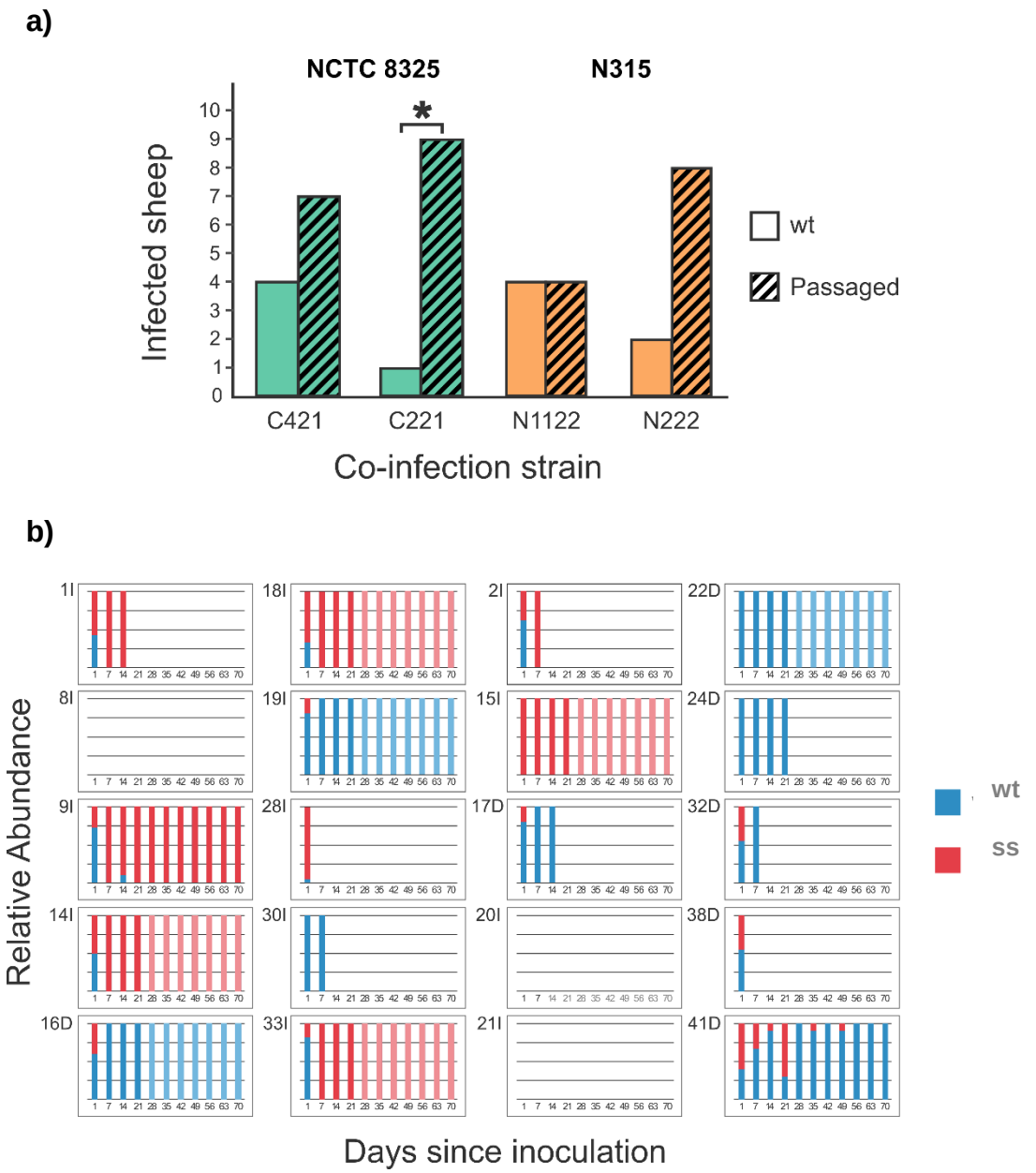


Figure 3.4. Coinfection experiment results. a) Number of infected sheep with the *wt* strain or the passaged strain (striped) 40 d post co-infection. Significant differences were obtained when compared to the 50-50% outcome expected for the null hypothesis with no adaptation. (p=0.039, one-tailed fisher exact test; p=0.027 Barnard's test). b) Proportions of the *wt* strain (blue) and the strain with a synonymous SNP (*ss*) for co-infection experiments of 20 sheep.

As expected, the *wt* and *ss* strains were recovered in similar proportions (8 wt, 7 *ss*). Of note, both strains coexisted in most sheep at day one but later during the first week, the equilibrium was lost and the udder was only colonized by a single strain. In one case (sheep 41D) both strains coexisted for more than 50 d, until a single strain became dominant.

3.4.3. Adaptive mutations acquired during the infections and passages

From the final isolation plates of every lineage, we performed whole-genome sequencing of three colonies and identified the genomic variation *S. aureus* had accumulated during the passages. Passaged strains presented a range of mutation types, including SNPs, short indels and large deletions. However, large insertions, acquisition of MGEs or genomic rearrangements were not observed in any isolate. The number and variety of mutations for independent lineages were variable but evenly distributed across the genome (Figure 3.5a). Of the 99 SNPs identified in total among the 51 isolates (27 derived from NCTC8325 and 24 from N315), 25 were in intergenic regions, 52 were non-synonymous, 18 synonymous and 4 caused the disruption of genes by introducing premature stop codons. Of the short insertions and deletions (indels) 9 resulted in frame shifts, 4 in disruptive frame changes and 14 were intergenic (Figure 3.5b). Most mutations had moderate effects (change in protein structure) and modifier effects (affecting intergenic and regulatory regions) (Figure 3.5c), but we were interested in those that affect protein structure, which are more likely to be involved in host adaptation. The majority of these mutations were in genes encoding proteins involved in virulence, regulation, signal transduction, or transport and metabolism (Figure 3.5a). Of these, 47 % affected genes encoding secreted products

or proteins related with the bacterial surface, most of which participate in cell wall biosynthesis, pathogenesis and host-pathogen interactions. Some notable examples include the capsule gene *cap5B*, involved in the biosynthesis of capsular polysaccharide (O'Riordan and Lee, 2004); the EssB transporter, required for secretion of the proteins EsxA and EsxB involved in virulence (Chen *et al.*, 2012); the staphylococcal complement inhibitor SCIN, that counteracts the host immune defence (Rooijakkers *et al.*, 2005); and hyaluronate lyase, involved in tissue invasion (Makris *et al.*, 2004). We also identified missense SNPs in genes encoding regulatory proteins, such as the *walK* gene of the two-component regulatory system WalK/WalR, which controls autolysis, biofilm formation and cell wall metabolism (Dubrac *et al.*, 2007), mutations in the virulence regulators *agrC* and *agrA*, or in the *ccpA* gene, which encodes the catabolite control protein A, required for carbon catabolite repression and virulence regulation (Seidl *et al.*, 2006). These mutations in regulators of global processes are of interest, as they may enable bacteria to rapidly adapt to environmental changes. Additionally, we detected a frameshift in the *vraS* locus, suggesting that adaptation to a new host-species might also be mediated by changes in two-component signal transduction systems that enable bacteria to sense, and respond to stimuli and induce changes in transcription. Proteins involved in replication were affected also, with mutations in the *recU* and *rpoB* genes, which respectively encode the Holliday junction resolvase and the β-subunit of RNA polymerase. Lastly, several mutations affected transporters (multidrug efflux transporter Acr, putative drug exporter SA2339; ABC transporter SAOUHSC_00634 or tricarboxylate transporter SA0664) and metabolic enzymes (adenine-specific DNA methylase, glyceraldehyde-3-phosphate dehydrogenase, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase).

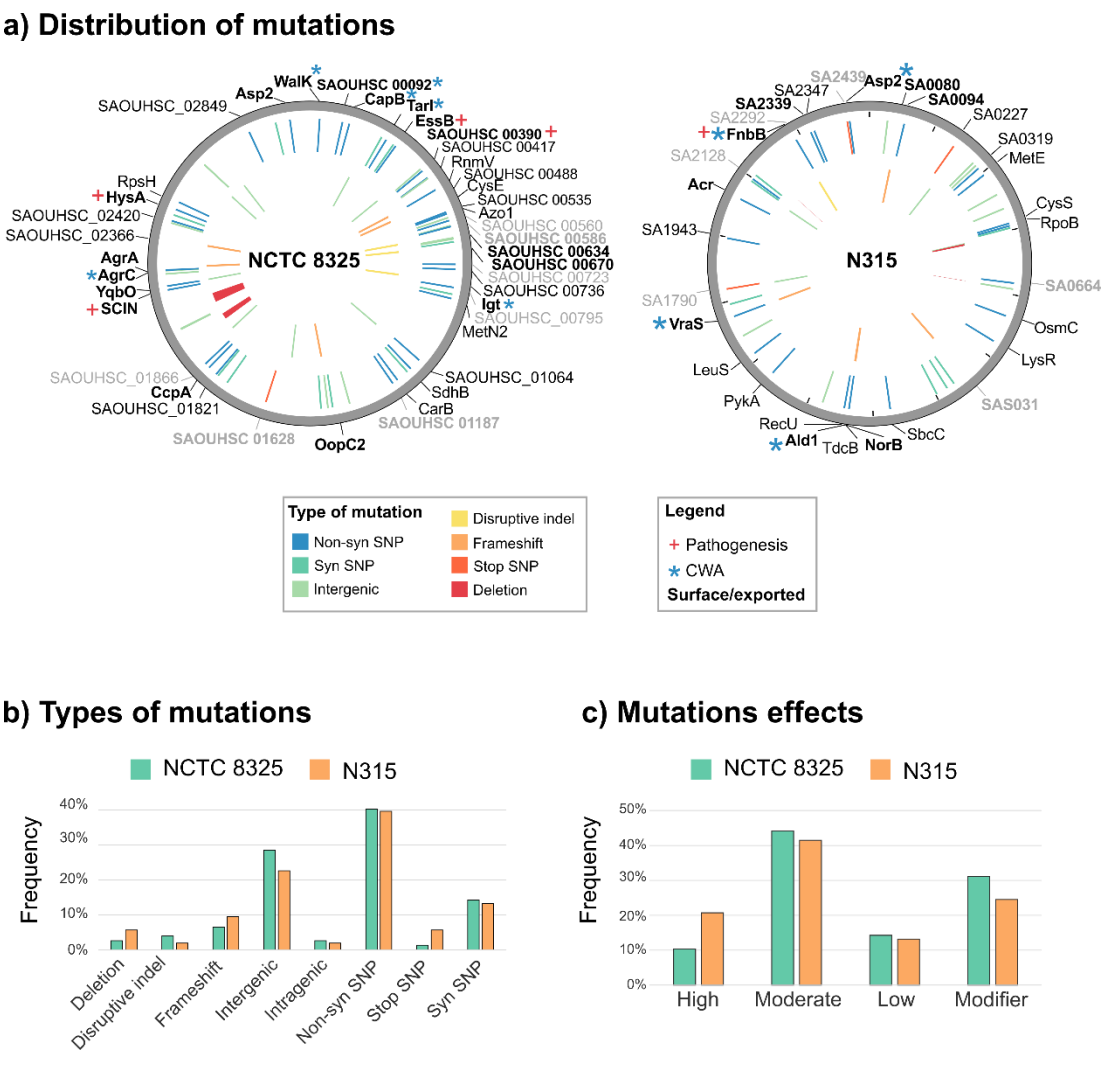


Figure 3.5. Mutations acquired during the infections and passages. a) Distribution of different types of mutations across the genomes and genes affected. b) Frequency of various types of mutations. c) Mutations effects on the protein level: high (protein disrupted or shortened), moderate (low impact variations in protein sequence such as amino acid changes), low (mutations do not affect the protein) or modifier (in intergenic regions).

Consistent with previous studies that identified gene loss and pseudogene accumulation during *S. aureus* host-adaptation (Guinane *et al.*, 2010; Lowder *et al.*, 2009); we also detected large deletions in isolates derived from both the NCTC8325 and N315 strains, which corresponded with well-characterized elements. The typical transducing prophage Φ11 was excised in an early infection of the C2 lineage, since it was absent in all the derived isolates. In addition, a single isolate from the clone C221 lacked a 28 kb region of the variable genome island νSaβ that contains all the 6 *spl* genes and a cluster of 4 genes encoding a flavoprotein and the lantibiotics epidermin and gallidermin. In addition, all the isolates from the sublineage C4 had a 60bp deletion in a STAR-like false-CRISPR element (Zhang and Ye, 2017), which has been proposed to play a functional role in cell physiology and pathogenesis (Purves *et al.*, 2012). Similarly, the N121 clonal isolates had a 60 bp deletion of a repeat situated in an intergenic region, suggesting these elements linked with regulatory functions may play a role during adaptation. Finally, the clone N21 lacked the *sdr* gene cluster, which encodes the serine-aspartate repeat-containing proteins that participate in adherence to human epithelial cells (Corrigan *et al.*, 2009), interaction with the innate immune system cells (Sitkiewicz *et al.*, 2011) and the infection process (Tung *et al.*, 2000).

Although the experiment was designed to allow the potential acquisition of MGEs from resident sheep bacteria, possibly mediated by lambs suckling multiple ewes, none was identified. Considering that strains from different host-species exchange bacteriophages and plasmids when put together in the same niche (McCarthy *et al.*, 2014), we conducted BLAST searches of mobile genetic elements from ruminant-

adapted *S. aureus* strains (Guinane *et al.*, 2010) in all the isolates passaged, but such elements were not detected, suggesting robust barriers to genetic acquisition.

Convergent evolution has been reported during rapid within-host adaptation in other bacterial species (Lieberman *et al.*, 2011; Marvig *et al.*, 2013). We searched for host-adaptive signatures or genes that acquired related mutations in independent lineages and found the *asp2* gene to be mutated in two passaged isolates from the N315 and NCTC8325 strains. *asp2* encodes the protein Asp2, which is part of the accessory secretory system involved in the export of serine-rich glycoproteins to the bacterial surface, that are involved in pathogenesis in a number of Gram-positive species (Siboo *et al.*, 2008).

Next, by combining the results from the coinfection experiments with the genomic changes identified, we aimed to estimate the contributions of different sets of mutations to the fitness enhancement during passage. Although the number and types of mutations for the four passaged strains were very different, the two isolates with the highest competition rates (C221 and N222), which outcompeted the wt in 90% and 80% of cases, presented the highest proportions of non-synonymous SNPs in their genomes, 9/10 and 4/4 respectively. C221 missense SNPs were located in genes encoding a polysaccharide biosynthesis protein, a glyceraldehyde-3-phosphate dehydrogenase or a hyaluronate lyase, among others. The 4 non-synonymous SNPs of N222 affected a pyruvate kinase, a multidrug efflux transport, an aminotransferase and the Asp2 protein. However, in the other strains that also presented a fitness increment compared to the wt, 80% of SNPs were intergenic and synonymous, and only 1 non-

synonymous SNP was identified, impacting the gene encoding the type VII secretion protein EssB. In addition, it contained an indel disrupting an ABC transporter.

For the *in vitro* passaged lineages, isolates had accumulated a greater numbers of mutations, but they were highly variable for each lineage, with the most divergent presenting 23 SNPs and the most conserved having only 2 SNPs and 4 indels. SNPs were primarily non-synonymous affecting genes encoding enzymes involved in metabolic pathways. In addition, 4 SNPs in the *lpl3*, *narG*, *oatA* and *alr1* genes resulted in premature stop codons of a lipoprotein, a nitrate reductase, an O-acetyltransferase and an alanine racemase. We also found large deletions of phages and a 17 kb cluster of genes involved in the biosynthesis of staphyloxanthin. Taken together, these results suggest that during the *in vitro* passages, due to the availability of nutrients in the media, non-essential metabolic pathways tend to undergo genetic diversification leading to a loss of function (Leiby *et al.*, 2014).

3.4.4. Short-term and long-term host adaptation follow different evolutionary pathways.

Some of the genes mutated during the passages, such as the serine aspartate-rich containing proteins C and E or the Fibronectin-binding protein A, had been previously characterized to be undergoing adaptive evolution in ruminants (Guinane *et al.*, 2010). Thus, we investigated if during the *in vivo* passages, genomes from the human-associated strains acquired similar features to *S. aureus* strains adapted to sheep and goats. For this purpose, we performed BLAST searches of every mutated gene against 14 ruminant associated genomes from the CC133 clade (Figure 3.8) and then aligned

homologous sequences and compared them in search of similarities with the mutated genes. We did not identify any genetic variants in human genes that were converting to resemble those from ruminant strains.

During the first few thousand generations after a host-jump, bacteria require a rapid adaptation that may involve different genes than those affected by the long-term adaptive evolution, a slower process of refinement to the host (Rohmer *et al.*, 2011). We examined protein-coding genes under positive selection in *S. aureus* strains that specifically infect sheep and goats (CC133) and compared them to the genes mutated during passages. We also checked for genes under positive selection in the human-associated clades of the strains used for the experiments. Several of the genes under selection in ruminant and human associated strains also encoded cell wall-associated (CWA) proteins, which are involved in host-pathogen interactions and therefore more likely to be under selective pressure to adapt to polymorphic receptors in different host species (Guinane *et al.*, 2010). Comparisons of functional categories showed that genes mutated during the passages were more enriched in transport and metabolism while genes under positive selection presented higher proportions of COGs involved in information storage and processing (Figure 3.6).

We also tried to use the non-synonymous/synonymous ratio for identifying which mutated genes were under selection, but this measure is not adequate for samples drawn from a single population (Kryazhimskiy and Plotkin, 2008).

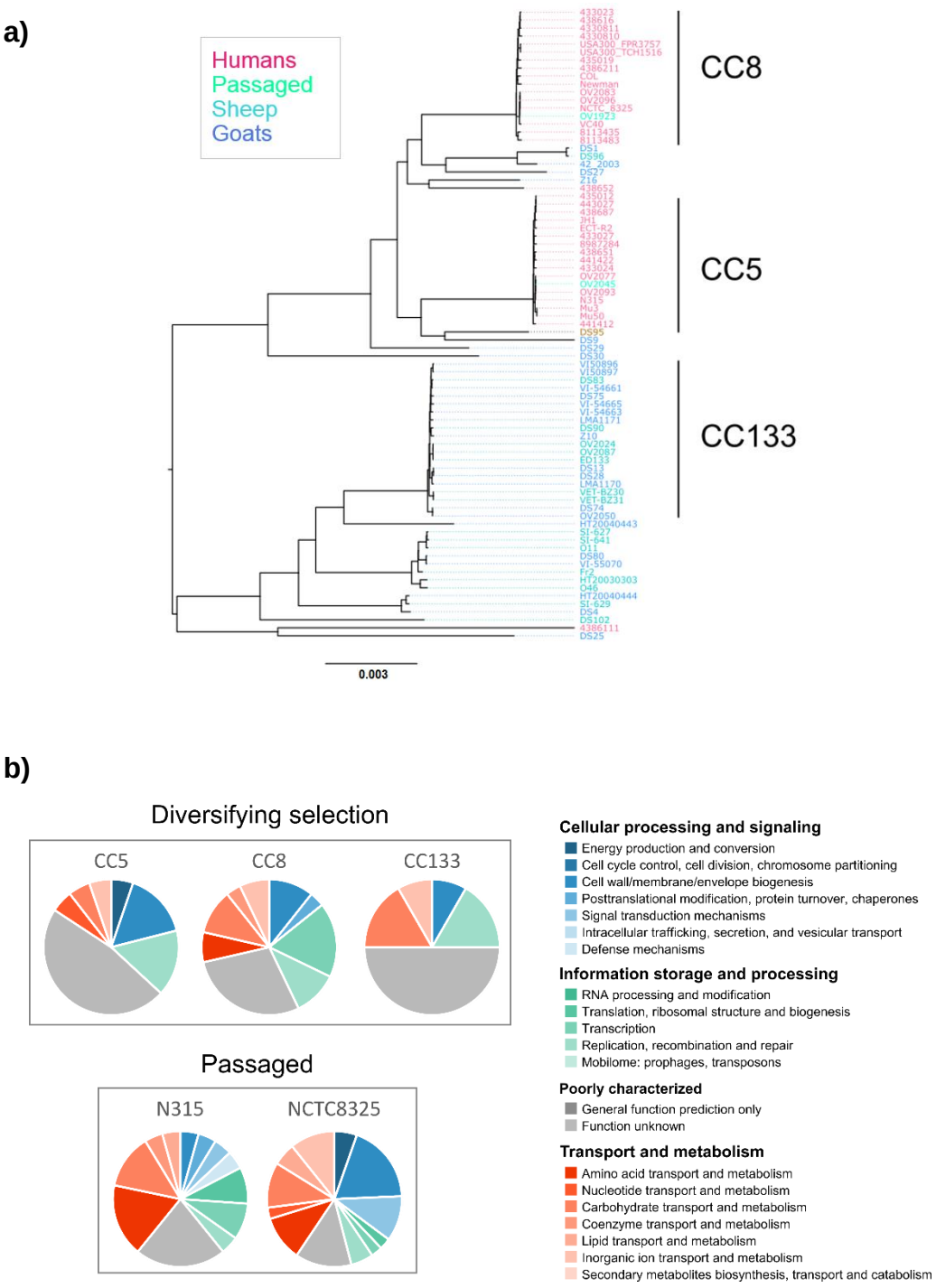


Figure 3.6. Differences between short and long-term adaptation to ruminants. a) Maximum likelihood phylogenetic tree of *S. aureus* strains from humans and ruminants. b) Clusters of orthologous groups (COGs) for genes under positive selection in the human-associated CC5 and CC8 clades, the ruminant associated CC133, and for genes mutated in passages.

3.4.5. Population genetics and dynamics of within-host evolution and transmissions

We constructed minimum evolution trees for *S. aureus* NCTC8325 and N315 based on the core SNPs of all the isolates derived from each passage experiment (Figure 3.7). Tree topologies were consistent with the transmission chains simulated during the infection experiments (Figure 3.2). To infer the population dynamics from the trees we need to account for both within-host evolution and transmissions events. Depending on the rate at which new mutations arise and the fitness benefit effect of those mutations, within-host evolution may reflect dominance of a single strain at any time-point due to periodic selection, or coexistence of multiple genotypes within the host due to clonal interference (Barrick and Lenski, 2013). Simulations of transmission events between sheep introduced tight bottlenecks representing 50-100 colonies picked from the primary isolation plate of one animal that were used to infect the next animal. Considering the effect of transmission events, we postulate three possible scenarios: (i) dominance of a single strain before and after transmission; (ii) diverse genotypes coexist within individuals but only a single clone dominates after transmission; (iii) diverse genotypes coexist within a host and after transmission. Examination of the topology of the phylogenetic trees (Figure 3.7) is consistent with scenario (ii), where different branch lengths representing multiple isolates from individual sheep indicate the accumulation of genetic diversity within those animals, followed by dominance of a single lineage after a passage to a new host (Figure 3.7). This is further supported by sequencing of intermediate isolates, which represent distinct sub-branches of the tree (Figure 3.7 in red). Differences in branch lengths of ML trees indicate that distinct genotypes accumulate mutations at different rates.

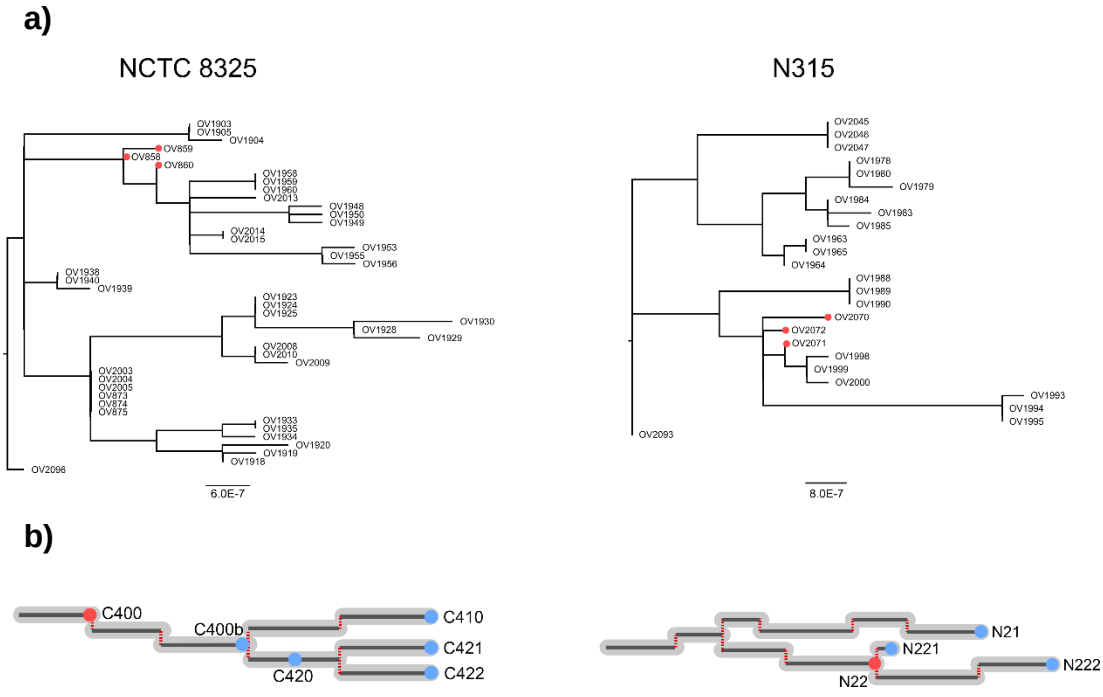


Figure 3.7. Minimum evolution trees of the passages. a) Tree topologies and lineages are consistent with the transmission chains simulated in the experiments. Differences in branch lengths for isolates from specific sheep indicate the co-existence of different alleles within the animals, supporting the scenario of 'genetic variability arising from single clones after a transmission'. b) Lineages from which intermediate isolates were sequenced are marked in red.

In order to understand the molecular processes contributing to host-adaption after a host-switch, we examined the evolutionary dynamics of *S. aureus* during experimental infection of sheep. Bacteria infecting sheep exhibited an average substitution rate of 2.78 mutations per Mb per year (Figure 3.8a), similar to previous estimations for *S. aureus* infecting humans (Young *et al.*, 2012). The estimated rates for independent lineages were normally distributed and there was no evidence for hypermutators, which have been previously associated with rapid host adaptation (Lieberman *et al.*, 2014). In contrast, bacteria cultured in nutrient-rich conditions *in vitro* had an average substitution rate of 12.3 SNPs/Mb/year, a rate 4.5 times higher than the *in vivo* rate. We can attribute this difference to the higher replication rate of *S. aureus* in nutrient broth compared to intra-mammary infections (see methods), where suboptimal environmental conditions and the host immune response deaccelerate bacterial growth. Furthermore, for the *in vitro* passages, bacterial populations reached very high numbers (~1e9 cfu/mL) and serial transfers were performed every 12 h, with much broader bottlenecks, leading to higher numbers of fixed SNPs.

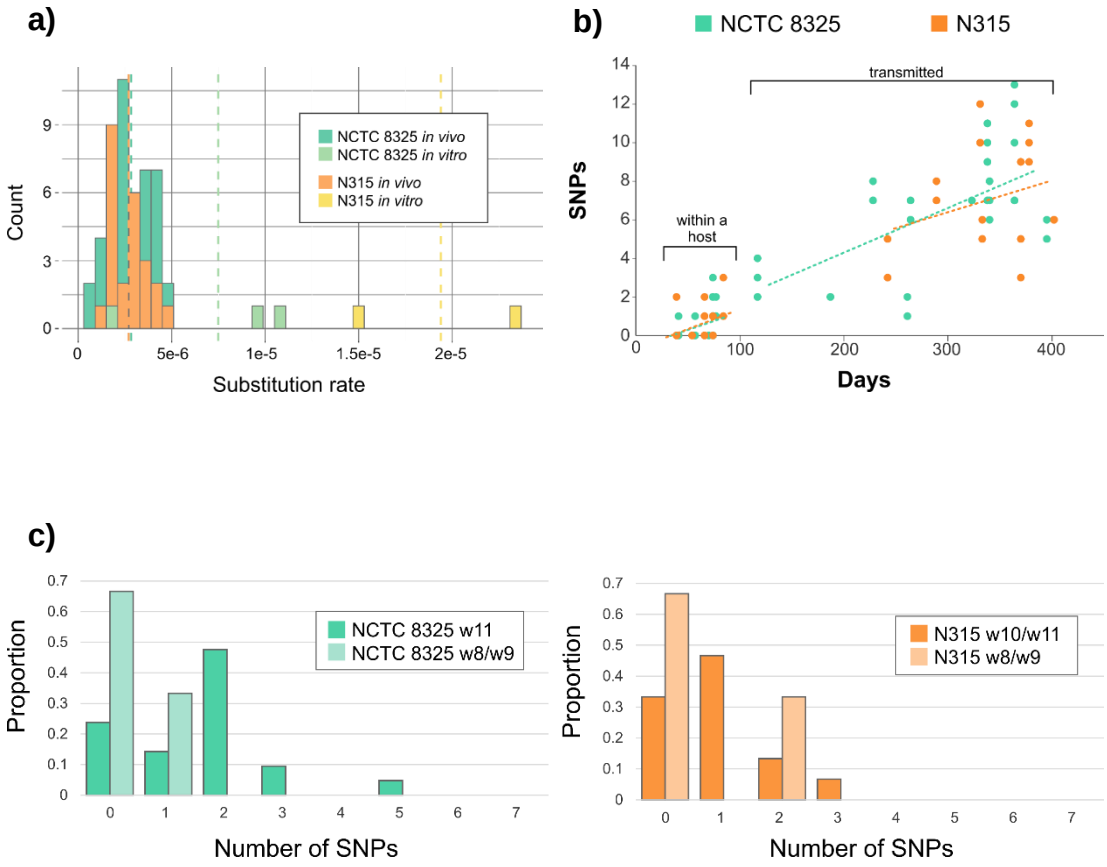


Figure 3.8. Evolutionary dynamics summary analysis. a) Substitution rates for *in vivo* and *in vitro* passages. b) SNP accumulation over time during transmissions (NCTC 8325: R^2=0.44, correlation=0.66; N315: R^2=0.10, correlation=0.31) and/or within host only (NCTC 8325: R^2=0.06, correlation=0.24; N315: R^2=0.10, correlation=0.32). c) Pairwise genetic distance distribution for weeks 8 and 9 (early) and weeks 10 and 11 (late).

In order to further explore within-host dynamics and to infer the impact of transmissions on genetic diversity, we plotted the number of SNPs accumulated in all isolates sequenced during the study versus the total infection days (Figure 3.8b). The graph shows a strong temporal correlation and the molecular clock lines crosses the x-axis at around 0, consistent with the use of single genotypes for the inoculation of the initial hosts. In contrast, when we only plotted SNPs identified between isolates from individual animals, trend lines crossed the time axis at around day 30, consistent with a delay in the appearance of genetic diversity. Since the rate at which new mutations arise in the populations is constant (Figure 3.8), the absence of the expected diversity in the time after the initial inoculation/transmission simulation can be explained by either strong selection of ancestral genotypes or by genetic drift. As we have demonstrated that passaged isolates are generally fitter than their progenitors, we suggest that the reduction in genetic diversity is due to the continuous bottlenecks resulting from feeding lambs. This was further supported by the distributions of pairwise genetic distances between isolates sampled from within individual hosts at early (8-9 w) or late (10-11 w) time points (Figure 3.8c). At early time points, infections from clonal populations are expected to follow a geometric distribution, which then turn into a geometric-Poisson approximation as time passes on (Worby *et al.*, 2014a; Worby *et al.*, 2014b). These patterns were observed in our data consistent with the existence of continuous bottlenecks purging the accumulated population diversity.

To provide a high resolution picture of the within-host population dynamics, we sequenced 100 *S. aureus* isolates representing colonies isolated from sheep 40 d after

co-infection with isogenic *wt* and derivative *ss* that differed by a single synonymous mutation in the *vwbp* pseudogene. Among the 100 isolates examined, 6 SNPs were identified, of which only 1 non-synonymous SNP in the locus *aacA*, encoding the acetyl-CoA carboxylase alpha subunit, was fixed in the population. We performed PCRs of this genomic region in 10 colonies isolated every week after the infection, revealing this mutation achieved fixation during the first week post-inoculation. The other five SNPs were present in very low frequency, representing 1-3% of the population. Sequencing of the mutation-containing region in 3% of the population from 400 colonies at times 14, 22 and 29 did not detect the presence of the mutation, indicating it only appeared during the last 12 d of the infection. These data are consistent with our previous analysis indicating that bottlenecks limit the fixation of mutations by purging of accumulated variation due to genetic drift.

3.4.6. Beneficial mutations emerge in the face of regular bottlenecks

In order to further examine the impact of transmission and feeding-associated bottlenecks on the genomic diversity and the nature of mutations selected, we carried out simulations of models of the evolution of bacterial genomes using Genomepop2 (Carvajal-Rodriguez, 2008). To account for differences between the different bottlenecks, we reproduced four scenarios: a constant population size (1), transmissions between individuals (2), feeding bottlenecks (3) and a combination of both (4). Given that the software allowed us to specify the selection coefficients for every mutation in the genome, we simulated two models, one with all mutations as neutral and another with selection coefficients following a hypothetical previously determined distribution (Barrick and Lenski, 2013; Eyre-Walker and Keightley, 2007)

(see methods). Simulations were run for over 17,000 generations, similar to the estimated replications for the *in vivo* experiments, and we sampled 100 isolates every 1000 generations. Pairwise genetic distances increased over time for the neutral model and remained constant when selection occurred (Figure 3.9a), indicating that in the absence of selection, mutations accumulate steadily in the population. In both situations, bottlenecks resulted in a reduction of the population diversity. Next, we looked at the accumulation of variable and fixed SNPs over time (Figure 3.9b). In lack of selection, variable SNPs increase logarithmically towards an equilibrium, consistent with the pairwise genetic distances observed. As expected, bottlenecks considerably reduced the variable SNPs, but still allowed certain genetic variation to remain in the population without ever reaching fixation. However, once we introduced some level of selection as would be expected during infection, some mutations swept through the population and became fixed, causing a drastic reduction in the number of variable SNPs. Nevertheless, the number of SNPs that became fixed in the simulations was much higher than in our experiments, possibly due to differences between the model and the experimental infections in relation to the size of the bottlenecks, mutation rates, the generation time within sheep or the selection coefficients distribution.

Finally, we determined the types of coefficients associated with the variable and fixed SNPs in the four scenarios (Figure 3.9c). Within a host, variable SNPs represent the diversity on which selection acts to fix them into the population. As one would expect, in lack of bottlenecks, beneficial mutations outcompete neutral and deleterious SNPs, and tend to accumulate during the generations.

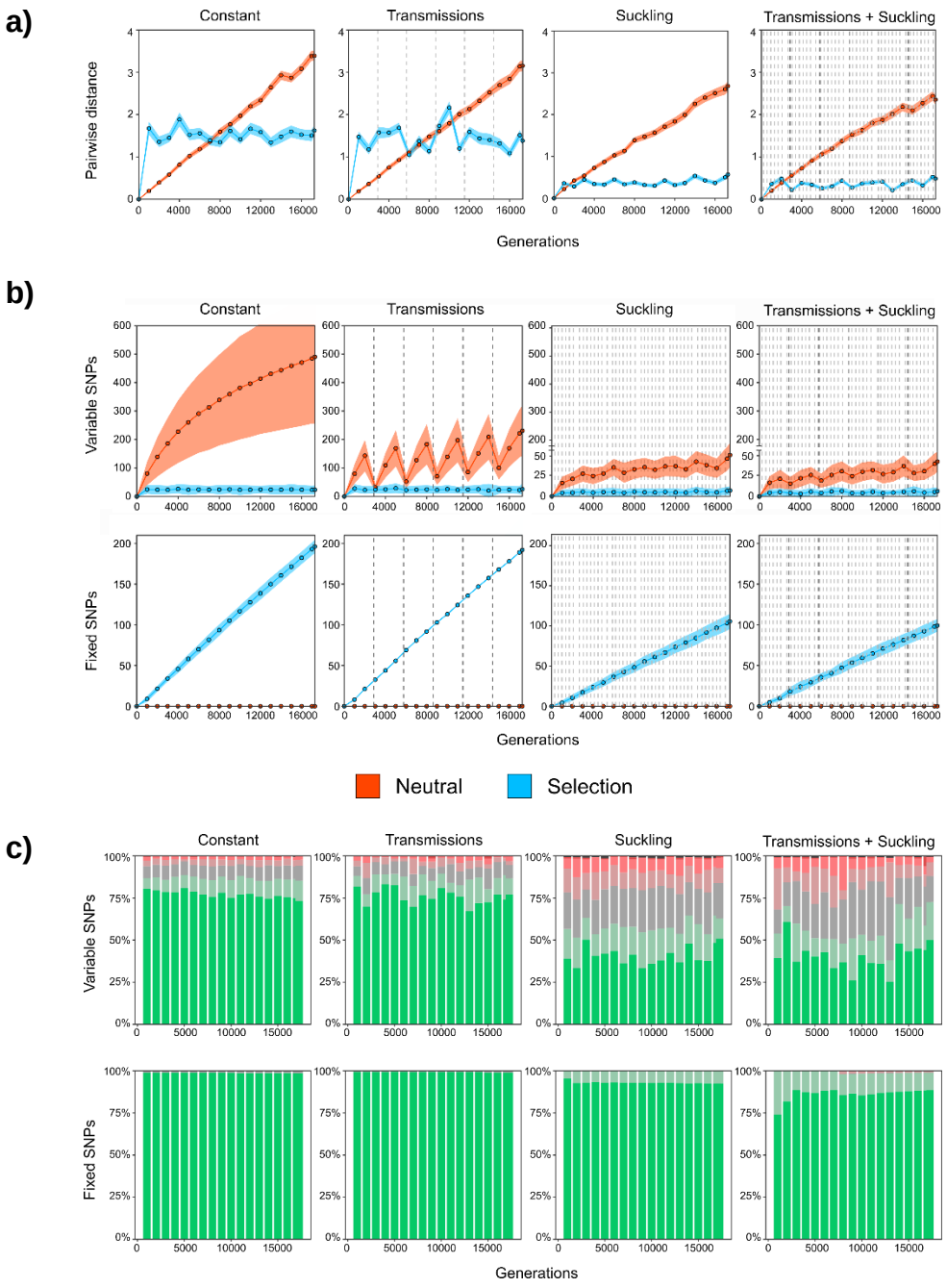


Figure 3.9. Simulations of genomic populations. a) Average pairwise genetic distances between random selected isolates from the populations simulated. b) Accumulation of fixed and variable SNPs over time. c) Types of variable and fixed SNPs determined from the selection coefficients associated with every nucleotide.

Although deleterious and lethal mutants are rapidly purged, transmission and feeding bottlenecks increase the power of genetic drift, leading to an accumulation of neutral and deleterious SNPs, which, after natural selection, result into an accumulation of some neutral mutations. This is consistent with our co-infection results, as one of the four lineages resulted in similar fitness to the ancestral strain, suggesting that mutations acquired were at the best, neutral.

3.5. Discussion

In this study, we investigated the molecular basis of adaptation to a new host-species and revealed new insights into within-host evolution and transmission dynamics associated with the adaptive process in a natural environment. Although evidence for *S. aureus* host adaptation has been reported (Guinane *et al.*, 2010; Lowder *et al.*, 2009; Resch *et al.*, 2013; Spoor *et al.*, 2013), these studies generally used comparative genomics and phylogenetic analysis of contemporary clinical isolates to identify genetic signatures associated with transition to a new hosts-species. However, our understanding of the population dynamics associated with a shift in host-species is lacking. Here we examined, for the first time, the evolutionary genetic changes responsible for *S. aureus* adaptation during experimental infection of a new host-species.

Adaptive evolution to a new host-species is a slow process, starting when pathogens interact with novel host-types leading to transient spill-over events (Engering *et al.*, 2013). If pathogens are able to transmit to further individuals of the new host-species, these events can be considered successful host jumps (Woolhouse *et al*, 2005). This

has occurred historically through domestication and most recently because of industrial livestock production, providing increased interactions and new opportunities for host-switching (Wolfe *et al.*, 2007). Using an *in vivo* long-term experimental evolution approach, we reproduced *S. aureus* host-jumps and transmissions in the new host, which simulated the natural infection scenario while giving us control of the direction of transmissions.

Once pathogens are established in the new host-species, they may develop novel traits to exploit distinct nutrients available, which may result in increased pathology (Engering *et al.*, 2013). In the current study, sheep injected with human-associated strains only developed subclinical infections and infection rates did not increase for strains that had been associated for longer with the new host, suggesting passaged strains did not acquire such traits in the examined timeframe. However, using a competitions analysis approach (de Visser and Lenski, 2002), we demonstrated that passaged strains were more likely to persist than the evolutionary progenitors, indicating increased fitness was acquired during the passage experiments.

Examination of the nature of the genes that acquired mutations revealed a large number of encoding proteins involved in *S. aureus* pathogenesis and host-interactions. Although these categories may contribute to the adaptation to the new environment, they are also likely to be under immune selection (Grumann *et al.*, 2014). The identification of polymorphisms in regulators of transcription may have a profound effect on host-pathogen interactions and just a few genetic changes can result in the

remodelling of global regulatory networks, determining the ecology, physiology, and fitness of bacteria (Damkiær *et al.*, 2013; Howden *et al.*, 2011).

The large deletions we observed for some lineages is consistent with gene loss as a mechanism for host-adaptation (Guinane *et al.*, 2010; Lowder *et al.*, 2009). Although this kind of genome structural variation is less common than small SNPs and indels, loss of function and genome reduction have been demonstrated to be important strategies for adaptation to new hosts (Hottes *et al.*, 2013; Toft and Andersson, 2010). Although MGEs have been strongly implicated in the adaptation of *S. aureus* to new host species (Guinane *et al.*, 2010; Spoor *et al.*, 2013), we did not identify acquisition of MGEs among isolates obtained from the experimental infections. Similarly, other studies with different bacterial species also identified large-scale deletions but not insertions during within-host evolution (Price *et al.*, 2013; Smith *et al.*, 2006).

We also examined whether *S. aureus* uses similar mechanisms during short-term and long-term adaptation. We detected positive selection by comparing the ratio of non-synonymous to synonymous mutations, as this method has been extensively used in other bacterial genomes (Alves *et al.*, 2013; Lefébure and Stanhope, 2009; Orsi *et al.*, 2008; Xu *et al.*, 2011). We identified that several genes mutated during the passages and genes under positive selection in ruminant-associated strains encode proteins that are exposed on the bacterial cell surface or associated with the membrane. This was expected considering the encoded antigens interact with the host immune system, causing positive selection to favour novel epitopes that can evade immune recognition (Didelot *et al.*, 2016). Nevertheless, genes that acquired high-effect mutations in the

experimental evolution assays were not similar to those undergoing positive selection in ovine-specific *S. aureus* clones, which could be interpreted as short and long-term adaptation following different evolutionary strategies. Another explanation could be that the adaptive evolution landscape is so extensive that strains achieve host-adaptation through multiple distinct pathways. However, we only tested "site models" of positive selection and for many genes these can be still conservative (Yang, 2006). In addition, one could also think that adaptation does not occur at gene level but at the metabolic pathway level (Su *et al.*, 2013), which was not investigated.

The overall substitution rate was consistent with previous estimates for *S. aureus* (Young *et al.*, 2012). Within-host variation was also observed in *S. aureus* infecting humans (Golubchik *et al.*, 2013; Young *et al.*, 2012) and is in accordance with clonal interference, with bacterial populations accumulating multiple beneficial mutations that continuously compete, slowing down the fixation in the population (Campos and Wahl, 2010; Fogle *et al.*, 2008; Sniegowski and Gerrish, 2010). Within-host variation could also reflect the emergence of subtypes, which are adapting to multiple different niches within the mammary gland, which represents numerous several small ducts leading to thousands of alveoli where bacterial subpopulations could acquire specific-niche adaptations.

The capacity of *S. aureus* to colonize and persist in a new host-species depends on the dynamics of its interaction with the host (Wilson, 2012). We discovered that after a transmission, the subsequent infection is founded by a single clone. Similar population dynamics patterns have been described in intravenous infections of pneumococci,

where after an inoculum of bacteria the resulting septicaemia originated from a single clone that posteriorly diversified into new lineages (Gerlini *et al.*, 2014). McVicker *et al.* (2014) also described population bottlenecks and clonal expansions following initial inoculums of zebrafish with *S. aureus*. In our study, we attribute these bottlenecks to two major events that produced regular shrinking of the population size and subsequent foundational effects. During such scenarios, the power of genetic drift is so overwhelming that mutations accumulate in an effectively neutral fashion (Lynch *et al.*, 2016), allowing deleterious and neutral mutations to fixate in higher proportions (Muller's Ratchet). In addition, due to the stochastic nature of evolutionary processes and the diminishing advantages of increased perfection of molecular adaptations, selection can only achieve an upper limit of refinement in smaller populations (Lynch *et al.*, 2016). However, our data indicate that in spite of the effect of regular bottlenecks on population structure, the infection passages favoured the fixation of beneficial mutations.

In silico evolution analysis supported our results, showing that beneficial mutations increase over time even when bottlenecks are frequent, suggesting that the fitness gain of beneficial SNPs is high enough to overcome genetic drift and sweep through the population. When the population size is constant, multiple beneficial mutations are expected to arise, and competition between them would slow down their progression towards fixation. However, although continuous bottlenecks produce random genetic drift and beneficial mutation loss, as new population emerge from very few individuals, beneficial mutations can achieve fixation quickly after the bottlenecks.

The simulation models are limited by the fact that did not consider differences in replication time, as bacteria with shorter generation time will have a fitness advantage. In addition, the mutation rate and the selection coefficients distribution we used were hypothetical. The variation within each sheep would also depend on the effective population size and the frequency of the sweeps occurring within them. Finally, in natural transmission, genetic drift probably has a higher impact, since probably just one colony is able to penetrate the teat of the next animal. Nevertheless, we had control over most of the parameters used for the simulations, which were consistent with the experimental infection data.

3.6. Conclusions

Experimental infection of animals with human-associated bacterial strains offers an excellent framework to understand adaptive evolution after a host-switch event. Having full control of the environment and the direction of the transmissions allowed us to characterize the molecular mechanisms of the early stages of bacterial adaptive evolution to a new host-species.

Most of the studies on within-host evolution have been performed in static environments such as the lungs or the gut. Here we provide new insights on the adaptive evolution in a dynamic environment, which can help to understand the development of infections in similar environments, such as the urinary tract. The evolutionary dynamics in the face of regular bottlenecks suggests that the fitness gain of beneficial mutations is high enough to overcome genetic drift and sweep through the population.

4

Population genomics of *Legionella longbeachae* and hidden complexities of infection source attribution.

4.1. Introduction

Legionellosis presents as 2 clinically distinct forms: an influenza-like illness called Pontiac fever and a severe pneumonia known as Legionnaires' disease (Fields *et al.*, 2002). In Europe and the United States, most legionellosis cases are caused by *Legionella* pneumophila serogroup 1 (Edelstein, 1982; Fields *et al.*, 2002); <5% of cases are caused by non-pneumophila *Legionella spp*. (Joseph and Ricketts, 2010; Marston *et al.*, 1994). In Australasia, New Zealand, and some countries in Asia, infections caused by *L. longbeachae* occur at comparable levels to infections caused by *L. pneumophila* (Cramp *et al.*, 2010; J. S. Li, *et al.*, 2002; Whiley and Bentham, 2011). Unlike *L. pneumophila* infections, which are typically linked to artificial water systems, *L. longbeachae* infections are associated with exposure to soil, compost, and potting mixes (Yu *et al.*, 2002).

The number of legionellosis cases caused by *L. longbeachae* is increasing worldwide (Whiley and Bentham, 2011), with a notable rise reported across Europe (Den Boer *et al.*, 2007; García, *et al.*, 2004; Potts *et al.*, 2013). Within the United Kingdom, most *L. longbeachae* infections have been identified in Scotland, where 6 cases were diagnosed during 2008–2012 (Lindsay *et al.*, 2012) and another 6 were diagnosed in the summer of 2013 and represented a singular increased incidence or cluster with all patients requiring intensive care hospitalization (Potts *et al.*, 2013). Epidemiologic investigation revealed that most patients from the 2013 cluster were avid gardeners, and *L. longbeachae* was isolated from respiratory secretions and from samples of the growing media they had used for gardening before becoming ill (Lindsay *et al.*, 2012; Potts *et al.*, 2013). However, an investigation into the provenance of the growing

media did not reveal a single commercial or manufacturing source that would suggest a common origin for the *L. longbeachae* associated with the outbreak (Lindsay *et al.*, 2012).

Molecular typing methods used to discriminate between *L. longbeachae* and other *Legionella* spp. and between the 2 *L. longbeachae* serogroups have limited efficacy, and although considerable evidence supports growing media as a source for *L. longbeachae* infections (Koide *et al.*, 2001; Steele *et al.*, 1990), there is still a lack of genetic evidence for an epidemiologic link. Furthermore, a population genomic study involving large numbers of *L. pneumophila* isolates has been conducted (Rao *et al.*, 2013; Reuter *et al.*, 2013), but the same has not been done for *L. longbeachae*, so the diversity of environmental and pathogenic genotypes and the relationship between them remains unknown for *L. longbeachae*. To examine the aetiology of the 2013 cluster of legionellosis cases in Scotland in the context of *L. longbeachae* species diversity, we analysed the genomes of 70 *Legionella* spp. isolates from 4 countries over 18 years.

4.2. Aims

- To investigate the epidemiology of the cluster of legionellosis cases in Scotland using population genomic approaches.
- To examine the population structure of *L. longbeachae* from Scotland in the context of the global species diversity.
- To investigate the impact of recombination and intraspecies and interspecies gene flow on the genetic diversity of this pathogen.

4.3. Material and Methods

4.3.1. Bacterial isolates

We sequenced 65 isolates that had previously been identified as *L. longbeachae*. These isolates were obtained during 1996–2014 from several patients, growing media samples (including compost and soil), and a hot water supply. Of these isolates, 55 were from Scotland (29 from the 2013 cluster of infections and 26 from other clinical and environmental samples) and 10 were from patients and environmental compost samples in New Zealand (Supplementary Table 6).

In our analysis, we also included all publicly available genome sequences for *L. longbeachae*: *L. longbeachae* NSW150 (serogroup 1) and *L. longbeachae* C-4E7 (serogroup 2) isolated from patients in Australia; and *L. longbeachae* D-4968 (serogroup 1), *L. longbeachae* ATCC39642 (serogroup 1), and *L. longbeachae* 98072 (serogroup 2) isolated from patients in the United States (Cazalet *et al.*, 2010; Gomez-Valero *et al.*, 2011; Kozak *et al.*, 2010). We sequenced multiple isolates (n = 2 to 5) for each of 3 patients and their linked growing media samples from the 2013 outbreak in Scotland and for 2 additional compost samples. The species of all isolates had been determined by serotyping or macrophage infectivity potentiator (*mip*) gene sequencing (Fallon and Abraham, 1983; Ratcliff *et al.*, 1998).

4.3.2. Bacterial culture, genomic DNA isolation, and WGS

We cultured *Legionella* spp. isolates in a microaerophilic and humid environment at 37°C on BCYE (buffered charcoal yeast extract) agar plates for 48 h. We then picked individual colonies from the plates and grew them in ACES-buffered yeast extract

broth containing *Legionella* BCYE Growth Supplement (Oxoid Ltd., Basingstoke, UK) with shaking at 37°C for 24–48 h. We extracted genomic DNA from fresh cultures by using the QIAGEN DNeasy Blood and Tissue Kit (QIAGEN Benelux B.V., Venlo, the Netherlands).

We prepared sequencing libraries by using the Nextera XT kit for MiSeq or HiSeq (all from Illumina, San Diego, CA, USA) sequencing at Edinburgh Genomics, University of Edinburgh (Edinburgh, Scotland, UK). For each isolate, one 2 × 250–bp or two 2 × 200–bp paired-end sequencing runs were carried out using the MiSeq and HiSeq technologies, respectively. Raw reads were quality checked using FastQC v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and primers were trimmed by using Cutadapt (Martin, 2011). We used wgsim software (https://github.com/lh3/wgsim) to simulate sequence reads for publicly available, complete whole-genome sequences.

4.3.3. Genome assemblies and variant calling

De novo assemblies of the *Legionella* isolates were produced using SPAdes 2.5.1. (Bankevich *et al.*, 2012) (k values of 21, 33, 55, 77, 99 and 127), generating a median of 106 contigs per genome (range, 38– 402 contigs), with an average of 4.16 Mb in length (3.98–4.52 Mb) and an average N50 of 130 kb (29 kb-291 kb).

The error-corrected reads produced by SPAdes were mapped against the *Legionella longbeachae* reference genome of strain NSW150 (GenBank accession number NC_013861) using BWA 0.5.9 (Li and Durbin, 2010) with default parameters. SNPs

were called using Samtools 1.18 (Li *et al.*, 2009) and those absent in at least 30% of the reads, with quality below 30 and depth below 3 were filtered out. The output from this filtering was used to construct consensus genomes of all the isolates for further phylogenetic analyses.

4.3.4. Analysis of genome content

The contigs were annotated using Prokka v1.10 (Seemann, 2014) and orthologous genes were clustered using the algorithm OrthoMCL (Li *et al.*, 2003) integrated in the software Get_homologs (Contreras-Moreira and Vinuesa, 2013). We selected the options –f 50 (filters by 50% length difference within clusters) and –t 0 (for reporting all the clusters), resulting in 1801 core genome clusters. This program was also run using the Sg1 isolates only as input, specifying the options minimum percentage coverage (-C 80) and percentage identity (-S 85), which generated a core genome of 2574 gene clusters.

We also used JSpecies (Richter and Rosselló-Móra, 2009) to compute the average nucleotide identity values (BLAST; ANIb) between several pairs of isolates. These ANIb results were represented on a plot where isolates were clustered according to the 16S rRNA phylogenetic tree. In addition, a pangenomic tree of the OMCL binary matrix from get_homologs.pl using all the *Legionella* isolates was constructed using the compare_clusters.pl script.

4.3.5. Evolutionary and phylogenetic analysis

To confirm the identity of the isolates, a Neighbour-Joining tree based on the 16S rRNA gene of the sequenced genomes and all the cultured type *Legionella* strains available in the Ribosomal Database Project (J. R. Cole *et al*., 2014) (as of 01/06/2015) was constructed. The RNAmmer 1.2 server (Lagesen *et al.*, 2007) was used to identify the 16S rRNA genes in the *de novo* assemblies, which were then aligned using MUSCLE with default parameters (Edgar, 2004). The Neighbor-Joining tree was estimated using the Hasegawa–Kishino–Yano model and 1000 bootstrap resampling replicates using the program Geneious 5.4.6 (Kearse *et al.*, 2012).

To construct a phylogeny based on the whole genome sequence data, the 1801 orthologous open reading frames identified using OrthoMCL were aligned using MUSCLE 3.8.31 (Edgar, 2004). Individual protein alignments were translated back to DNA alignments using pal2nal v14 (Suyama *et al.*, 2006) and the resultant alignments were concatenated using catfasta2phyml.pl (https://github.com/nylander/catfasta2phyml) into an 1110024 bp long super-alignment. A ML phylogenetic tree was estimated based on this alignment using RAxML. C-4E7 was excluded from the original clustering as the low quality of the assembly significantly reduced the size of the core genome.

The *L. longbeachae* phylogeny was reconstructed using a Neighbour-Joining approach in Splitstree4 (Huson and Bryant, 2006). Phylogenies of *L. longbeachae* Sg1 isolates before and after removing recombination were reconstructed from the genome alignments using RAxML 7.2.6 (Stamatakis, 2006).

4.3.6. Detection of recombination

Recombination was examined using the SplitsTree4 program (version 4.13.1) (Huson and Bryant, 2006). A phylogenetic network was computed on the *L. longbeachae* Sg1 multiple genome alignment using the Neighbour-Net method implemented in this software. The statistical significance of the tree was confirmed using a Phi test (Bruen *et al.*, 2006). Recombination was detected on the core genome alignment of the Sg1 isolates using BratNextGen (Marttinen *et al.*, 2012). After drawing a PSA tree, we selected a cutoff of 0.042, which split the tree into 8 clusters. We used 20 iterations for the recombination learning algorithm and after performing 100 replicate runs in a single processor we selected a threshold of 5% for estimating the significance of recombination. Finally, the *L. longbeachae* Sg1 ML tree and the whole genome alignment were used as input for ClonalFrameML (Didelot and Wilson, 2015) to generate a phylogeny with branch lengths corrected for recombination. 100 pseudo-bootstrap replicates were used to estimate the uncertainty in the EM model and the option - ignore_user_sites with the list of non-core coordinates was parsed.

4.3.7. Plasmid analysis

We used PLACNET, a software that constructs a network of contigs interactions, for the identification and visualization of plasmids (Lanza *et al.*, 2014). Bowtie2 v2.0.6 (Langmead and Salzberg, 2012) was first used to find all possible scaffold links of the contigs by mapping the reads to them. Length and insert sizes of the reads mapped were calculated using Picard-Tools v1.90 (https://sourceforge.net/projects/picard/files/picard-tools). These files and metrics were parsed as input for placnet.pl, which produced a plasmid network from which we extracted the scaffolds. We then

performed a BLAST search of the contigs assembly files to a database containing all the bacteria and plasmids genomes available in the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/) (created in March 2015) and results were filtered as follows: contigs longer than 200 bp, with a bitscore below 1e-26 and that had at least a blast hit over 5% of the contig size. The results were further analysed to classify the nodes into one of these categories: "hit completely to a single reference genome," "split nodes that hit to a single reference genome" and "nodes that hit several genomes." The hits and the scaffolds were combined into a network that was uploaded into Cytoscape (Smoot *et al.*, 2011). Recommendations given in the PLACNET manual were followed to visualize the chromosome and plasmids networks. BLAST was finally used to search for *Legionella* spp. plasmid related sequences in the contigs.

4.4. Results

4.4.1. Limitations of current typing approaches for *Legionella* spp. identification

We sequenced 65 isolates obtained from several patients and environmental samples over 18 years in different countries and previously identified as *L. longbeachae*. To confirm the species identity of the *Legionella* isolates, we constructed a phylogenetic tree that included all *Legionella* type strains for which cultures are available, based on the 16S rRNA gene sequence (Cole *et al.*, 2014). We also built phylogenetic trees based on the whole-genome content and core-genome diversity. For each approach, 64 of the 70 isolates examined co-segregated within the *L. longbeachae*–specific clade, 4 isolates clustered with *Legionella anisa*, and 2 belonged to a separate clade that was distinct from all known *Legionella* spp. (Figure 4.1, Figure 4.3, Figure 4.3).

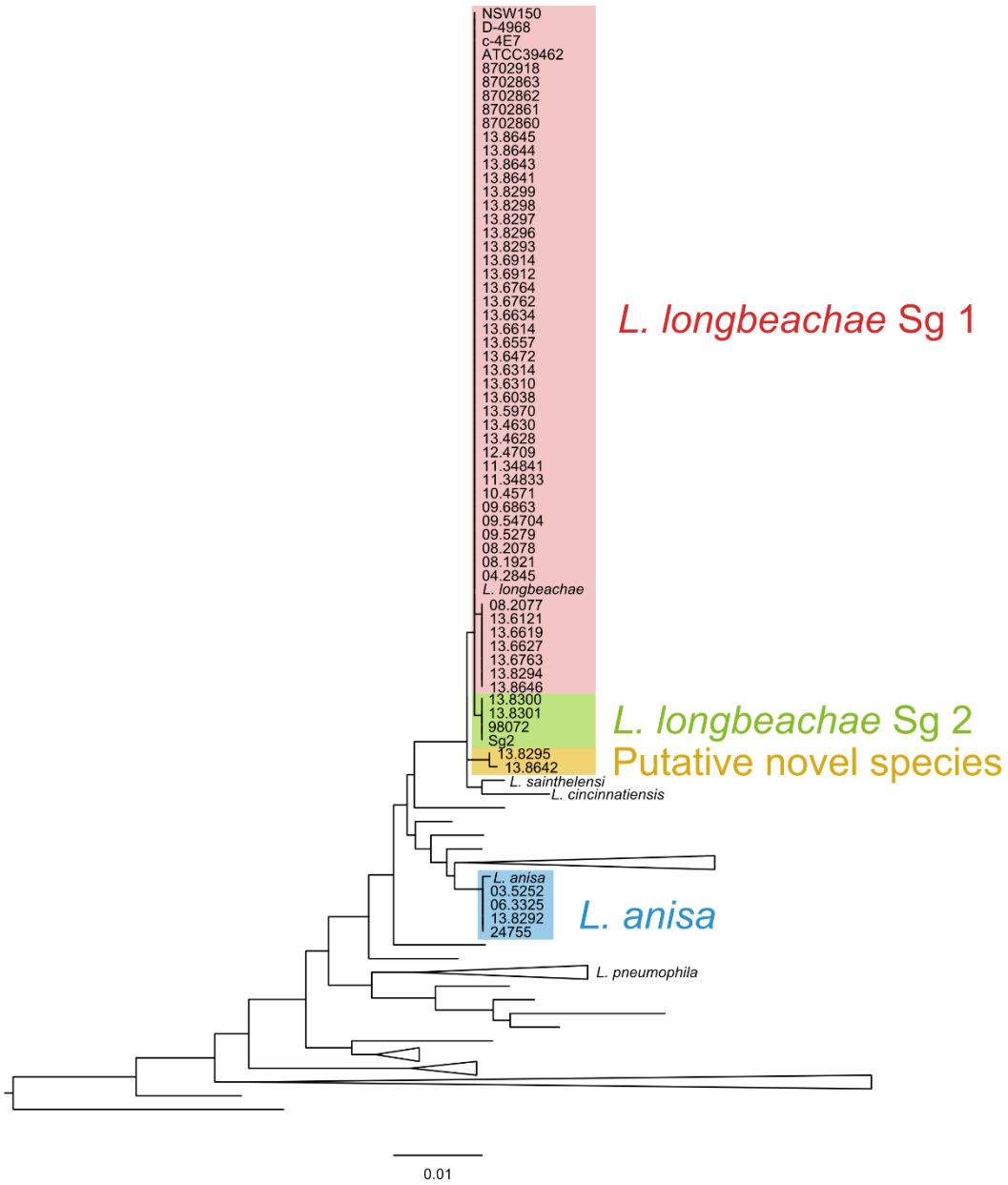


Figure 4.1. **16S rRNA gene–based phylogenetic tree**. Sequenced genomes and all the cultured and type *Legionella* spp. strains available in the ribosomal database project (http://rdp.cme.msu.edu/), as accessed in May 2015, were included. Scale bar indicates the mean number of nucleotide substitutions per site. The isolates and species identified in this study are coloured, in contrast to the *Legionella spp*. groups that have not bee identified among our samples.

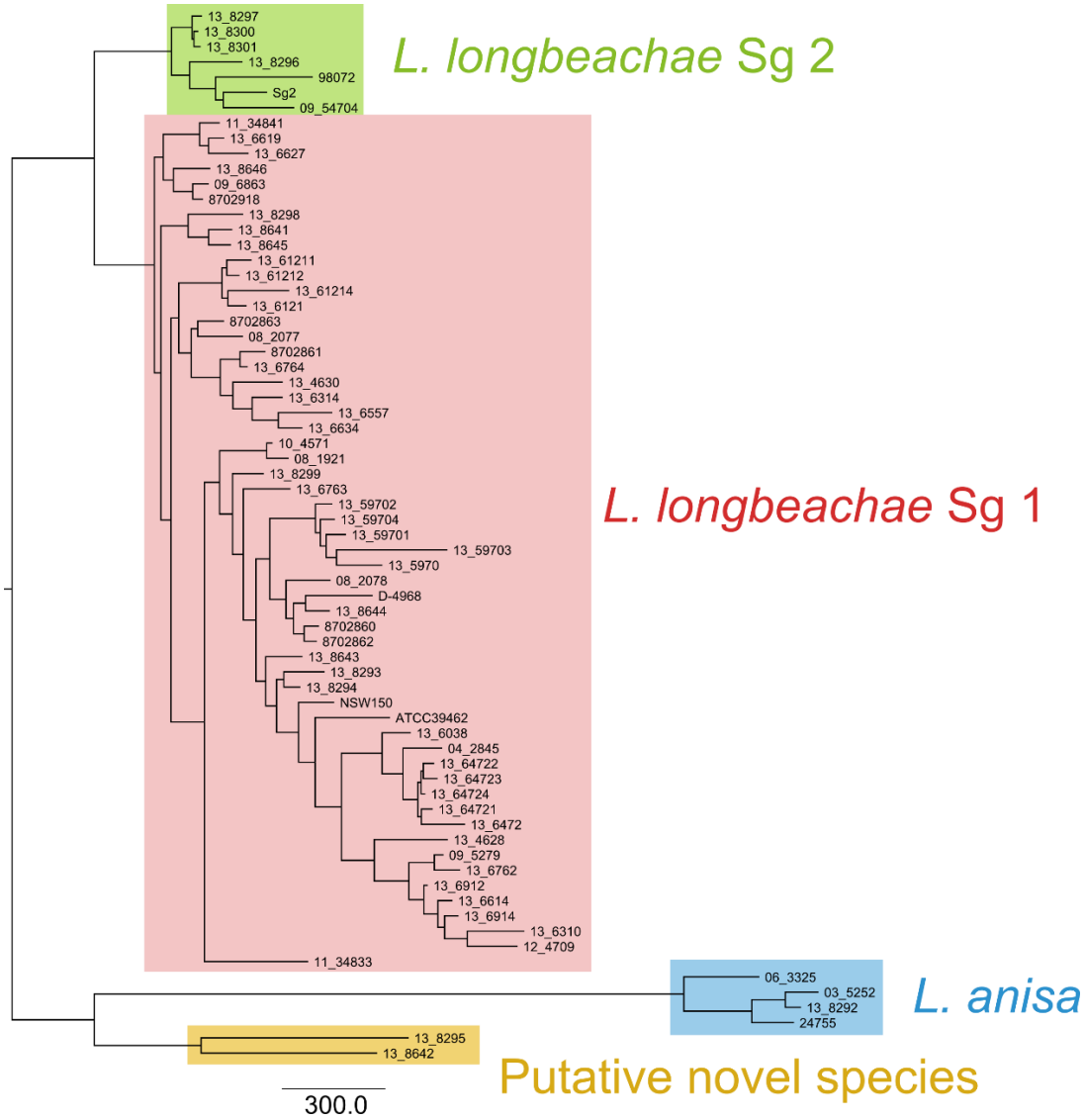


Figure 4.2. Parsimony based tree of the OMCL pangenomic matrix obtained for all the sequenced genomes. Scale bar indicates the gene content differences. The isolates are coloured according to their species or serogroup.

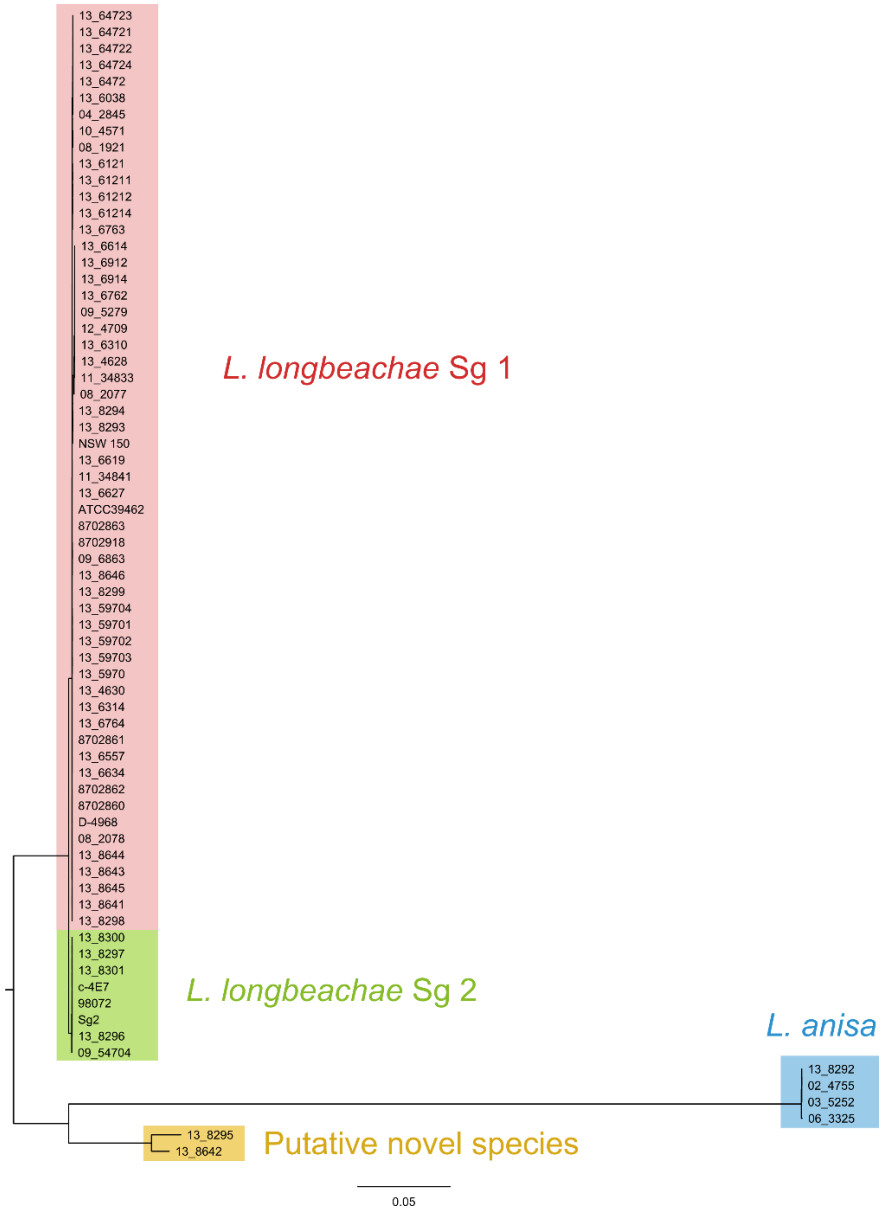


Figure 4.3. Maximum likelihood tree of a core gene alignment of all the isolates included in the study. The tree shows the same clusters as the 16S rRNA gene based tree and the parsimony pangenome tree. Scale bar indicates the mean number of nucleotide substitutions per site. The isolates are coloured according to their species or serogroup.

The species identities were further supported by determination of the average nucleotide identity values, a widely used method for bacterial species delineation based on genomic relatedness (Kim, *et al.*, 2014). Of note, *L. anisa* is the most common non-pneumophila *Legionella* spp. in Europe (HPS, 2015; Mee-marquet *et al.*, 2006; Svarrer and Uldum, 2012). In addition, *L. longbeachae* isolates 13.8642 (from a compost sample from Scotland) and 13.8295 (from a patient in New Zealand) belong to a putative novel *Legionella* spp. Overall, the data indicate that current serotyping methods and mip gene sequencing are limited in their capacity to identify *L. longbeachae* to the species level.

To investigate the genetic relatedness of *L. longbeachae* strains associated with the 2013 outbreak to temporally and geographically distinct isolates, we constructed a core genome–based neighbour-joining tree of the 64 confirmed *L. longbeachae* isolates obtained from 4 countries over 18 years (Figure 4.4). This phylogenetic tree presents a comet-like pattern, with 2 distinct clades separated by 9,911 single-nucleotide polymorphisms, representing the major serogroups (serogroups 1 and 2) previously identified for *L. longbeachae* (Ratcliff *et al.*, 1998), each containing isolates from patient and environmental samples from different years. In contrast with findings from a previous analysis of 2 isolates of *L. longbeachae* serogroup 1 (Gomez-Valero *et al.*, 2011), we observed a higher diversity among the 56 isolates within serogroup 1 (Figure 4.1, Figure 4.3); this finding is not unexpected, given the difference in the number of genomes examined.

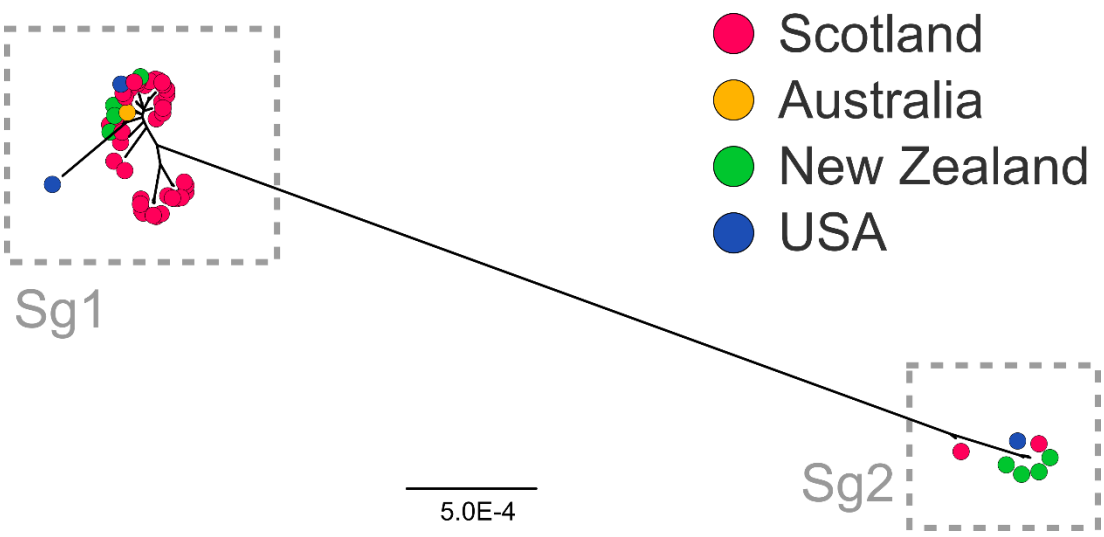


Figure 4.4. Neighbour-Joining phylogeny based on the core genome of *Legionella longbeachae* isolates. Isolates are coloured by geographic source, and dashed boxes indicate the defined or predicted serogroups to which the isolates belong. The two serogroups Sg1 and Sg2 can be easily differentiated, Sg1 on the left and Sg2 on the right side of the tree.

Nevertheless, compared with isolates from the same serogroup in other *Legionella* spp., such as *L. pneumophila* serogroup 1 (2% polymorphism), *L. longbeachae* serogroup 1 exhibits a lower diversity (<0.1% polymorphism). Although serogroup 1 and 2 clades contained isolates from Scotland, Australasia, and the United States, 96% of the isolates from Scotland (including all of the 2013 outbreak isolates) belonged to serogroup 1, suggesting that serogroup 1 may be more clinically relevant in Scotland than in some other countries where *L. longbeachae* is a more established cause of legionellosis. However, analysis of more isolates from different countries would be required to investigate this observation further.

4.4.2. Effect of recombination on *L. longbeachae* serogroup 1 population structure

It is established that recombination has played a key role in shaping the evolutionary history of *L. pneumophila*, but its effect on *L. longbeachae* population structure is unknown (Gomez-Valero *et al.*, 2011; Underwood *et al.*, 2013). This knowledge is critical because for highly recombinant bacteria, recombination networks may represent evolutionary relationships more explicitly than traditional phylogenetic trees. Therefore, we constructed a recombination network of all serogroup 1 isolates by using the neighbour-net algorithm of SplitsTree4 (Huson and Bryant, 2006). The resultant network displayed a reticulate topology with an extensive reticulated background from which clusters of isolates emerge, supporting an evolutionary history involving recombination ($p < 0.01$ by ϕ test) (Bruen *et al.*, 2006), followed by clonal expansion and subsequent additional recombination events among some lineages (Figure 4.5). Using BratNextGen (33), we identified a total of 94 predicted

recombination events affecting more than half of the core genome (1.74 Mb of 3.36 Mb) and representing recent and ancient recombination events of different sizes (range 1,350 bp–350 kb) distributed across the phylogeny (Figure 4.6). Given the reported limitation in sensitivity of BratNextGen for the identification of all recombination events (de Been *et al.*, 2013), we also used ClonalFrameML (Didelot and Wilson, 2015), an algorithm that uses maximum likelihood inference to simultaneously detect recombination in bacterial genomes and account for it in phylogenetic reconstruction. The estimated average length of the recombined fragments was 8,047 bp, and the ratio of recombination to mutation was 1.42, indicating a greater role for recombination over mutation in the diversification of *L. longbeachae*. This estimate is in accordance with early estimates for *L. pneumophila* based on multiple gene sequence data (Coscollá *et al.*, 2011), but it is low compared with recent estimates based on whole-genome sequence data (recombination to mutation ratios of 16.8 (Underwood *et al.*, 2013) or 47.93 (Sánchez-Busó, *et al.*, 2014)). Differences in the clonal diversity of *Legionella* spp. sequence datasets used to determine recombination rates could affect the estimates. Reconstruction of the phylogeny after removal of all predicted recombinant sequences resulted in a tree with largely similar clusters of isolates but with reduced branch lengths and variation in the position of nodes deep in the phylogeny (Figure 4.7).

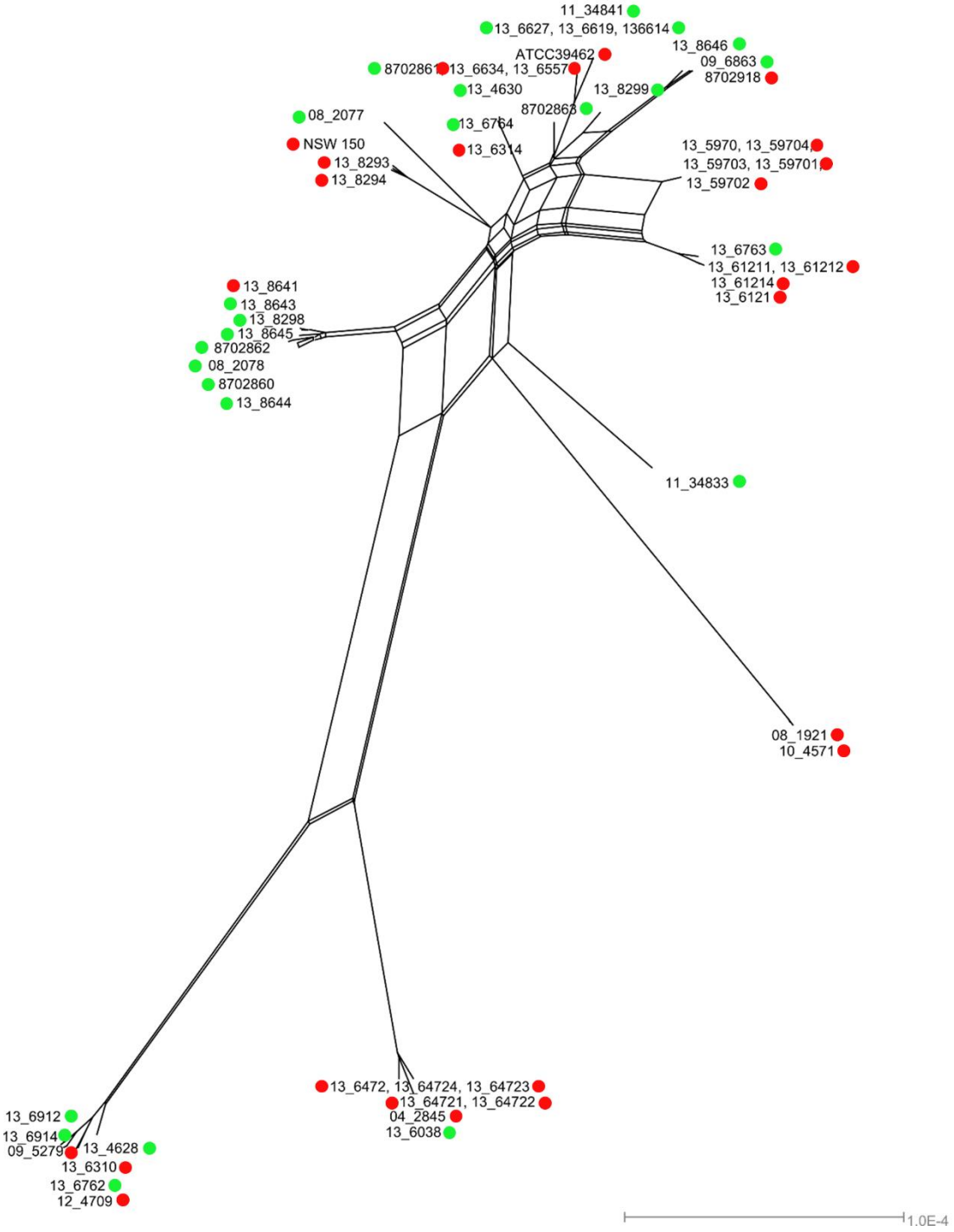


Figure 4.5. **Neighbour-Joining split network**. Only the *Legionella longbeachae* Serogroup 1 isolates are included (patients-related in red, environmental-related in green), the network is based on the consensus alignment obtained from mapping every isolate to the reference chromosome NSW150. Scale bar indicates the mean number of nucleotide substitutions per site.

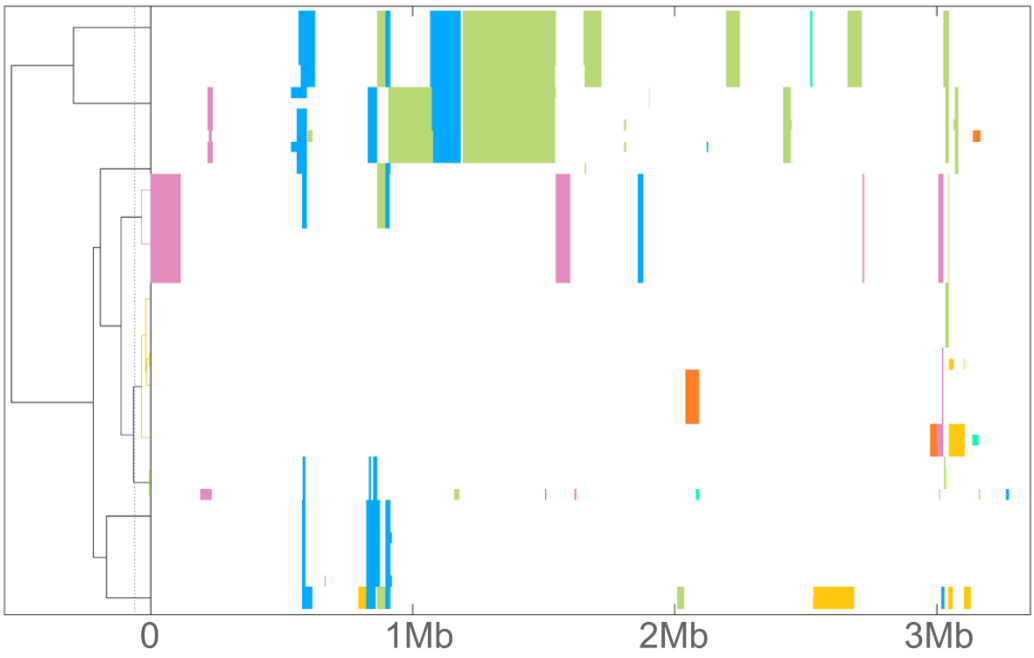


Figure 4.6. Recombinant regions of the core genome alignment of 55 *L. longbeachae* Sg1 isolates as identified using BratNextGen. On the left, a clustering tree of the isolates with coloured branches indicating cluster relationships. On the right, significant recombinant segments predicted, with similar colour in a column representing recombinant regions for those isolates have the same origin.

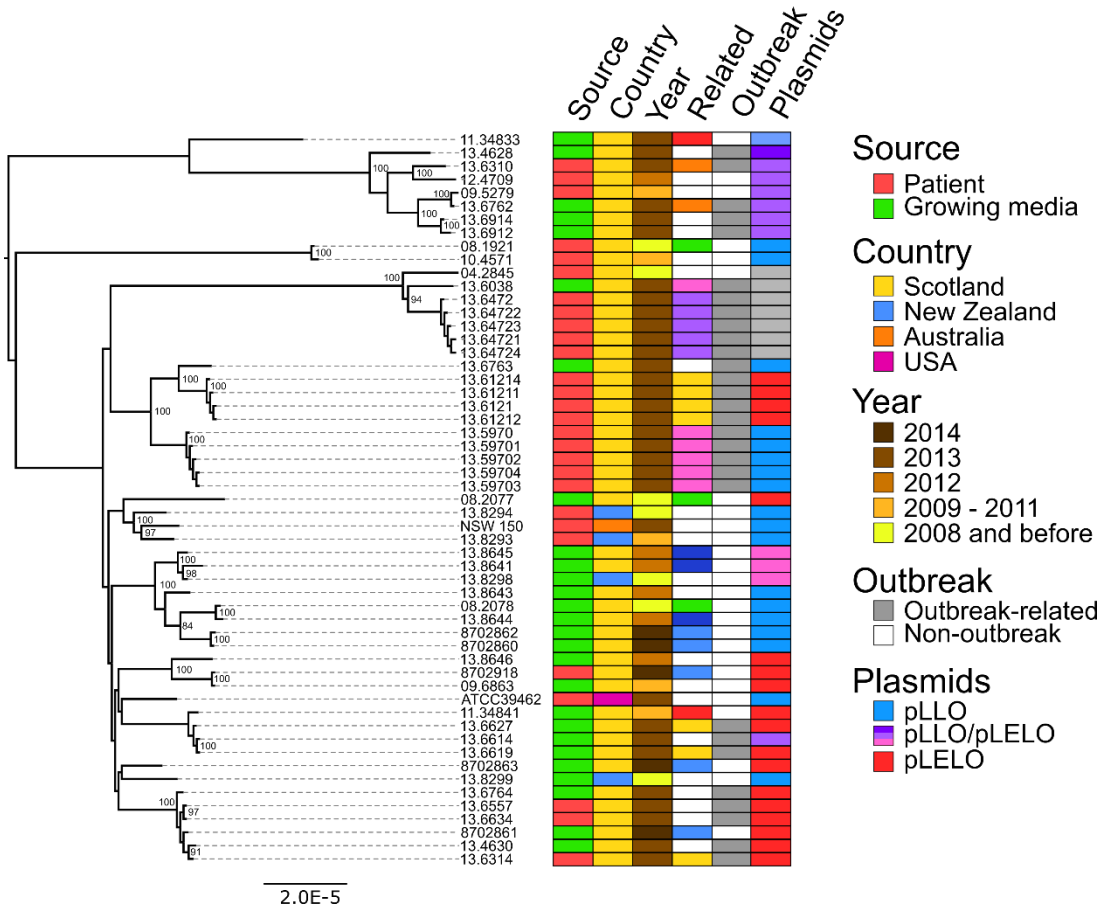


Figure 4.7. Core genome–based maximum-likelihood phylogeny of *Legionella longbeachae* serogroup 1 isolates corrected for recombination. Source, country, year of isolation, relatedness and plasmid carriage are indicated. Related isolates include those from the same patient or their cognate environment and are shown in the same colour. Isolates from the 2013 outbreak are indicated in grey. Of note, isolates from the same patients are clustered together but do not co-segregate with their respective compost samples. Scale bar indicates the mean number of nucleotide substitutions per site.

4.4.3. Accessory genome analysis indicates extensive interspecies and intraspecies gene flow

The extent to which horizontal gene transfer occurs among *L. longbeachae* isolates and between *L. longbeachae* and other *Legionella* spp. is unknown. In our study, the pangenome of *L. longbeachae* represented by the 56 serogroup 1 isolates was 6,890 genes, including a core genome of 2,574 genes; the average gene content was 3,558 genes per strain. The accessory genome, which included only strain-dependent genes varied from 809 to 1,155 genes, depending on the strain. A parsimony clustering analysis based on the presence or absence of all genes classified the isolates in a manner distinct from that in a core genome–based maximum-likelihood tree, suggesting extensive horizontal gene transfer among *L. longbeachae* isolates. BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis of all assembled contigs was used to filter for plasmid-related homologous sequences, revealed 2 major plasmids: pLLO, described previously in *L. longbeachae* NSW150 (Cazalet *et al.*, 2010), and pLELO, originally identified in *L. pneumophila* subsp. *pneumophila* (Gomez-Valero *et al.*, 2011). Of the 55 serogroup 1 isolates, 36 contained sequences for the pLLO and pLELO plasmids. Of note, the distribution of these plasmids among the *L. longbeachae* isolates correlated with the gene content–based clustering, whereas the distribution of plasmids in the core genome–based tree was independent of the phylogeny (Figure 4.7). In addition, 11 isolates appeared to contain plasmids with sequences homologous to those for pLLO and pLELO, which is indicative of recombinant forms of the plasmid.

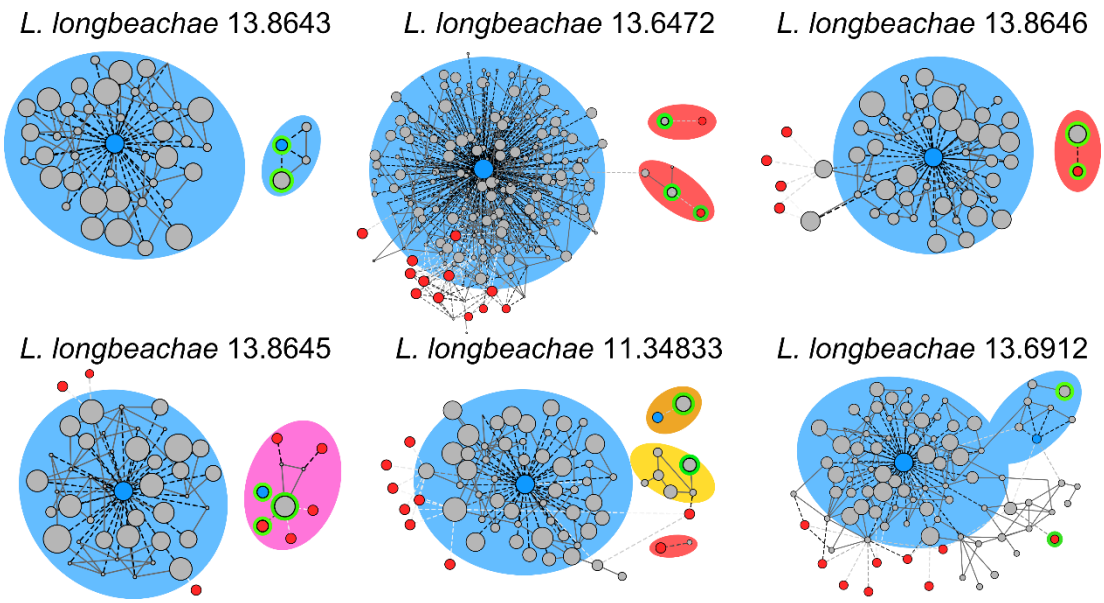


Figure 4.8. *Legionella longbeachae* plasmid analysis. Contigs networks reconstructions for 6 representative *L. longbeachae* types of plasmid content. The networks of the contigs representing the main chromosome and plasmids comprising the genome obtained by using PLACNET (38), a program enabling reconstruction of plasmids from whole-genome sequence datasets. The sizes of the contig nodes (in grey) are proportional to their lengths; continuous lines correspond to scaffold links. Dashed lines represent BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) hits to the *L. longbeachae* (blue) or *L. pneumophila* (red) strains; intensity of the line is proportional to the hit (white indicates low, black indicates high). Green lines correspond to plasmid contigs. Background colours indicate species relatedness for the main chromosome and plasmids (blue for *L. longbeachae*, red for *L. pneumophila*, pink for a combination of both, and yellow for previously unidentified genomic content).

Further examination of plasmid diversity using a modified version of PLACNET (Lanza *et al.*, 2014), a program enabling reconstruction of plasmids from whole-genome sequence datasets, confirmed that some plasmids consisted of a mosaic of recombinant fragments homologous to pLELO, pLLO, or other unknown plasmids (Figure 4.8). Overall, these data indicate the high prevalence of specific plasmids among *L. longbeachae* isolates and reveal extensive recombination and horizontal gene transfer among different *Legionella spp.* (Cazalet *et al.*, 2004). The high prevalence of plasmids in *L. longbeachae* is notable, considering these elements may be less common in *L. pneumophila* (Underwood *et al.*, 2013).

To examine the possibility that clinical and environmental isolates of *L. longbeachae* contained genomic differences reflecting their distinct origins, we compared their accessory genome content. For isolates obtained from a single patient sample, the accessory genome was highly conserved compared with those for environmental isolates from a single compost sample or closely related environmental isolates from distinct compost samples (Figure 4.9a). In addition, considering the average gene content of all sequenced isolates (28 clinical and 27 environmental), the gene content for *L. longbeachae* from growing media samples (3,586 genes) was significantly higher than that for isolates from patients (3,533 genes; 2-sample t-test, $t = 2.5213$; d.f. = 53; $p = 0.01474$) (Figure 4.9b). The data imply that gene loss occurs during human infection or that *L. longbeachae* strains with reduced gene content have enhanced human infectivity. However, we did not identify a specific enriched gene or functional category in clinical or environmental samples (data not shown).

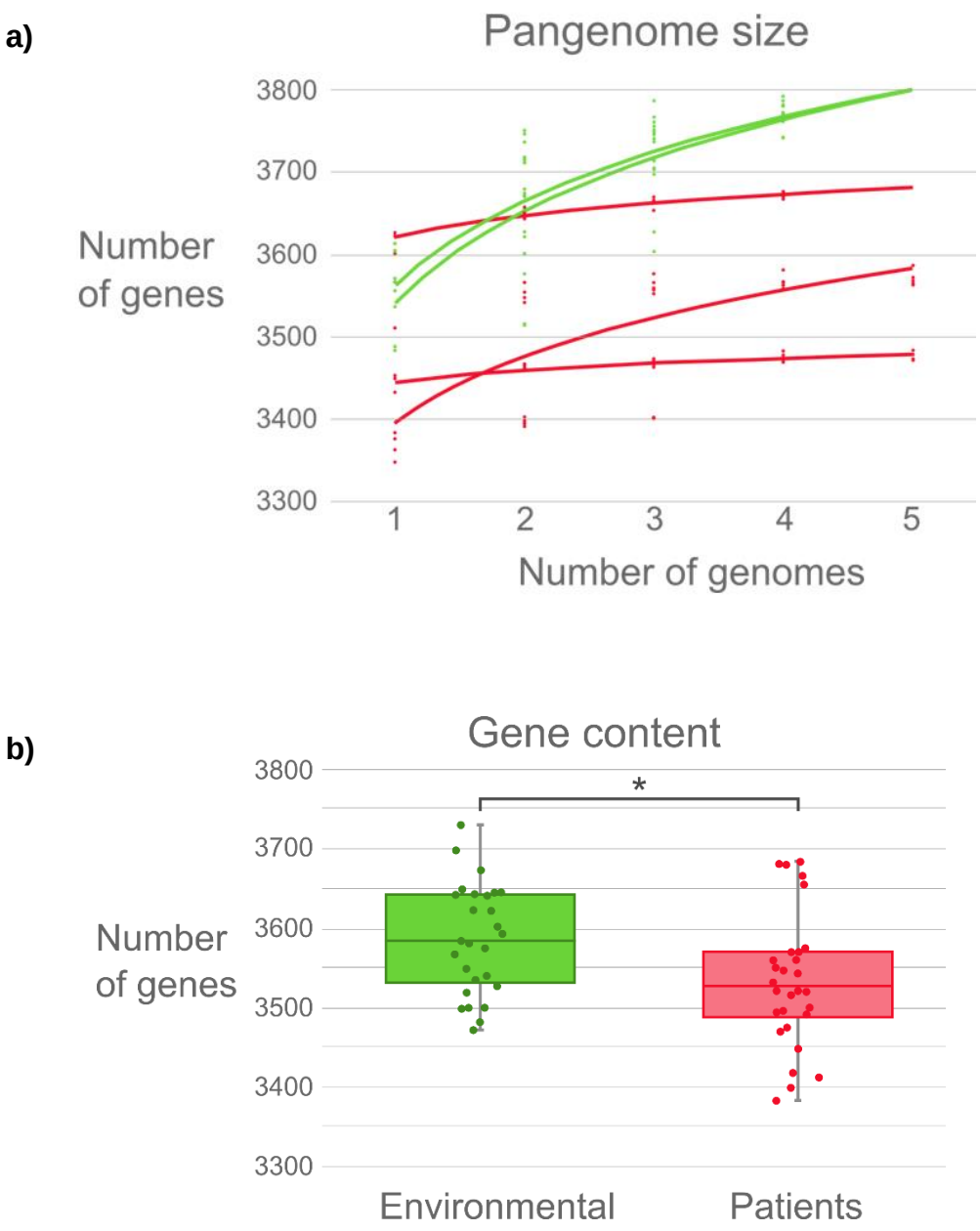


Figure 4.9. Variation in gene content between environmental and patient *Legionella longbeachae* samples. a) Increase in pangenome size with every addition of a *L. longbeachae* genome. Environmental isolates pangenomes (green) are larger and continue increasing after the addition of 5 genomes, consistent with an open pangenome, but the within-patient pangenome plateaus quickly, consistent with a more closed pangenome. b) Average gene content of environmental isolates is significantly higher than that of clinical isolates ($p = 0.01474$).

4.4.4. Source attribution confounded by complex serogroup 1 populations within environmental samples

Having accounted for the influence of recombination on the phylogeny of *L. longbeachae*, we investigated the diversity of isolates associated with 5 patients and their linked compost samples obtained during 2008–2014, including 3 patients from the 2013 outbreak in Scotland. Of note, isolates from the 2013 outbreak were distributed across several subclades of the tree, indicating that the infections were caused by different strains (Figure 4.7). However, all isolates from a single patient clustered together, consistent with a monoclonal aetiology of each infection. Of note, for all 5 patients, clinical isolates were not closely allied to the environmental isolates obtained from linked compost samples, and therefore a genetic link between patient and compost samples could not be established. Most subclades included isolates of diverse geographic origin, consistent with a wide distribution for *L. longbeachae* strains; however, 3 *L. longbeachae* isolates originating from Australasia (strains 13.8294, 13.8293, and NSW150) belonged to their own region-specific cluster (Figure 4.7).

We hypothesized that the lack of genetic relatedness between *L. longbeachae* isolates from patients and linked compost samples could be explained by a highly diverse population of *L. longbeachae* in growing media samples compounded by a sampling strategy consisting of a single sequenced isolate. All 5 compost samples for which we had >1 isolate contained isolates distributed across multiple clades in the phylogenetic tree. In particular, 4 isolates from the same growing media sample linked to a patient infected in Edinburgh in 2014 were distributed across 4 distinct clades, demonstrating

that within a single environmental sample, considerable species diversity may be represented (Figure 4.7). Taken together, these data suggest that for future outbreak investigations, extensive sampling of environmental samples may be required to identify genotypes responsible for episodes of legionellosis infection, if indeed they are present.

4.5. Discussion

Our findings reveal the population genomic structure for *L. longbeachae*, an emerging pathogen in Europe and the United States, and includes a genome-scale investigation into an outbreak of *L. longbeachae* legionellosis. We provide evidence for extensive recombination and lateral gene transfer among *L. longbeachae*, including the presence of widely distributed mosaic plasmids that have likely recombined with plasmids from other *Legionella* spp., suggesting an ecologic overlap or shared habitat. Our analysis highlights the need to account for recombination events when determining the genetic relatedness of *L. longbeachae* isolates.

Our application of whole-genome sequencing for diagnostic purposes revealed the misidentification, using current serotyping methods, of several *L. anisa* isolates as *L. longbeachae* and led to the identification of a putative novel *Legionella* sp. linked to legionellosis. These findings highlight the limitations of current typing methods for differentiation of *Legionella* spp. and accurate identification of legionellosis aetiology.

We used whole-genome sequencing to attempt to establish a genetic link between legionellosis infections and associated compost samples. Our inability to establish a

link probably reflects the traditional strategy of single isolate sampling, which when applied to a highly diverse pool of *L. longbeachae* genotypes fails to detect the infecting genotype. We suggest that the approach to investigating the source of future legionellosis cases linked to growing media will require a radical revision of sampling protocols to maximize the chances of isolating the infecting strain, if present.

4.6. Conclusions

Taken together, our findings provide a view of the population structure of *L. longbeachae* and highlight the complexities of tracing the origin of legionellosis associated with growing media. Overall, our findings demonstrate the resolution afforded by whole-genome sequencing for understanding the biology underpinning legionellosis and provide information that should be considered for future epidemiologic investigations.

5

General discussion

Pathogenic bacteria are a main cause of human disease and mortality worldwide. The history of humanity has been interweaved with the impact of infectious diseases, and major epidemics of plague, smallpox and yellow fever, among others, have caused millions of deaths over the centuries (Tognotti, 2013; Wolfe *et al.*, 2007). The industrialization and urbanization during the 19th century contributed to improved sanitation and hygiene, which combined with control and prevention measures introduced during the 20th century, led to the reduction of the incidence of several infectious diseases (Bloomfield *et al.*, 2006). In addition, the discovery of antibiotics and the development of vaccines permitted the eradication of smallpox virus and the elimination of other infections in specific regions (Bazin, 2003). With such promising successes, the end of infectious diseases seemed a feasible reality at one stage (Brachman, 2003). However, the appearance of AIDS in the 1980s, the re-emergence of typhus, cholera and tuberculosis, and an overall increase in mortality associated with other infections, indicated that bacterial pathogens would remain a burden of disease (Fonkwo, 2008). Demographic, socio-economic and environmental factors are contributing to the emergence of new pathogens, and misuse of antibiotics both in humans and animals have driven the re-emergence of diseases that had been successfully controlled (Michael *et al.*, 2014; Morse, 1995). Notably, most emerging human infectious diseases have been traced to an animal origin (Taylor *et al.*, 2001; Wiethoelter *et al.*, 2015). Understanding the ecology and evolution of bacterial pathogens at the population level is essential to unveil the molecular basis of infectious diseases. In this thesis, WGS and population genomic approaches provide new insights into the evolutionary dynamics of pathogens, molecular basis of host-adaptation and the aetiology of a new emerging disease, which represent valuable information that

could be applied to the development of novel therapies or the design of effective surveillance and control measures.

The current study provided broad insights into the evolutionary landscape of the *S. aureus* species in the context of its host-associations. Traditionally, molecular typing approaches were used, but the low resolution achieved by such methods did not provide sufficient discriminatory power to infer certain transmissions. With the high resolution afforded by WGS, population genomic analyses have contributed substantially to our understanding of the evolutionary and epidemiological processes promoting the transfer of pathogens from animals to humans (Assefa *et al.*, 2015; Haagmans *et al.*, 2009; Weis *et al.*, 2016), and numerous studies have traced the dynamics of human diseases with zoonotic origins (Kilpatrick and Randolph, 2012; Wiethoelter *et al.*, 2015). In this work, the high-resolution phylogeny of hundreds of isolates associated with multiple host-species supported a stratification of the clonal host-specificity. This has important public health implications for assessing the relevance of future emerging epidemic clones. Knowing if an emerging strain descends from a generalist or a host-restricted lineage can provide information relevant to understanding the likelihood of that strain expanding in the new host population. This information can be relevant for other pathogens showing similar host-association relationships. For example, *C. jejuni*, *E. coli* and species of the *Salmonella* genus are also characterized by presenting sub-lineages able to infect an ample range of host-species and specialist sub-clones restricted to a single host (Llarena *et al.*, 2016; Strachan *et al.*, 2015). In addition, knowing the level of host-specialization for individual lineages can also provide information on the virulence of potential emerging

clones, since host-adapted *S. aureus* strains may often be more virulent than other generalist clones (Shepheard *et al*., 2013).

Furthermore, identifying the most common routes for *S. aureus* host-species switches between animals and humans can be helpful for the development of more effective control measures. Generally, animals act as reservoirs for the transmission of pathogens into humans (Day *et al.*, 2012), and changing livestock management methods, global trade and increasing interaction between humans, wildlife and domesticated animals represent risk factors for zoonotic transmissions (Marano and Pappaioanou, 2004). However, the present study identified humans as the major hub for the spread of *S. aureus* to livestock, suggesting a role of human activities, such as domestication of animals, in subsequent opportunities for cross-species transmission. In fact, an increasing number of infectious diseases transmitting from humans into animals have been reported, such as influenza A virus, *Cryptosporidium parvum*, and *Ascaris lumbricoides*, and population and evolutionary genomic analyses have the capacity to elucidate the factors underlying these reverse zoonosis (Carroll *et al.*, 2014; Messenger *et al.*, 2014). Nevertheless, even though humans are the largest reservoir of *S. aureus*, cows may represent the second largest reservoir, with 3-5% of global bovine mastitis being caused by this pathogen. This is clearly sufficient to support multiple transmissions of *S. aureus* back into humans (Sakwinska *et al.*, 2011).

An understanding of the molecular mechanisms underlying host-specificity is important to inform on the potential for specific lineages to cross the species-barrier and infect new host-species. In the current study, the putative host-adaptive genes

identified provide new avenues for investigating the mechanisms of bacterial host-adaptation, and these genes may also represent novel therapeutic targets for controlling future human and animal infections. Previous studies using comparative genomic analyses of pathogens from multiple host-species revealed various genetic mechanisms mediating host-adaptation (Foley *et al.*, 2013; Sheppard *et al.*, 2013; Viana *et al.*, 2015; Almeida *et al.*, 2016). In the current work, the identification of prophages and pathogenicity islands is consistent with the role played by the exchange of MGE through lateral gene transfer in *S. aureus* host-adaptation (Malachowa and Deleo, 2010).

GWAS on single SNPs did not reveal any associations with humans or animal groups, suggesting that such small changes in the core genome probably do not mediate host-tropism at the entire species level. Although Viana and colleagues identified that a single nucleotide mutation was required to convert a human-specific *S. aureus* strain into one that could infect rabbits (Viana *et al.*, 2015), that study was restricted to a specific clonal complex and a single species, compared to the wider groups investigated in this work. Nevertheless, considering that adaptation to a novel host-species is a continuous evolutionary process (Rohmer *et al.*, 2011), SNPs in the core genome might mediate adaptation in the long-term. In the present study, positive selection analyses identified genes and functional categories undergoing selection involved in host-adaptation, suggesting that infecting a new host over thousands of generations can further optimize the metabolism of pathogens through mutations in several genes.

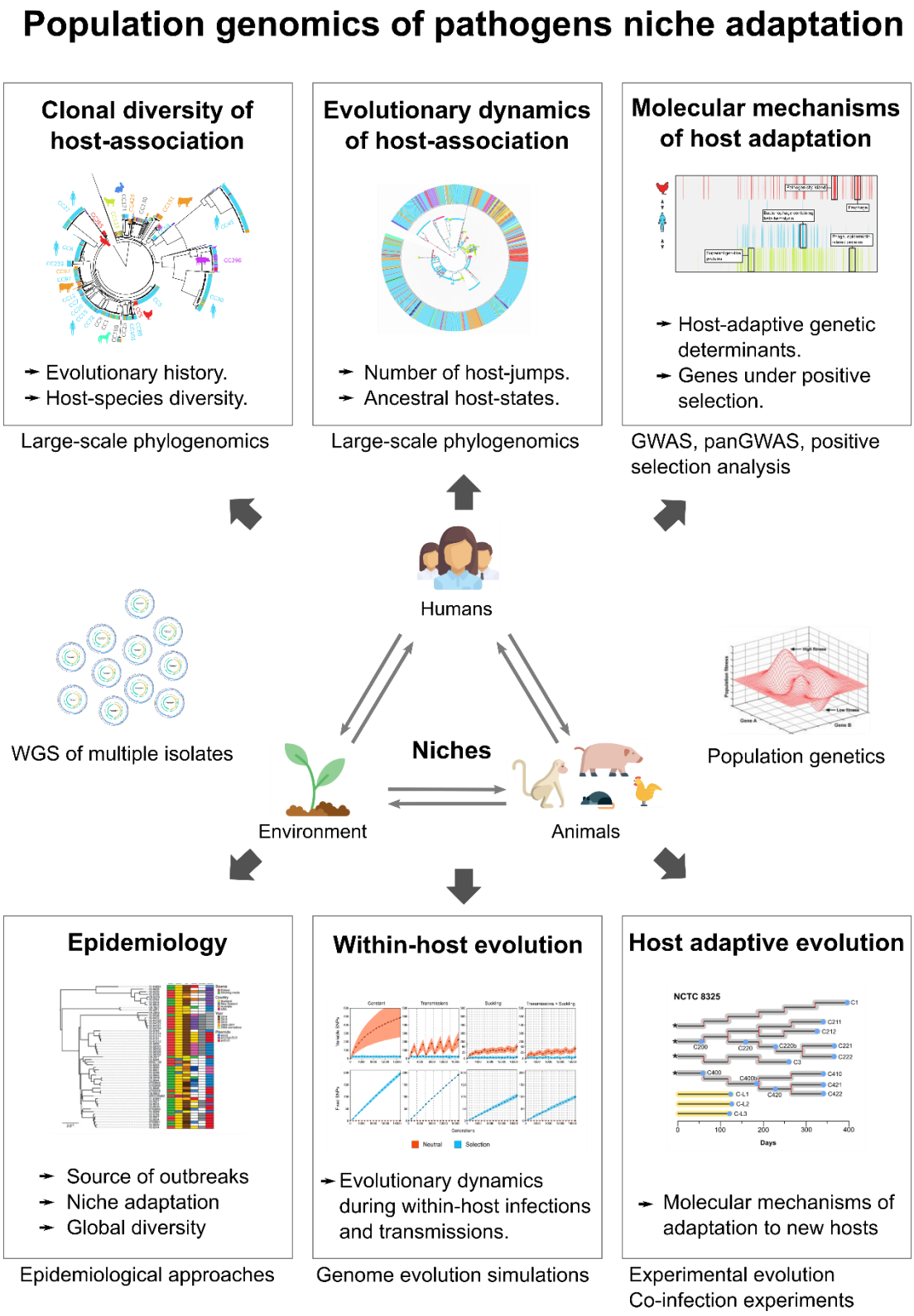


Figure 5.1. Population genomic analyses to investigate bacterial pathogen niche adaptation. The figure includes examples used in this work.

Comparative genomics and phylogenetic analyses of *S. aureus* isolates from multiple species have provided a long-term overview of the evolution and genetic basis of host-specificity (Lowder *et al.*, 2009; Viana *et al.*, 2010; Uhlemann *et al.*, 2012). In this work, a better comprehension of the adaptive process to a new host-species was gained by investigating the evolutionary processes occurring during the early events of a host-switch. Following a transmission from humans to sheep, *S. aureus* had to colonize the new host, overcome the immune system defences and produce effectors that promoted the invasion of the host tissues. During these initial stages, *S. aureus* underwent a series of genomic changes that increased its fitness in the new niche. Contrary to the key role played by MGE in host adaptation during longer evolutionary scales, short-term adaptation was mediated by small genomic changes, a finding that could have important implications for targeting determinants in vaccine development programs.

Most of the mutations identified in this work were in genes encoding proteins involved in virulence, regulation, transport and metabolism, which is consistent with other studies examining the within-host adaptation of pathogens (Lieberman *et al.*, 2011; Marvig *et al.*, 2015; Price *et al.*, 2013). In addition, many genes with mutations encoded secreted products or proteins related with the bacterial surface, which have been shown to be under diversifying selection, favouring the generation of new variants that may promote evasion by the host immune system (Kennemann *et al.*, 2011; Marvig *et al.*, 2015; Price *et al.*, 2013). In addition, the spatial heterogeneity provided by the microenvironments within the mammary gland may promote the genomic diversification of *S. aureus* populations, which is consistent with other studies

that reported populations diverging into distinct sub-lineages within the same or different hosts (Markussen *et al.*, 2014; Zdziarski *et al.*, 2010).

Pathogenic bacteria infecting individuals of a new species represent the first step of a potential successful host-jump. Many infections transmitted from one host-species to another are acute in nature and are rapidly cleared, but pathogens able to mediate onward transmissions to other individuals may expand to become an epidemic clone in the new host-species (Woolhouse *et al*, 2005; Wolfe *et al.*, 2007). Experimental evolution replicating the transmission of *S. aureus* between animals permitted us to investigate the impact of such transmissions on the genetic diversity of the pathogen population, suggesting that continuous bottlenecks limit the fixation of mutations by purging the accumulated diversity. This is relevant for understanding the within-host evolution of pathogens in other environments with complex dynamics, such as the urinary tract (Cole *et al.*, 2014), especially considering most studies have been performed in static environments such as the lungs. Body systems that accumulate and eject media colonized by bacteria impose strong genetic drift in pathogen populations, leading to accumulation of mutations in an effectively neutral fashion (Lynch *et al.*, 2016). Theoretically, this would allow deleterious and neutral mutations to fixate in high proportions (Muller's Ratchet process), but the current work demonstrated that in the face of regular bottlenecks the fitness gain of beneficial mutations can be high enough to overcome genetic drift and sweep through the population. On the contrary, other studies investigating the impact of bottlenecks on pathogen populations have shown similar results to theoretical predictions, with the genetic diversity being stochastically reduced (Kono *et al.*, 2016). However, the inoculum and founding

population sizes of those studies were much smaller than the 100 cfus used in our experiments. This suggests that very tight bottlenecks lead to the Muller's Ratchet process to drive the loss of beneficial mutations, while wider bottleneck allow the fixation of beneficial mutations. In addition, the findings of our work highlight the importance of validating theoretical predictions made by population genomic models using *in vivo* systems. To date, due to logistical difficulties and high costs of within-host experimental evolution experiments, very few studies have used these approaches to investigate the adaptation of pathogens to new species, and these have been performed using plants and the pathogens *Pseudomonas syringae* and *Ralstonia solanacearum* (Guidot *et al.*, 2014; Meaden and Koskella, 2017). However, the evolutionary dynamics of pathogens during the adaptation process to plants and animals are likely very distinct due to major anatomical and physiological differences between these types of organisms.

Finally, the environment also represents an enormous reservoir for bacterial species, and many emerging pathogens are opportunistic bacteria with a versatile life-style. Adaptation to the environment and human associated niches is a major challenge for pathogens, and usually involves intermediate host organisms that allow a pre-adaptation to humans, such as amoebas in the case of *Legionella* species (Aujoulat *et al.*, 2012). In the current study, genome reduction was observed in *L. longbeachae* isolates recovered from humans, which is consistent with other studies suggesting gene loss is a main force for the emergence of pathogens from the environment (Merhej *et al.*, 2009).

In addition, the sequencing of several isolates from both the environment and single individuals revealed the open pangenome of the environmental isolates compared to the closed pangenome of the human-associated ones. This would not have been possible if WGS had been applied to a single isolate per individual, as traditionally performed, and such low level of sampling is not likely to provide enough resolution to infer the rates and routes of transmission in epidemiological investigations (Eyre *et al.*, 2013; Worby *et al.*, 2014). In the current study, the sequencing of single isolates from the compost samples impeded the identification of the source of the infections (Bacigalupe *et al.*, 2017). However, population genomic analyses identified the evolutionary relationships between the Scottish *L. longbeachae* isolates with strains from other countries around the world, which shows the value of WGS for investigation of the global spread of bacterial pathogens. In addition, the extensive levels of recombination and lateral gene transfer between *L. longbeachae* isolates and other species highlights the importance of accounting for these events in epidemiological investigations. These evolutionary mechanisms may impact enormously on the population structure of pathogens, hindering the identification of the source of the infections and the characterization of the transmission chains (Tibayrenc and Ayala, 2012).

This study demonstrates the application of WGS and population genomic approaches to investigate the evolutionary history of bacterial pathogens, the molecular mechanisms of host adaptation and the epidemiology of infectious diseases in the context of their global diversity. These data provided new insights into the ecology and evolution of infectious diseases, which is relevant to other major bacterial

pathogens able to spread between different niches. However, this work also presents some limitations that potentially affect our findings. First, the large collection of *S. aureus* was considerable biased towards isolates from human origin, while other host-types were underrepresented. Although it is often assumed that genomic determinants found in divergent lineages associated with specific hosts and absent in their ancestors represent host-adaptive factors (Murray *et al.*, 2017), the aforementioned bias might have led to identification of genetic changes associated with drift rather than with adaptation. Despite the bias in our dataset, our results are probably not very affected as most human clones were highly clonal. Secondly, host-adaptation can be also driven by the combined effect of several genes, as recently identified in the adaptation of influenza virus to humans and various animals (Khaliq *et al.*, 2016). The effect of combinations of genes or the epistatic interactions of SNPs remains to be explored, which could reveal more complex pathways for host-adaptation than the investigated so far. In addition, genetic features of the host can influence host-pathogen interactions, determining the capacity of bacteria to colonize and persist in the new host-species. It has been reported that different human ethnicities have different rates of *S. aureus* infection (Messina *et al.*, 2016). Similarly, different strains of sheep and cattle present distinct susceptibility to *S. aureus* infection, sepsis, and death (Bonnefont *et al*., 2012; Griesbeck-Zilch *et al*., 2009). Thus, investigating the impact of both *S. aureus* and host genetic variation could improve the understanding of the molecular mechanisms of host-adaptation. Furthermore, potential genomic differences between isolates from host-restricted and generalist sub-lineages were not analysed, which could result in the identification of traits associated with the ability for a multi-host ecology. This aspect is also relevant in our experimental evolution work, since we only

used strains from the generalist CC5 and CC8 clonal complexes. We did not observe exchange of MGE encoding host-specificity determinants, usually required by specialist pathogens to cross the host-species boundaries, but probably not needed by generalists for switching between host-types. Thus, future studies should examine the evolution of host-restricted clones during the adaptation process to a new host-species. Furthermore, although we tried to reproduce natural conditions during the passage experiments, the inoculum size used to colonize new sheep could be different to the actual amount of bacteria transmitted from sheep-to-sheep by natural means. The size of this bottleneck was important for our population genetic inferences, which were also limited by the number of isolates we sampled, sequenced and analyzed. Additionally, the replication time estimated in the laboratory conditions might also differ to that within sheep, considering the immune system imposes pressures that limit bacterial grow. Moreover, the selection coefficients and other parameters used in the *in silico* evolution experiments coud also deviate from the real ones, but we thoroughly revised the bibliography to use the most precise ones. Finally, as we concluded in our genomic epidemiology study, the single-isolate sampling strategy of *L. longbeachae* impeded us to detect the infecting genotype and therefore the source of the infections. Furthermore, the number of isolates from other countries other than Scotland, England and New Zealand was very small. We were constrained by the availalble genomic sequences and our study provided a snapshot of the known global *L. longbeachae* diversity in 2014.

In the coming years, as sequencing technologies continue to improve and routine WGS is implemented in research centres and health service facilities across the world,

hundreds of thousands and millions of isolates of different bacterial species from different environments, host-groups and multiple countries will be available in public databases. Subsequently, future investigations including very large numbers of isolates from additional host-types and other continents will improve our understanding of the evolutionary dynamics of bacteria in the context of their host-associations. In addition, novel sequencing technologies will further improve the quality of bacterial genomes, producing highly accurate assemblies that only contain the full chromosomes and plasmids. Sequencing of single cells rather than populations obtained from colonies will bring population genomic studies to an unprecedented extent, providing an excellent framework for applying big data population genomics and epidemiological strategies. These advances will allow the scientific community to better understand the evolutionary dynamics of pathogens, their reservoirs and transmission routes, the phenotype-to-genotype relationships and human predisposition to disease, which will ultimately permit the development of new diagnostics, vaccines and therapies for treatment of infectious diseases.

References

Aanensen, D.M., Feil, E.J., Holden, M.T.G., Dordel, J., Yeats, C.A., Fedosejev, A., Goater, R., Castillo-Ramírez, S., Corander, J., Colijn, C., Chlebowicz, M.A., Schouls, L., Heck, M., Pluister, G., Ruimy, R., Kahlmeter, G., Åhman, J., Matuschek, E., Friedrich, A.W., Parkhill, J., Bentley, S.D., Spratt, B.G., Grundmann, H., European SRL Working Group, E.S.W., Mittermayer, H., Krziwanek, K., Stumvoll, S., Koller, W., Denis, O., Struelens, M., Nashev, D., Budimir, A., Kalenic, S., Pieridou-Bagatzouni, D., Jakubu, V., Zemlickova, H., Westh, H., Larsen, A.R., Skov, R., Laurent, F., Ettienne, J., Strommenger, B., Witte, W., Vourli, S., Vatopoulos, A., Vainio, A., Vuopio-Varkila, J., Fuzi, M., Ungvári, E., Murchan, S., Rossney, A., Miklasevics, E., Balode, A., Haraldsson, G., Kristinsson, K.G., Monaco, M., Pantosti, A., Borg, M., Santen-Verheuvel, M. van, Huijsdens, X., Marstein, L., Jacobsen, T., Simonsen, G.S., Airesde-Sousa, M., Lencastre, H. de, Luczak-Kadlubowska, A., Hryniewicz, W., Straut, M., Codita, I., Perez-Vazquez, M., Iglesias, J.O., Spik, V.C., Mueller-Premru, M., Haeggman, S., Olsson-Liljequist, B., Ellington, M., Kearns, A., 2016. Whole-Genome Sequencing for Routine Pathogen Surveillance in Public Health: a Population Snapshot of Invasive *Staphylococcus aureus* in Europe. *MBio* 7, e00444–16.

Aarestrup, F.M., Brown, E.W., Detter, C., Gerner-Smidt, P., Gilmour, M.W., Harmsen, D., Hendriksen, R.S., Hewson, R., Heymann, D.L., Johansson, K., Ijaz, K., Keim, P.S., Koopmans, M., Kroneman, A., Lo Fo Wong, D., Lund, O., Palm, D., Sawanpanyalert, P., Sobel, J., Schlundt, J.J., Wong, D.L.F., Lund, O., Palm, D., Sawanpanyalert, P., Sobel, J., Schlundt, J.J., 2012. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerg. Infect. Di*s. 18, e1–e1.

Achtman, M., 2008. Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. *Annu. Rev. Microbiol*. 62, 53–70.

Acton, D.S., Tempelmans Plat-Sinnige, M.J., Van Wamel, W., De Groot, N., Van Belkum, A., 2009. Intestinal carriage of *Staphylococcus aureus:* How does its frequency compare with that of nasal carriage and what is its clinical impact? *Eur. J. Clin. Microbiol. Infect. Dis.* 28, 115–127.

Alam, M.T., Petit, R.A., Crispell, E.K., Thornton, T.A., Conneely, K.N., Jiang, Y., Satola, S.W., Read, T.D., 2014. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biol. Evol*. 6, 1174–1185.

Alibayov, B., Baba-Moussa, L., Sina, H., Zdeňková, K., Demnerová, K., 2014. *Staphylococcus aureus* mobile genetic elements. *Mol. Biol. Rep*. 41, 5005–5018.

Almeida, A., Alves-Barroco, C., Sauvage, E., Bexiga, R., Albuquerque, P., Tavares, F., Santos-Sanches, I., Glaser, P., 2016. Persistence of a dominant bovine lineage of group B *Streptococcus* reveals genomic signatures of host adaptation. *Environ. Microbiol*. 18, 4216–4229.

Alves, J.M.P., Serrano, M.G., Da Silva, F.M., Voegtly, L.J., Matveyev, A. V., Teixeira, M.M.G., Camargo, E.P., Buck, G.A., 2013. Genome evolution and phylogenomic analysis of *Candidatus kinetoplastibacterium*, the betaproteobacterial endosymbionts of *Strigomonas* and *Angomonas*. *Genome Biol. Evol*. 5, 338–350.

Amodeo, M.R., Murdoch, D.R., Pithie, A.D., 2010. Legionnaires' disease caused by *Legionella longbeachae* and *Legionella pneumophila*: Comparison of clinical features,

host-related risk factors, and outcomes. *Clin. Microbiol. Infect.* 16, 1405–1407.

Andam, C.P., Challagundla, L., Azarian, T., Hanage, W.P., Robinson, D.A., 2017. 3 – Population Structure of Pathogenic Bacteria. In: Genetics and Evolution of Infectious Diseases. pp. 51–70.

Arenas, M., Posada, D., 2014. Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. *Mol. Biol. Evol*. 31, 1295–1301.

Argimón, S., Abudahab, K., Goater, R.J.E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E.J., Holden, M.T.G., Yeats, C.A., Grundmann, H., Spratt, B.G., Aanensen, D.M., 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* 2, e000093.

Arias, C.A., Murray, B.E., 2012. The rise of the *Enterococcus*: beyond vancomycin resistance. *Nat. Rev. Microbiol*. 10, 266–78.

Assefa, S., Lim, C., Preston, M.D., Duffy, C.W., Nair, M.B., Adroub, S. a, Kadir, K. a, Goldberg, J.M., Neafsey, D.E., Divis, P., Clark, T.G., Duraisingh, M.T., Conway, D.J., Pain, A., Singh, B., 2015. Population genomic structure and adaptation in the zoonotic malaria parasite *Plasmodium knowlesi*. *Proc. Natl. Acad. Sci*. U. S. A. 112, 13027–13032.

Aujoulat, F., Roger, F., Bourdier, A., Lotthé, A., Lamy, B., Marchandin, H., Jumas-Bilak, E. 2012. From environment to man: Genome evolution and adaptation of human opportunistic bacterial pathogens. *Genes*, 3(2), 191–232.

Azevedo, M., Sousa, A., De Sousa, J.M., Thompson, J.A., Proença, J.T., Gordo, I., 2016. Trade-offs of *Escherichia coli* adaptation to an intracellular lifestyle in macrophages. *PLoS One* 11, e0146123.

Baba, T., Takeuchi, F., Kuroda, M., Yuzawa, H., Aoki, K.I., Oguchi, A., Nagai, Y., Iwama, N., Asano, K., Naimi, T., Kuroda, H., Cui, L., Yamamoto, K., Hiramatsu, K., 2002. Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* 359, 1819–1827.

Baele, G., Suchard, M.A., Rambaut, A., Lemey, P., 2017. Emerging concepts of data integration in pathogen phylodynamics. *Syst. Biol*. 66, e47–e65.

Baker, S., Hanage, W.P., Holt, K.E., 2010. Navigating the future of bacterial molecular epidemiology. *Curr. Opin. Microbiol*. 13, 640–645.

Balloux, F., 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. *J. Hered.* 92, 301–302.

Bäumler, A., & Fang, F. C. 2013. Host specificity of bacterial pathogens. *Cold Spring Harbor Perspectives in Medicine*, 3(12), a010041.

Bank, C., Ewing, G.B., Ferrer-Admettla, A., Foll, M., Jensen, J.D., 2014. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends Genet*. 30, 540–546.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a., Pevzner, P. a., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477.

Barber, M., Rozwadowska-Dowzenko, M., 1948. Infection By Penicillin-Resistant

Staphylococci. *Lancet* 252, 641–644.

Barrick, J., Yu, D., Yoon, S., Jeong, H., Oh, T., Schneider, D., Lenski, R., Kim, J., 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461, 1243–1247.

Barrick, J.E., Lenski, R.E., 2013. Genome dynamics during experimental evolution. *Nat. Rev. Genet*. 14, 827–839.

Barroso-Batista, J., Sousa, A., Lourenço, M., Bergman, M.L., Sobral, D., Demengeot, J., Xavier, K.B., Gordo, I., 2014. The First Steps of Adaptation of *Escherichia coli* to the Gut Are Dominated by Soft Sweeps. PLoS Genet. 10.

Barthe, C., Hermans, K., Haesebrouck, F., 2009. Main pathologies associated with *Staphylococcus aureus* infections in rabbits: A review. *World Rabbit Sci.* 17, 115–125.

Barton, M.D., 2014. Impact of antibiotic use in the swine industry. *Curr. Opin. Microbiol*. 19, 9–15.

Bayles, K.W., Iandolo, J.J., 1989. Genetic and molecular analyses of the gene encoding staphylococcal enterotoxin D. *J. Bacteriol*. 171, 4799–4806.

Bazin, H., 2003. A brief history of the prevention of infectious diseases by immunisations. *Comp. Immunol. Microbiol. Infect. Dis*.

Bentley, S.D., Parkhill, J., 2015. Genomic perspectives on the evolution and spread of bacterial pathogens. *Proc. R. Soc. B Biol. Sci.* 282, 20150488.

Berg, T., Firth, N., Apisiridej, S., Hettiaratchi, A., Leelaporn, A., Skurray, R.A., 1998. Complete nucleotide sequence of pSK41: evolution of staphylococcal conjugative multiresistance plasmids. *J. Bacteriol*. 180, 4350–9.

Bergonier, D., de Crémoux, R., Rupp, R., Lagriffoul, G., Berthelot, X., 2003. Mastitis of dairy small ruminants. *Vet. Res.* 34, 689–716.

Bibb, W.F., Sorg, R.J., Thomason, B.M., Hicklin, M.D., Steigerwalt, A.G., Brenner, D.J., Wulf, M.R., 1981. Recognition of a second serogroup of *Legionella longbeachae*. *J. Clin. Microbiol.* 14, 674–677.

Biek, R., Pybus, O.G., Lloyd-Smith, J.O., Didelot, X., 2015. Measurably evolving pathogens in the genomic era. Trends Ecol. Evol. 30, 306–313.

Bierowiec, K., Płoneczka-Janeczko, K., Rypuła, K., 2016. Is the colonisation of *Staphylococcus aureus* in pets associated with their close contact with owners? *PLoS One* 11, e0156052.

Binder, S., Levitt, A.M., Sacks, J.J., Hughes, J.M., 1999. Emerging Infectious Diseases : Public Health Issues for the 21st Century. *Science* (80). 284, 1311–1314.

Black, W.C., Baer, C.F., Antolin, M.F., DuTeau, N.M., 2001. Population genomics: genome-wide sampling of insect populations. *Annu. Rev. Entomol*. 46, 441–469.

Bloom, D.E., Black, S., Rappuoli, R., 2017. Emerging infectious diseases: A proactive approach. *Proc. Natl. Acad. Sci*. 114, 4055–4059.

Bloomfield, S.F., Stanwell-Smith, R., Crevel, R.W.R., Pickup, J., 2006. Too clean, or not too clean: The Hygiene Hypothesis and home hygiene. *Clin. Exp. Allergy* 36, 402–425.

Boetzer, M., Henkel, C. V., Jansen, H.J., Butler, D., Pirovano, W., 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.

Bonnefont, C.M., Rainard, P., Cunha, P., Gilbert, F.B., Toufeer, M., Aurel, M.R., Rupp, R., Foucras, G. 2012. Genetic susceptibility to *S. aureus* mastitis in sheep: differential expression of mammary epithelial cells in response to live bacteria or supernatant. *Physiological Genomics*; 44(7):403–16.

Boundy, S., Safo, M.K., Wang, L., Musayev, F.N., O'Farrell, H.C., Rife, J.P., Archer, G.L., 2013. Characterization of the *Staphylococcus aureus* rRNA methyltransferase encoded by orfx, the gene containing the staphylococcal chromosome cassette mec (SCCmec) insertion site. *J. Biol*. Chem. 288, 132–140.

Bowman, T., Jacobson, E., 1980. Cloacal Flora of Clinically Healthy Psittacine Birds. *J. Zoo Wildl. Med.* 11, 81–85.

Boyle-Vavra, S., Daum, R.S., 2007. Community-acquired methicillin-resistant *Staphylococcus aureus:* the role of Panton–Valentine leukocidin. *Lab. Investig.* 87, 3–9.

Brachman, P.S., 2003. Infectious diseases - Past, present, and future. *Int. J. Epidemiol.* 32, 684–686.

Brown, T., Didelot, X., Wilson, D.J., De Maio, N., 2016. SimBac: simulation of whole bacterial genomes with homologous recombination. *Microb. Genomics* 2, 1–6.

Bruen, T.C., Philippe, H., Bryant, D., 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172, 2665–2681.

Brussow, H., Canchaya, C., Hardt, W.D., 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* 68, 560–602.

Bryant, J.M., Harris, S.R., Parkhill, J., Dawson, R., Diacon, A.H., van Helden, P., Pym, A., Mahayiddin, A.A., Chuchottaworn, C., Sanne, I.M., Louw, C., Boeree, M.J., Hoelscher, M., McHugh, T.D., Bateson, A.L.C., Hunt, R.D., Mwaigwisya, S., Wright, L., Gillespie, S.H., Bentley, S.D., 2013. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *Lancet. Respir. Med*. 1, 786–92.

Brynildsrud, O., Bohlin, J., Scheffer, L., Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17, 238.

Buchholz, U., Bernard, H., Werber, D., Böhmer, M.M., Remschmidt, C., Wilking, H., Deleré, Y., an der Heiden, M., Adlhoch, C., Dreesman, J., Ehlers, J., Ethelberg, S., Faber, M., Frank, C., Fricke, G., Greiner, M., Höhle, M., Ivarsson, S., Jark, U., Kirchner, M., Koch, J., Krause, G., Luber, P., Rosner, B., Stark, K., Kühne, M., 2011. German Outbreak of *Escherichia coli* O104:H4 Associated with Sprouts. *N. Engl. J. Med.* 365, 1763–1770.

Buchrieser, C., Hilbi, H., 2013. *Legionella* : methods and protocols. Humana Press.

Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J. a, Pupko, T., Shuman, H. a, Segal, G., 2016. Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat. Genet.* 48, 167–175.

Byrne, L., Fisher, I., Peters, T., Mather, A., Thomson, N., Rosner, B., Bernard, H., McKeown, P., Cormican, M., Cowden, J., Aiyedun, V., Lane, C., International Outbreak Control Team, 2014. A multi-country outbreak of *Salmonella* Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012. *Euro Surveill.* 19, 6–13.

Cameron, R.L., Pollock, K.G.J., Lindsay, D.S.J., Anderson, E., 2016. Comparison of *Legionella longbeachae* and *Legionella pneumophila* cases in Scotland; implications for diagnosis, treatment and public health response. *J. Med. Microbiol*. 65, 142–146.

Cameron, S., Roder, D., Walker, C., Feldheim, J., 1991. Epidemiological characteristics of *Legionella* infection in South Australia: implications for disease control. *Aust. N. Z. J. Med.* 21, 65–70.

Campos, P.R.A., Wahl, L.M., 2010. The adaptation rate of asexuals: Deleterious mutations, clonal interference and population bottlenecks. *Evolution* (N. Y). 64, 1973–1983.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., Brüssow, H., 2003. Prophage genomics. Microbiol. *Mol. Biol. Rev.* 67, 238–276.

Carroll, S.P., Jørgensen, P.S., Kinnison, M.T., Bergstrom, C.T., Denison, R.F., Gluckman, P., Smith, T.B., Strauss, S.Y., Tabashnik, B.E., 2014. Applying evolutionary biology to address global challenges. *Science* 346, 1245993.

Carvajal-Rodriguez, A., 2008. GENOMEPOP: A program to simulate genomes in populations. *BMC Bioinformatics* 9, 223.

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S.R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R.D., Brown, T., Drobniewski, F., 2014. Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet*. 46, 279–286.

Casewell, M.W., Hill, R.L., 1986. The carrier state: methicillin-resistant *Staphylococcus aureus*. *J. Antimicrob. Chemother*. 18 Suppl A, 1–12.

Cazalet, C., Gomez-Valero, L., Rusniok, C., Lomma, M., Dervins-Ravault, D., Newton, H.J., Sansom, F.M., Jarraud, S., Zidane, N., Ma, L., Bouchier, C., Etienne, J., Hartland, E.L., Buchrieser, C., 2010. Analysis of the *Legionella longbeachae* genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. *PLoS Genet*. 6, e1000851.

Cazalet, C., Rusniok, C., Brüggemann, H., Zidane, N., Magnier, A., Ma, L., Tichit, M., Jarraud, S., Bouchier, C., Vandenesch, F., Kunst, F., Etienne, J., Glaser, P., Buchrieser, C., 2004. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. *Nat. Genet*. 36, 1165–1173.

Centers for Disease Control and Prevention (CDC). Antibiotic resistance threats in the United States, 2013. Atlanta: Available from: http://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf

Chaguza, C., Cornick, J.E., Everett, D.B., 2015. Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. *Comput. Struct. Biotechnol. J*. 13, 241–247.

Chambers, H.F., Deleo, F.R., 2009. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. *Nat. Rev. Microbiol.* 7, 629–41.

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–303.

Charusanti, P., Conrad, T.M., Knight, E.M., Venkataraman, K., Fong, N.L., Xie, B., Gao, Y., Palsson, B., 2010. Genetic basis of growth adaptation of *Escherichia coli* after deletion of pgi, a major metabolic gene. *PLoS Genet.* 6, e1001186.

Chen, P.E., Shapiro, B.J., 2015. The advent of genome-wide association studies for bacteria. Curr. Opin. Microbiol. 25, 17–24.

Chen, Y.-H., Anderson, M., Hendrickx, A.P., Missiakas, D., 2012. Characterization of EssB, a protein required for secretion of ESAT-6 like proteins in *Staphylococcus aureus*. *BMC Microbiol.* 12, 219.

Chewapreecha, C., Marttinen, P., Croucher, N.J., Salter, S.J., Harris, S.R., Mather, A.E.,

Hanage, W.P., Goldblatt, D., Nosten, F.H., Turner, C., Turner, P., Bentley, S.D., Parkhill, J., 2014. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet.* 10, e1004547.

Chua, K.Y.L., Seemann, T., Harrison, P.F., Monagle, S., Korman, T.M., Johnson, P.D.R., Coombs, G.W., Howden, B.O., Davies, J.K., Howden, B.P., Stinear, T.P., 2011. The dominant Australian community-acquired methicillin-resistant *Staphylococcus aureus* clone ST93-IV [2B] is highly virulent and genetically distinct. *PLoS One* 6, e25887.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., Tiedje, J.M., 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, 633–642.

Cole, S.J., Records, A.R., Orr, M.W., Linden, S.B., Lee, V.T., 2014. Catheter-associated urinary tract infection by *Pseudomonas aeruginosa* is mediated by exopolysaccharide-independent biofilms. *Infect. Immun.* 82, 2048–2058.

Coleman, D.C., Sullivan, D.J., Russell, R.J., Arbuthnott, J.P., Carey, B.F., Pomeroy, H.M., 1989. *Staphylococcus aureus* bacteriophages mediating the simultaneous lysogenic conversion of beta-lysin, staphylokinase and enterotoxin A: molecular mechanism of triple conversion. *J. Gen. Microbiol.* 135, 1679–1697.

Collins, C., Didelot, X., 2017. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *bioRxiv* 1–23.

Contreras-Moreira, B., Vinuesa, P., 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* 79, 7696–7701.

Corander, J., Connor, T.R., O'Dwyer, C.A., Kroll, J.S., Hanage, W.P., 2012. Population structure in the *Neisseria*, and the biological significance of fuzzy species. *J. R. Soc. Interface* 9, 1208–15.

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.

Corrigan, R.M., Miajlovic, H., Foster, T.J., 2009. Surface proteins that promote adherence of *Staphylococcus aureus* to human desquamated nasal epithelial cells. *BMC Microbiol.* 9, 22.

Coscollá, M., Comas, I., González-Candelas, F., 2011. Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila*. *Mol. Biol. Evol.* 28, 985–1001.

Cramp, G.J., Harte, D., Douglas, N.M., Graham, F., Schousboe, M., Sykes, K., 2010. An outbreak of Pontiac fever due to *Legionella longbeachae* serogroup 2 found in potting mix in a horticultural nursery in New Zealand. *Epidemiol. Infect.* 138, 15–20.

Croucher, N.J., Didelot, X., 2015. The application of genomics to tracing bacterial pathogen transmission. *Curr. Opin. Microbiol*. 23, 62–67.

Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P., Lipsitch, M., 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* 45, 656–63.

Croxatto, A., Prod'hom, G., Greub, G., 2012. Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiol. Rev.* 36, 380–407.

Currie, S.L., Beattie, T.K., Knapp, C.W., Lindsay, D.S.J., 2014. *Legionella spp.* in UK

composts-a potential public health issue? *Clin. Microbiol. Infect.* 20, 1–6.

Cybis, G.B., Sinsheimer, J.S., Bedford, T., Mather, A.E., Lemey, P., Suchard, M.A., 2015. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* 9, 969–991.

Dallman, T., Inns, T., Jombart, T., Ashton, P., Loman, N., Chatt, C., Messelhaeusser, U., Rabsch, W., Simon, S., Nikisins, S., Bernard, H., Hello, S. le, Da-Silva, N.J., Kornschober, C., Mossong, J., Grant, K., Clear, P., 2016. Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. *Microb. Genomics* 2, 1–8.

Damkiaer, S., Yang, L., Molin, S., Jelsbak, L., 2013. Evolutionary remodeling of global regulatory networks during long-term bacterial adaptation to human hosts. *Proc Natl Acad Sc*i U S A 110, 7766–7771.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

Day, M.J., Breitschwerdt, E., Cleaveland, S., Karkare, U., Khanna, C., Kirpensteijn, J., Kuiken, T., Lappin, M.R., McQuiston, J., Mumford, E., Myers, T., Palatnik-de-Sousa, C.B., Rubin, C., Takashima, G., Thiermann, A., 2012. Surveillance of zoonotic infectious disease transmitted by small companion animals. *Emerg. Infect. Dis.* 18, e1.

De Been, M., Van Schaik, W., Cheng, L., Corander, J., Willems, R.J., 2013. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium. Genome Biol.* Evol. 5, 1524–1535.

De Maio, N., Wilson, D.J., 2017. The bacterial sequential markov coalescent. *Genetics* 206, 333–343.

De Maio, N., Wu, C.H., O'Reilly, K.M., Wilson, D., 2015. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet.* 11, e1005421.

De Meeûs, T., Prugnolle, F., Agnew, P., 2007. Asexual reproduction: Genetics and evolutionary aspects. *Cell. Mol. Life Sci.* 64, 1355–1372.

Delport, W., 2006. COALFACE: A graphical user interface program for the simulation of coalescence. *Mol. Ecol. Notes* 6, 281–284.

Den Boer, J.W., Yzerman, E.P.F., Jansen, R., Bruin, J.P., Verhoef, L.P.B., Neve, G., Van Der Zwaluw, K., 2007. Legionnaires' disease and gardening. *Clin. Microbiol. Infect.* 13, 88–91.

Deng X, den Bakker HC, H.R., 2016. Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci Technol* 7, 1–22.

Deringer, J.R., Ely, R.J., Monday, S.R., Stauffacher, C. V, Bohach, G.A., 1997. Vbeta-dependent stimulation of bovine and human T cells by host-specific staphylococcal enterotoxins. *Infect. Immun.* 65, 4048–54.

Devriese, L.A., Devos, A.H., Beumer, J., Maes, R., 1972. Characterisation of Staphylococci Isolated from Poultry. *Poult. Sci.* 51, 389–397.

Didelot, X., Bowden, R., Street, T., Golubchik, T., Spencer, C., McVean, G., Sangal, V., Anjum, M.F., Achtman, M., Falush, D., Donnelly, P., 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet.* 7, e1002191.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E.A., Crook, D.W., 2012. Transforming clinical

microbiology with bacterial genome sequencing. *Nat Rev Genet* 13, 601–612.

Didelot, X., Nell, S., Yang, I., Woltemate, S., van der Merwe, S., Suerbaum, S., 2013. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proc. Natl. Acad. Sci.* U. S. A. 110, 13880–5.

Didelot, X., Walker, A.S., Peto, T.E., Crook, D.W., Wilson, D.J., Sarah Walker, A., Peto, T.E., Crook, D.W., Wilson, D.J., 2016. Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* 14, 150–162.

Didelot, X., Wilson, D.J., 2015. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* 11, e1004041.

Diekema, D.J., Pfaller, M. a, Jones, R.N., Doern, G. V, Winokur, P.L., Gales, a C., Sader, H.S., Kugler, K., Beach, M., 1999. Survey of Infections Due to Staphylococcus Species: Frequency of Occurrence and Antimicrobial Susceptibility of Isolates Collected in the United States, Canada, Latin America, Europe, and the Western Pacific Region for the SENTRY Antimicrobial Surveillanc. *Clin. Infect. Dis.* 29, 595–607.

Dobrindt, U., Hochhut, B., Hentschel, U., Hacker, J., 2004. Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* 2, 414–424.

Donati, C., Hiller, N.L., Tettelin, H., Muzzi, A., Croucher, N.J., Angiuoli, S. V, Oggioni, M., Dunning Hotopp, J.C., Hu, F.Z., Riley, D.R., Covacci, A., Mitchell, T.J., Bentley, S.D., Kilian, M., Ehrlich, G.D., Rappuoli, R., Moxon, E.R., Masignani, V., 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol.* 11, R107.

Dubrac, S., Boneca, I.G., Poupel, O., Msadek, T., 2007. New insights into the WalK/WalR (YycG/YycF) essential signal transduction pathway reveal a major role in controlling cell wall metabolism and biofilm formation in *Staphylococcus aureus*. *J. Bacteriol.* 189, 8257–8269.

Dukic, V.M., Lauderdale, D.S., Wilder, J., Daum, R.S., David, M.Z., 2013. Epidemics of Community-Associated Methicillin-Resistant *Staphylococcus aureus* in the United States: A Meta-Analysis. *PLoS One* 8, e52722.

Dye, C., 2014. After 2015: infectious diseases in a new era of health and development. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 369, 20130426.

Edelstein, P.H., 1982. Comparative study of selective media for isolation of *Legionella pneumophila* from potable water. *J. Clin. Microbiol*. 16, 697–699.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32, 1792–1797.

Eldholm, V., Norheim, G., von der Lippe, B., Kinander, W., Dahle, U.R., Caugant, D.A., Mannsåker, T., Mengshoel, A.T., Dyrhol-Riise, A.M., Balloux, F., 2014. Evolution of extensively drug-resistant *Mycobacterium tuberculosis* from a susceptible ancestor in a single patient. *Genome Biol*. 15, 490.

Elena, S.F., Lenski, R.E., 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet*. 4, 457–69.

Ellington, M.J., Hope, R., Livermore, D.M., Kearns, A.M., Henderson, K., Cookson, B.D., Pearson, A., Johnson, A.P., 2009. Decline of EMRSA-16 amongst methicillin-resistant *Staphylococcus aureus* causing bacteraemias in the UK between 2001 and 2007. *J. Antimicrob. Chemother.* 65, 446–448.

Engering, A., Hogerwerf, L., Slingenbergh, J., 2013. Pathogen-host-environment interplay and

disease emergence. *Emerg. Microbes Infect.* 2, e5.

Enright, M.C., Day, N.P., Davies, C.E., Peacock, S.J., Spratt, B.G., 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* 38, 1008–15.

Ensminger, A.W., 2013. Experimental Evolution of Pathogenesis: "Patient" Research. *PLoS Pathog.* 9, e1003340.

Espadinha, D., Faria, N.A., Miragaia, M., Lito, L.M., Melo-Cristino, J., de Lencastre, H., Network, M.S., 2013. Extensive Dissemination of Methicillin-Resistant *Staphylococcus aureus* (MRSA) between the Hospital and the Community in a Country with a High Prevalence of Nosocomial MRSA. *PLoS One* 8, e59960.

Espinosa-Gongora, C., Moodley, A., Lipinska, U., Broens, E.M., Hermans, K., Butaye, P., Devriese, L.A., Haesebrouck, F., Guardabassi, L., 2014. Phenotypes and Genotypes of Old and Contemporary Porcine Strains Indicate a Temporal Change in the *S. aureus* Population Structure in Pigs. PLoS One 9, e101988.

Everitt, R.G., Didelot, X., Batty, E.M., Miller, R.R., Knox, K., Young, B.C., Bowden, R., Auton, A., Votintseva, A., Larner-Svensson, H., Charlesworth, J., Golubchik, T., Ip, C.L.C., Godwin, H., Fung, R., Peto, T.E.A., Walker, A.S., Crook, D.W., Wilson, D.J., 2014. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. Nat. Commun. 5, 1–9.

Ewing, G., Hermisson, J., 2010. MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26, 2064–2065.

Excoffier, L., Foll, M., 2011. fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics 27, 1332–1334.

Eyre, D.W., Cule, M.L., Griffiths, D., Crook, D.W., Peto, T.E.A., Walker, A.S., Wilson, D.J., Clark, T., Luong, K., Song, Y., Tsai, Y., Boitano, M., Dayal, J., Brooks, S., Schmidt, B., Young, A., Thomas, J., Bouffard, G., Blakesley, R., Mullikin, J., Korlach, J., Henderson, D., Frank, K., Palmore, T., Segre, J., 2013. Detection of Mixed Infection from Bacterial Whole Genome Sequence Data Allows Assessment of Its Role in *Clostridium difficile* Transmission. PLoS Comput. Biol. 9, e1003059.

Eyre, D.W., Cule, M.L., Wilson, D.J., Griffiths, D., Vaughan, A., O'Connor, L., Ip, C.L.C., Golubchik, T., Batty, E.M., Finney, J.M., Wyllie, D.H., Didelot, X., Piazza, P., Bowden, R., Dingle, K.E., Harding, R.M., Crook, D.W., Wilcox, M.H., Peto, T.E. a, Walker, a S., 2013. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N. Engl. J. Med. 369, 1195–205.

Eyre-Walker, a, Keightley, P., 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8, 610–8.

Fallon, R.J., Abraham, W.H., 1983. Experience with heat-killed antigens of *L. longbeachae* serogroups 1 and 2, and *L. jordanis* in the indirect fluorescence antibody test. Zentralbl. Bakteriol. Mikrobiol. Hyg. A. 255, 8–14.

Falush, D., Bowden, R., 2006. Genome-wide association mapping in bacteria? *Trends Microbiol*. 14, 353–355.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D.F., Wilson, D.J., Achtman, M., 2006. Mismatch induced speciation in *Salmonella*: model and data. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361, 2045–53.

Farhat, M.R., Shapiro, B.J., Kieser, K.J., Sultana, R., Jacobson, K.R., Victor, T.C., Warren, R.M., Streicher, E.M., Calver, A., Sloutsky, A., Kaur, D., Posey, J.E., Plikaytis, B., Oggioni, M.R., Gardy, J.L., Johnston, J.C., Rodrigues, M., Tang, P.K.C., Kato-Maeda, M., Borowsky, M.L., Muddukrishna, B., Kreiswirth, B.N., Kurepina, N., Galagan, J., Gagneux, S., Birren, B., Rubin, E.J., Lander, E.S., Sabeti, P.C., Murray, M., 2013. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis. Nat. Genet.* 45, 1183–9.

Feasey, N.A., Dougan, G., Kingsley, R.A., Heyderman, R.S., Gordon, M.A., 2012. Invasive non-typhoidal *Salmonella* disease: An emerging and neglected tropical disease in Africa. *Lancet* 379, 2489–2499.

Feil, E.J., Cooper, J.E., Grundmann, H., Robinson, D.A., Enright, M.C., Berendt, T., Peacock, S.J., Smith, J.M., Murphy, M., Spratt, B.G., Moore, C.E., Day, N.P.J., 2003. How clonal is *Staphylococcus aureus*? *J. Bacteriol.* 185, 3307–3316.

Feldman, M., Zusman, T., Hagag, S., Segal, G., 2005. Coevolution between nonhomologous but functionally similar proteins and their conserved partners in the *Legionella* pathogenesis system. *Proc Natl Acad Sci* 102, 12206–12211.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.

Feliziani, S., Marvig, R.L., Luján, A.M., Moyano, A.J., Di Rienzo, J.A., Krogh Johansen, H., Molin, S., Smania, A.M., 2014. Coexistence and Within-Host Evolution of Diversified Lineages of Hypermutable *Pseudomonas aeruginosa* in Long-term Cystic Fibrosis Infections. *PLoS Genet.* 10, e1004651.

Feßler, A.T., Kadlec, K., Schwarz, S., 2011. Novel apramycin resistance gene apmA in bovine and porcine methicillin-resistant *Staphylococcus aureus* ST398 isolates. *Antimicrob. Agents Chemother.* 55, 373–375.

Fields, B.S., Benson, R.F., Besser, R.E., 2002. *Legionella* and Legionnaires' disease: 25 years of investigation. *Clin. Microbiol. Rev.* 15, 506–26.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res*. 44, D279–D285.

Firth, C., Lipkin, W.I., 2013. The genomics of emerging pathogens. *Annu Rev Genomics Hum Genet* 14, 281–300.

Fitzgerald, J.R., 2012. Livestock-associated *Staphylococcus aureus*: Origin, evolution and public health threat. *Trends Microbiol.* 20, 192–198.

Fitzgerald, J.R., Holden, M.T.G., 2016. Genomics of Natural Populations of *Staphylococcus aureus*. *Annu. Rev. Microbiol.* 70, 102215–095547.

Fitzgerald, J.R., Meaney, W.J., Hartigan, P.J., Smyth, C.J., Kapur, V., 1997. Fine-structure molecular epidemiological analysis of *Staphylococcus aureus* recovered from cows. *Epidemiol. Infect.* 119, 261–269.

Fitzgerald, J.R., Monday, S.R., Foster, T.J., Bohach, G.A., Hartigan, P.J., Meaney, W.J., Smyth, C.J., 2001. Characterization of a putative pathogenicity island from bovine *Staphylococcus aureus* encoding multiple superantigens. *J. Bacteriol.* 183, 63–70.

Fitzgerald, J.R., Reid, S.D., Ruotsalainen, E., Tripp, T.J., Liu, M.Y., Cole, R., Kuusela, P., Schlievert, P.M., Järvinen, A., Musser, J.M., 2003. Genome diversification in *Staphylococcus aureus*: Molecular evolution of a highly variable chromosomal region

encoding the staphylococcal exotoxin-like family of proteins. *Infect. Immun.* 71, 2827–2838.

Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., Musser, J.M., 2001. Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci.*. 98, 8821–8826.

Fogle, C.A., Nagle, J.L., Desai, M.M., 2008. Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics* 180, 2163–2173.

Foley, S.L., Johnson, T.J., Ricke, S.C., Nayak, R., Danzeisen, J., 2013. *Salmonella* pathogenicity and host adaptation in chicken-associated serovars. *Microbiol. Mol. Biol. Rev.* 77, 582–607.

Fonkwo, P.N., 2008. Pricing infectious disease. The economic and health implications of infectious diseases. *EMBO Rep*. 9, S13–S17.

Foxman, B., Riley, L., 2001. Molecular epidemiology: Focus on infection. *Am. J. Epidemiol.* 153, 1135–1141.

Foxman, B., Zhang, L., Koopman, J.S., Manning, S.D., Marrs, C.F., 2005. Choosing an appropriate bacterial typing technique for epidemiologic studies. *Epidemiol. Perspect. Innov.* 2, 10.

Frank, D.N., Feazel, L.M., Bessesen, M.T., Price, C.S., Janoff, E.N., Pace, N.R., 2010. The human nasal microbiota and *Staphylococcus aureus. PLoS One* 5, e10598.

Frazee, B.W., Lynn, J., Charlebois, E.D., Lambert, L., Lowery, D., Perdreau-Remington, F., 2005. High prevalence of methicillin-resistant *Staphylococcus aureus* in emergency department skin and soft tissue infections. *Ann. Emerg. Med.* 45, 311–320.

García, C., Ugalde, E., Campo, a B., Miñambres, E., Kovács, N., 2004. Fatal case of community-acquired pneumonia caused by *Legionella longbeachae* in a patient with systemic lupus erythematosus. *Eur. J. Clin. Microbiol. Infect. Dis.* 23, 116–118.

Gardy, J.L., Johnston, J.C., Sui, S.J.H., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M.K., Elwood, K., Jones, S.J.M., Brinkman, F.S.L., Brunham, R.C., Tang, P., 2011. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. *N. Engl. J. Med*. 364, 730–739.

Garner, M.J., Carson, C., Lingohr, E.J., Fazil, A., Edge, V.L., Waddell, J., 2015. An assessment of antimicrobial resistant disease threats in *Canada*. *PLoS One* 10, e0125155.

Gerlini, A., Colomba, L., Furi, L., Braccini, T., Manso, A.S., Pammolli, A., Wang, B., Vivi, A., Tassini, M., van Rooijen, N., Pozzi, G., Ricci, S., Andrew, P.W., Koedel, U., Moxon, E.R., Oggioni, M.R., 2014. The Role of Host and Microbial Factors in the Pathogenesis of Pneumococcal Bacteraemia Arising from a Single Bacterial Cell Bottleneck. *PLoS Pathog*. 10, e1004026.

Gillaspy A., Worrell, V., Orvis, J., Roe, B., Dyer, D., Iandolo, J. 2006. The *Staphylococcus aureus* NCTC 8325 Genome, p 381-412. *In* Fischetti, V., Novick, R., Ferretti, J., Portnoy, D., Rood, J. (ed), *Gram-Positive Pathogens, Second Edition*. ASM Press, Washington, DC. doi: 10.1128/9781555816513. Chapter 32.

Giraud, T., Koskella, B., Laine, A.L., 2017. Introduction: microbial local adaptation: insights from natural populations, genomics and experimental evolution. *Mol. Ecol.* 26, 1703–1710.

Goerke, C., Pantucek, R., Holtfreter, S., Schulte, B., Zink, M., Grumann, D., Bröker, B.M.,

Doskar, J., Wolz, C., 2009. Diversity of prophages in dominant *Staphylococcus aureus* clonal lineages. *J. Bacteriol.* 191, 3462–3468.

Goerke, C., Wirtz, C., Flückiger, U., Wolz, C., 2006. Extensive phage dynamics in *Staphylococcus aureus* contributes to adaptation to the human host during infection. *Mol. Microbiol*. 61, 1673–1685.

Goerke, C., Wolz, C., 2004. Regulatory and genomic plasticity of *Staphylococcus aureus* during persistent colonization and infection. *Int. J. Med. Microbiol.* 294, 195–202.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–36.

Golubchik, T., Batty, E.M., Miller, R.R., Farr, H., Young, B.C., Larner-Svensson, H., Fung, R., Godwin, H., Knox, K., Votintseva, A., Everitt, R.G., Street, T., Cule, M., Ip, C.L.C., Didelot, X., Peto, T.E. a, Harding, R.M., Wilson, D.J., Crook, D.W., Bowden, R., 2013. Within-host evolution of *Staphylococcus aureus* during asymptomatic carriage. *PLoS One* 8, e61319.

Gomez-Valero, L., Rusniok, C., Jarraud, S., Vacherie, B., Rouy, Z., Barbe, V., Médigue, C., Etienne, J., Buchrieser, C., 2011. Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics* 12, 536.

Gorak, E.J., Yamada, S.M., Brown, J.D., 2010. Community-Acquired Methicillin-Resistant *Staphylococcus aureus* in Hospitalized Adults and Children without Known Risk Factors. *Clin. Infect. Dis.* 29, 797–800.

Gordon, R.J., Lowy, F.D., 2008. Pathogenesis of Methicillin-Resistant *Staphylococcus aureus* Infection. *Clin. Infect. Dis.* 46, S350–S359.

Grad, Y.H., Lipsitch, M., 2014. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biol.* 15, 538.

Graham, F.F., White, P.S., Harte, D.J.G., Kingham, S.P., 2012. Changing epidemiological trends of legionellosis in New Zealand, 1979–2009. *Epidemiol. Infect.* 140, 1481–1496.

Graham, R.M.A., Doyle, C.J., Jennison, A. V, 2014. Real-time investigation of a *Legionella pneumophila* outbreak using whole genome sequencing. *Epidemiol. Infect*. 142, 2347–51.

Grave, K., Torren-Edo, J., Mackay, D., 2010. Comparison of the sales of veterinary antibacterial agents between 10 European countries. *J. Antimicrob. Chemother.* 65, 2037–2040.

Griesbeck-Zilch, B., Osman, M., Kühn, C., Schwerin, M., Bruckmaier, R. H., Pfaffl, M. W., Wellnitz, O. 2009. Analysis of key molecules of the innate immune system in mammary epithelial cells isolated from marker-assisted and conventionally selected cattle. *J. Dairy Sci.*, *92*(9), 4621–4633.

Grimstead, D., Tucker, D., Harris, K., Turner, D., 2015. Cutaneous *Legionella longbeachae* infection in immunosuppressed woman, United Kingdom. *Emerg. Infect. Dis.* 21, 1426–1428.

Grumann, D., Nübel, U., Bröker, B.M., 2014. *Staphylococcus aureus* toxins - Their functions and genetics. I*nfect. Genet. Evol.* 21, 583–592.

Guerrero, I., Ferrian, S., Penadés, M., García-Quirós, A., Pascual, J.J., Selva, L., Viana, D., Corpa, J.M., 2015. Host responses associated with chronic staphylococcal mastitis in rabbits. *Vet. J.* 204, 338–344.

Guidot, A., Jiang, W., Ferdy, J.B., Thébaud, C., Barberis, P., Gouzy, J., Genin, S., 2014.

Multihost experimental evolution of the pathogen *Ralstonia solanacearum* unveils genes involved in adaptation to plants. *Mol. Biol. Evol.* 31, 2913–2928.

Guinane, C.M., Zakour, N.L. Ben, Tormo-Mas, M.A., Weinert, L.A., Lowder, B. V., Cartwright, R.A., Smyth, D.S., Smyth, C.J., Lindsay, J.A., Gould, K.A., Witney, A., Hinds, J., Bollback, J.P., Rambaut, A., Penadés, J.R., Fitzgerald, J.R., 2010. Evolutionary genomics of *Staphylococcus aureus* reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biol. Evol*. 2, 454–466.

Gulcher, J., Stefansson, K., 1998. Population genomics: Laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med*. 36, 523–527.

Guttman, D.S., Stavrinides, J., 2010. Population Genomics of Bacteria. In: Bacterial Population Genetics in Infectious Disease. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 121–151.

Haagmans, B.L., Andeweg, A.C., Osterhaus, A.D.M.E., 2009. The application of genomics to emerging zoonotic viral diseases. *PLoS Pathog.* 5, e1000557.

Hadfield, J., Croucher, N.J., Goater, R.J., Abudahab, K., Aanensen, D.M., Harris, S.R., 2017. Phandango: an interactive viewer for bacterial population genomics. *bioRxiv* 119545.

Halasa, T., Nielen, M., Huirne, R.B.M., Hogeveen, H., 2009. Stochastic bio-economic model of bovine intramammary infection. *Livest. Sci.* 124, 295–305.

Hanage, W.P., Fraser, C., Spratt, B.G., 2005a. Fuzzy species among recombinogenic bacteria. *BMC Biol*. 3, 6.

Hanage, W.P., Kaijalainen, T., Herva, E., Saukkoriipi, A., Syrjänen, R., Spratt, B.G., 2005b. Using multilocus sequence data to define the *pneumococcus*. *J. Bacteriol.* 187, 6223–30.

Harkins, C.P., Pichon, B., Doumith, M., Parkhill, J., Westh, H., Tomasz, A., de Lencastre, H., Bentley, S.D., Kearns, A.M., Holden, M.T.G., 2017. Methicillin resistant *Staphylococcus aureus* emerged long before the introduction of methicillin in to clinical practice. *bioRxiv* 18.

Harrap, B.S., Woods, E.F., 1964. Soluble derivatives of feather keratin. 1. Isolation, fractionation and amino acid composition. *Biochem.* J. 92, 8–18.

Harris, S.R., Cartwright, E.J.P., Török, M.E., Holden, M.T.G., Brown, N.M., Ogilvy-Stuart, A.L., Ellington, M.J., Quail, M. a, Bentley, S.D., Parkhill, J., Peacock, S.J., 2013. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* 13, 130–136.

Harris, S. R., Robinson, C., Steward, K. F., Webb, K. S., Paillot, R., Parkhill, J., Holden, T. H. M., Waller, A. S. 2015. Genome specialization and decay of the strangles pathogen, *Streptococcus equ*i, is driven by persistent infection. *Genome Research*, *25*(9), 1360–71.

Harrison, E.M., Paterson, G.K., Holden, M.T.G., Larsen, J., Stegger, M., Larsen, A.R., Petersen, A., Skov, R.L., Christensen, J.M., Bak Zeuthen, A., Heltberg, O., Harris, S.R., Zadoks, R.N., Parkhill, J., Peacock, S.J., Holmes, M. a, 2013. Whole genome sequencing identifies zoonotic transmission of MRSA isolates with the novel *mec*A homologue *mec*C. *EMBO Mol. Med*. 5, 509–15.

Hartl, D.L., Dykhuizen, D.E., 1984. The population genetics of *Escherichia coli*. Annu. Rev. Genet. 18, 31–68.

Hasman, H., Moodley, A., Guardabassi, L., Stegger, M., Skov, R.L., Aarestrup, F.M., 2010. spa type distribution in *Staphylococcus aureus* originating from pigs, cattle and poultry. *Vet. Microbiol.* 141, 326–331.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C.A., Frimodt-Møller, N., Aarestrup, F.M., 2014. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol.* 52, 139–46.

Hedge, J., Wilson, D.J., 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio* 5, 5–8.

Hendriksen, R.S., Price, L.B., Schupp, J.M., Gillece, J.D., Kaas, R.S., Engelthaler, D.M., Bortolaia, V., Pearson, T., Waters, A.E., Upadhyay, B.P., Shrestha, S.D., Adhikari, S., Shakya, G., Keim, P.S., Aarestrup, F.M., 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio* 2, e00157–11.

Hendriksen, R.S., Vieira, A.R., Karlsmose, S., Lo Fo Wong, D.M. a, Jensen, A.B., Wegener, H.C., Aarestrup, F.M., 2011. Global monitoring of *Salmonella* serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. *Foodborne Pathog. Dis.* 8, 887–900.

Hennekinne, J.A., Kerouanton, A., Brisabois, A., De Buyser, M.L., 2003. Discrimination of *Staphylococcus aureus* biotypes by pulsed-field gel electrophoresis of DNA macro-restriction fragments. *J. Appl. Microbiol*. 94, 321–329.

Herold, B.C., 1998. Community-Acquired Methicillin-Resistant *Staphylococcus aureus* in Children With No Identified Predisposing Risk. *Jama* 279, 593.

Herron-Olson, L., Fitzgerald, J.R., Musser, J.M., Kapur, V., 2007. Molecular correlates of host specialization in *Staphylococcus aureus*. *PLoS One* 2, e1120.

Hindré, T., Knibbe, C., Beslon, G., Schneider, D., 2012. New insights into bacterial adaptation through in vivo and in silico experimental evolution. *Nat. Rev. Microbiol.* 10, 352–365.

Hoang, K.L., Morran, L.T., Gerardo, N.M., 2016. Experimental evolution as an underutilized tool for studying beneficial animal-microbe interactions. *Front. Microbiol*. 7, 1444.

Hoban, S., Bertorelle, G., Gaggiotti, O.E., 2011. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet* 13, 110–122.

Holden, M.T., Feil, E.J., Lindsay, J.A., Peacock, S.J., Day, N.P., Enright, M.C., Foster, T.J., Moore, C.E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S.D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., Doggett, J., Dowd, L., Feltwell, T., Hance, Z., Harris, B., Hauser, H., Holroyd, S., Jagels, K., James, K.D., Lennard, N., Line, A., Mayes, R., Moule, S., Mungall, K., Ormond, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Sanders, M., Sharp, S., Simmonds, M., Stevens, K., Whitehead, S., Barrell, B.G., Spratt, B.G., Parkhill, J., 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc. Nat. Acad. Sci.*. 101, 9786–9791.

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

Holden, M.T.G., Lindsay, J.A., Corton, C., Quail, M.A., Cockfield, J.D., Pathak, S., Batra, R., Parkhill, J., Bentley, S.D., Edgeworth, J.D., 2010. Genome sequence of a recently

emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant *Staphylococcus aureus*, sequence type 239 (TW). *J. Bacteriol.* 192, 888–92.

Holmes, M.A., Zadoks, R.N., 2011. Methicillin resistant *S. aureus* in human and bovine mastitis. J. Mammary Gland Biol. Neoplasia 16, 373–382.

Hong, J., Gresham, D., 2014. Molecular Specificity, Convergence and Constraint Shape Adaptive Evolution in Nutrient-Poor Environments. *PLoS Genet.* 10, e1004041.

Hongo, J.A., de Castro, G.M., Cintra, L.C., Zerlotini, A., Lobo, 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics* 16, 567.

Hottes, A.K., Freddolino, P.L., Khare, A., Donnell, Z.N., Liu, J.C., Tavazoie, S., 2013. Bacterial Adaptation through Loss of Function. *PLoS Genet.* 9, e1003617.

Howden, B.P., Holt, K.E., Lam, M.M.C., Seemann, T., Ballard, S., Coombs, G.W., Tong, S.Y.C., Grayson, M.L., Johnson, P.D.R., Stinear, T.P., 2013. Genomic insights to control the emergence of vancomycin-resistant enterococci. *MBio* 4, e00412–13.

Howden, B.P., McEvoy, C.R.E., Allen, D.L., Chua, K., Gao, W., Harrison, P.F., Bell, J., Coombs, G., Bennett-Wood, V., Porter, J.L., Robins-Browne, R., Davies, J.K., Seemann, T., Stinear, T.P., 2011. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. *PLoS Pathog.* 7, e1002359.

Howden, B.P., Peleg, A.Y., Stinear, T.P., 2014. The evolution of vancomycin intermediate *Staphylococcus aureus* (VISA) and heterogenous-VISA. I*nfect. Genet. Evol.* 21, 575–582.

Howell, K.J., Weinert, L. a, Chaudhuri, R.R., Luan, S.-L., Peters, S.E., Corander, J., Harris, D., Angen, Ø., Aragon, V., Bensaid, A., Williamson, S.M., Parkhill, J., Langford, P.R., Rycroft, A.N., Wren, B.W., Holden, M.T., Tucker, A.W., Maskell, D.J., 2014. The use of genome wide association methods to investigate pathogenicity, population structure and serovar in *Haemophilus parasuis*. *BMC Genomics* 15, 1179.

Hudson, P.J., Perkins, S.E., Cattadori, I.M., 2008. Ch. 16 -- The Emergence of Wildlife Disease and the Application of Ecology. Infect. Dis. Ecol. Eff. Ecosyst. Dis. Dis. Ecosyst. 347–367.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 278, 631–637.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., Polz, M.F., 2008. Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. *Science* (80). 320, 1081–1085.

Hunter, P.R., 1990. Reproducibility and indices of discriminatory power of microbial typing methods. *J. Clin. Microbiol.* 28, 1903–1905.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267.

Isberg, O'Connor, Heidtman, 2009. The *Legionella pneumophila* replication vacuole: making a cosy niche inside host cells. *Nat. Rev. Microbiol.* 7, 13–24.

Jevons, M.P., 1961. "Celbenin" - resistant Staphylococci. *Bmj* 1, 124–125.

Jia, H., Du, P., Yang, H., Zhang, Y., Wang, J., Zhang, W., Han, G., Han, N., Yao, Z., Wang,

H., Zhang, J., Wang, Z., Ding, Q., Qiang, Y., Barbut, F., Gao, G.F., Cao, Y., Cheng, Y., Chen, C., 2016. Nosocomial transmission of *Clostridium difficile* ribotype 027 in a Chinese hospital, 2012-2014, traced by whole genome sequencing. *BMC Genomics* 17, 405.

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., Aarestrup, F.M., 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 52, 1501–1510.

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.

Jones, B.A., Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M.Y., McKeever, D., Mutua, F., Young, J., McDermott, J., Pfeiffer, D.U., 2013. Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci.* 110, 8399–404.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. *Nature* 451, 990–U4.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.

Jose L. Raygada; Donald P. Levine, 2009. Methicillin-Resistant *Staphylococcus aureus*: A Growing Risk in the Hospital and in the Community. *Am Heal. Drug Benefits* 2, 86–95.

Joseph, C.A., Ricketts, K.D., 2010. Legionnaires disease in Europe 2007-2008. *Euro Surveill.* 15, 19493.

Juhas, M., 2013. Horizontal gene transfer in human pathogens. *Crit. Rev. Microbiol.* 7828, 1–8.

Kock, R., Becker, K., Cookson, B., van Gemert-Pijnen, J.E., Harbarth, S., Kluytmans, J., Mielke, M., Peters, G., Skov, R.L., Struelens, M.J., Tacconelli, E., Navarro Torn??, A., Witte, W., Friedrich, A.W., 2010. Methicillin-resistant *Staphylococcus aureus* (MRSA): burden of disease and control challenges in Europe. *Euro Surveill.* 15, 19688.

Kadlec, K., Schwarz, S., 2009. Novel ABC transporter gene, vga(C), located on a multiresistance plasmid from a porcine methicillin-resistant *Staphylococcus aureus* ST398 strain. *Antimicrob. Agents Chemother.* 53, 3589–3591.

Kalia, A., Bessen, D.E., 2004. Natural selection and evolution of streptococcal virulence genes involved in tissue-specific adaptations. *J. Bacteriol*. 186, 110–21.

Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E.H., MacHalaba, C.C., Thomas, M.J., Heymann, D.L., 2012. Ecology of zoonoses: Natural and unnatural histories. *Lancet* 380, 1936–1945.

Katayama, Y., Ito, T., Hiramatsu, K., 2000. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother*. 44, 1549–1555.

Katz, L.S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E.S., Turnsek, M.A., Guo, Y., Wang, S., Paxinos, E.E., Orata, F., Gladney, L.M., Stroika, S., Folster, J.P., Rowe, L., Freeman, M.M., Knox, N., Frace, M., Boncy, J., Graham, M., Hammer, B.K., Boucher, Y., Bashir, A., Hanage, W.P., Van Domselaar, G., Tarr, C.L., 2013. Evolutionary

dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *MBio* 4, e00398–13.

Kawecki, T.J., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I., Whitlock, M.C., 2012. Experimental evolution. *Trends Ecol. Evol. 27,* 547–560.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.

Keightley, P.D., 1998. Inference of genome-wide mutation rates and distributions of mutation effects for fitness traits: A simulation study. Genetics 150, 1283–1293.

Kenagy, E., Priest, P.C., Cameron, C.M., Smith, D., Scott, P., Cho, V., Mitchell, P., Murdoch, D.R., 2017. Risk Factors for *Legionella longbeachae* Legionnaires' Disease, New Zealand. *Emerg. Infect. Dis.* 23, 1148–1154.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., Reinhardt, R., Correa, P., Meyer, T.F., Josenhans, C., Falush, D., Suerbaum, S., 2011. *Helicobacter pylori* genome evolution during human infection. *Proc. Natl. Acad. Sci*.. 108, 5033–5038.

Khaliq, Z., Leijon, M., Belák, S., Komorowski, J., 2016. Identification of combinatorial host-specific signatures with a potential to affect host adaptation in influenza A H1N1 and H3N2 subtypes. *BMC Genomic*s 17.

Khan, S.A., 2005. Plasmid rolling-circle replication: Highlights of two decades of research. *Plasmid* 53, 126–136.

Kilpatrick, A.M., Randolph, S.E., 2012. Drivers, dynamics, and control of emerging vector-borne zoonotic diseases. *Lancet* 380, 1946–1955.

Kim, M., Oh, H.S., Park, S.C., Chun, J., 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol*. 64, 346–351.

Kirby, W.M., 1944. Extraction of a Highly Potent Penicillin Inactivator From Penicillin Resistant Staphylococci. *Science* 99, 452–3.

Klemm, E.J., Gkrania-Klotsas, E., Hadfield, J., Forbester, J.L., Harris, S.R., Hale, C., Heath, J.N., Wileman, T., Clare, S., Kane, L., Goulding, D., Otto, T.D., Kay, S., Doffinger, R., Cooke, F.J., Carmichael, A., Lever, A.M.L., Parkhill, J., Maclennan, C.A., Kumararatne, D., Dougan, G., Kingsley, R.A., 2016. Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nat. Microbiol.* 1, 1–6.

Kloos, W.E., 1980. Natural Populations of the Genus *Staphylococcus*. *Ann. Rev. Microbio*l 34, 559–592.

Kluytmans, J., Van Belkum, A., Verbrugh, H., 1997. Nasal carriage of *Staphylococcus aureus*: Epidemiology, underlying mechanisms, and associated risks. *Clin. Microbiol. Rev.* 10, 505–520.

Knox, N.C., Weedmark, K.A., Conly, J., Ensminger, A.W., Hosein, F.S., Drews, S.J., 2016. Unusual Legionnaires' outbreak in cool, dry Western Canada: an investigation using genomic epidemiology. *Epidemiol. Infect.* 145, 1–12.

Koide, M., Arakaki, N., Saito, A., 2001. Distribution of *Legionella longbeachae* and other legionellae in Japanese potting soils. J. *Infect. Chemother.* 7, 224–227.

Koop, G., Vrieling, M., Storisteanu, D.M.L., Lok, L.S.C., Monie, T., van Wigcheren, G.,

Raisen, C., Ba, X., Gleadall, N., Hadjirin, N., Timmerman, A.J., Wagenaar, J.A., Klunder, H.M., Fitzgerald, J.R., Zadoks, R., Paterson, G.K., Torres, C., Waller, A.S., Loeffler, A., Loncaric, I., Hoet, A.E., Bergström, K., De Martino, L., Pomba, C., de Lencastre, H., Ben Slama, K., Gharsa, H., Richardson, E.J., Chilvers, E.R., de Haas, C., van Kessel, K., van Strijp, J.A.G., Harrison, E.M., Holmes, M.A., 2017. Identification of LukPQ, a novel, equid-adapted leukocidin of *Staphylococcus aureus*. *Sci. Rep.* 7, 40660.

Köser, C.U., Ellington, M.J., Cartwright, E.J.P., Gillespie, S.H., Brown, N.M., Farrington, M., Holden, M.T.G., Dougan, G., Bentley, S.D., Parkhill, J., Peacock, S.J., 2012. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. PLoS Pathog. 8, e1002824.

Köser, C.U., Holden, M.T., Ellington, M.J., Cartwright, E.J., Brown, N.M., Ogilvy-Stuart, A.L., Hsu, L.Y., Chewapreecha, C., Croucher, N.J., Harris, S.R., Sanders, M., Enright, M.C., Dougan, G., Bentley, S.D., Parkhill, J., Fraser, L.J., Betley, J.R., Schulz-Trieglaff, O.B., Smith, G.P., Peacock, S.J., 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N.Engl.J.Med.* 366, 2267–2275.

Kozak, N.A., Buss, M., Lucas, C.E., Frace, M., Govil, D., Travis, T., Olsen-Rasmussen, M., Benson, R.F., Fields, B.S., 2010. Virulence factors encoded by *Legionella longbeachae* identified on the basis of the genome sequence analysis of clinical isolate D-4968. *J. Bacteriol.* 192, 1030–1044.

Kryazhimskiy, S., Plotkin, J.B., 2008. The population genetics of dN/dS. *PLoS Genet*. 4.

Kurkela, S., Brown, D.W.G., 2009. Molecular diagnostic techniques. *Medicine (Baltimore).* 37, 535–540.

Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., Nagai, Y., Lian, J., Ito, T., Kanamori, M., Matsumaru, H., Maruyama, A., Murakami, H., Hosoyama, A., Mizutani-Ui, Y., Takahashi, N.K., Sawano, T., Inoue, R., Kaito, C., Sekimizu, K., Hirakawa, H., Kuhara, S., Goto, S., Yabuzaki, J., Kanehisa, M., Yamashita, A., Oshima, K., Furuya, K., Yoshino, C., Shiba, T., Hattori, M., Ogasawara, N., Hayashi, H., Hiramatsu, K., 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* 357, 1225–1240.

Kwan, T., Liu, J., DuBow, M., Gros, P., Pelletier, J., 2005. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc. Natl. Acad. Sci*. 102, 5174–9.

Laabei, M., Recker, M., Rudkin, J.K., Aldeljawi, M., Gulay, Z., Sloan, T.J., Williams, P., Endres, J.L., Bayles, K.W., Fey, P.D., Yajjala, V.K., Widhelm, T., Hawkins, E., Lewis, K., Parfett, S., Scowen, L., Peacock, S.J., Holden, M., Wilson, D., Read, T.D., Van Den Elsen, J., Priest, N.K., Feil, E.J., Hurst, L.D., Josefsson, E., Massey, R.C., 2014. Predicting the virulence of MRSA from its genome sequence. *Genome Res.* 24, 839–849.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108.

Lan, R., Reeves, P.R., 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol*. 9, 419–24.

Landrum, M.L., Neumann, C., Cook, C., Chukwuma, U., Ellis, M.W., Hospenthal, D.R., Murray, C.K., 2012. Epidemiology of *Staphylococcus aureus* blood and skin and soft tissue infections in the US military health system, 2005-2010. *Jama* 308, 50–59.

Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D., Desai, M.M., 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast

populations. *Nature* 500, 571–4.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Method*s 9, 357–359.

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., Thomson, N. R. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc. Natl. Acad. Sci*., *112*(3), 863–8.

Lanza, V.F., de Toro, M., Garcillán-Barcia, M.P., Mora, A., Blanco, J., Coque, T.M., de la Cruz, F., 2014. Plasmid Flux in Escherichia coli ST131 Sublineages, Analyzed by Plasmid Constellation Network (PLACNET), a New Method for Plasmid Reconstruction from Whole Genome Sequences. *PLoS Genet*. 10.

Layton, M.C., Calliste, S.G., Gomez, T.M., Patton, C., Brooks, S., 1997. A mixed foodborne outbreak with *Salmonella heidelberg* and *Campylobacter jejuni* in a nursing home. *Infect. Control Hosp. Epidemiol.* 18, 115–21.

Lees, J.A., Vehkala, M., Välimäki, N., Harris, S.R., Chewapreecha, C., Croucher, N.J., Marttinen, P., Davies, M.R., Steer, A.C., Tong, S.Y.C., Honkela, A., Parkhill, J., Bentley, S.D., Corander, J., 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *bioRxiv* 7, 038463.

Lefébure, T., Stanhope, M.J., 2009. Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus *Campylobacter*. *Genome Res.* 19, 1224–1232.

Lefebure, T., Stanhope, M.J., Lefébure, T., Stanhope, M.J., 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* 8, R71.

Leggieri, N., Gouriet, F., Thuny, F., Habib, G., Raoult, D., Casalta, J.P., 2012. *Legionella longbeachae* and endocarditis. *Emerg. Infect. Dis.* 18, 95–97.

Leiby, N., Marx, C.J., 2014. Metabolic Erosion Primarily Through Mutation Accumulation, and Not Tradeoffs, Drives Limited Evolution of Substrate Specificity in *Escherichia coli*. *PLoS Biol.* 12, e1001789.

Lescat, M., Launay, A., Ghalayini, M., Magnan, M., Glodt, J., Pintard, C., Dion, S., Denamur, E., Tenaillon, O., 2017. Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an *Escherichia coli* natural isolate in the streptomycin-treated mouse gut. *Mol. Ecol.* 26, 1802–1817.

Lévesque, S., Plante, P.-L., Mendis, N., Cantin, P., Marchand, G., Charest, H., Raymond, F., Huot, C., Goupil-Sormany, I., Desbiens, F., Faucher, S.P., Corbeil, J., Tremblay, C., 2014. Genomic characterization of a large outbreak of *Legionella pneumophila* serogroup 1 strains in Quebec City, 2012. *PLoS One* 9, e103852.

Levin, B.R., 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99, 1–23.

Lewontin, R.C., 1985. Population genetics. *Annu. Rev. Genet*. 19, 81–102.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, J.S., O'Brien, E.D., Guest, C., 2002. A review of national legionellosis surveillance in

Australia, 1991 to 2000. *Commun. Dis. Intell*. 26, 461–468.

Li, L., Stoeckert, C.J.J., Roos, D.S., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res*. 13, 2178–2189.

Li, L.M., Grassly, N.C., Fraser, C., 2014. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biol.* 15, 541.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P., Kishony, R., 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* 46, 82–87.

Lieberman, T.D., Michel, J.-B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R., Skurnik, D., Leiby, N., LiPuma, J.J., Goldberg, J.B., McAdam, A.J., Priebe, G.P., Kishony, R., 2011. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* 43, 1275–1280.

Lindsay, D.S.J., Brown, A.W., Brown, D.J., Pravinkumar, S.J., Anderson, E., Edwards, G.F.S., 2012. *Legionella longbeachae* serogroup 1 infections linked to potting compost. *J. Med. Microbiol.* 61, 218–222.

Lindsay, J.A. 2008. *S. aureus* Evolution: Lineages and Mobile Genetic Elements (MGEs), p. 45-69. In: Lindsay, J.A. (ed.), *Staphylococcus* molecular genetics. Caister Academic Press, Norfolk, UK.

Lindsay, J.A., Holden, M.T.G., 2004. *Staphylococcus aureus*: Superbug, super genome? Trends Microbiol. 12, 378–385.

Lindsay, J.A., Holden, M.T.G., 2006. Understanding the rise of the superbug: Investigation of the evolution and genomic variation of *Staphylococcus aureus*. *Funct. Integr. Genomics* 6, 186–201.

Llarena, A. K., Zhang, J., Vehkala, M., Välimäki, N., Hakkinen, M., Hänninen, M.-L., Roasto, M., Maesaar, M., Taboada, E., Barker, D., Garofolo, G., Camma, C., Giannatale, E., Corander, J., Rossi, M. (2016). Monomorphic genotypes within a generalist lineage of Campylobacter jejuni show signs of global dispersion. *Microbial Genomics*, 2(10), e000088.

Loeffler, a, Pfeiffer, D.U., Lindsay, J. a, Magalhães, R.J.S., Lloyd, D.H., 2010. Prevalence of and risk factors for MRSA carriage in companion animals: a survey of dogs, cats and horses. *Epidemiol. Infect.* 1–10.

Loewe, L., Hill, W.G., 2010. The population genetics of mutations: good, bad and indifferent. Philos. *Trans. R. Soc. Lond. B. Biol. Sci.* 365, 1153–1167.

Long, A., Liti, G., Luptak, A., Tenaillon, O., 2015. Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nat. Rev. Genet*. 16, 567–582.

Lowder, B. V, Guinane, C.M., Ben Zakour, N.L., Weinert, L.A., Conway-Morris, A., Cartwright, R.A., Simpson, A.J., Rambaut, A., Nübel, U., Fitzgerald, J.R., 2009. Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proc.Natl.Acad.Sci.* 106, 19545–19550.

Lowy, F.D., 1998. *Staphylococcus aureus* Infections. N. Engl. J. Med. 339, 520–532.

Löytynoja, A., Goldman, N., 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci*. 102, 10557–62.

Lozano, C., García-Migura, L., Aspiroz, C., Zarazaga, M., Torres, C., Aarestrup, F.M., 2012. Expansion of a plasmid classification system for gram-positive bacteria and determination of the diversity of plasmids in *Staphylococcus aureus* strains of human,

animal, and food origins. *Appl. Environ. Microbiol.* 78, 5948–5955.

Luong, T.T., Ouyang, S., Bush, K., Lee, C.Y., 2002. Type 1 capsule genes of *Staphylococcus aureus* are carried in a staphylococcal cassette chromosome genetic element. *J. Bacteriol.* 184, 3623–3629.

Lupolova, N., Dallman, T.J., Matthews, L., Bono, J.L., Gally, D.L., 2016. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci.*113, 11312–11317.

Lynch, M., Ackerman, M.S., Gout, J.-F., Long, H., Sung, W., Thomas, W.K., Foster, P.L., 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* 17, 704–714.

Maere, S., Heymans, K., Kuiper, M., 2005. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 21, 3448–3449.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci*. 95, 3140–5.

Makris, G., Wright, J.D., Ingham, E., Holland, K.T., 2004. The hyaluronate lyase of *Staphylococcus aureus* - A virulence factor? *Microbiology* 150, 2005–2013.

Malachowa, N., Deleo, F.R., 2010. Mobile genetic elements of *Staphylococcus aureus*. *Cell. Mol. Life Sci.* 67, 3057–3071.

Marano, N., Pappaioanou, M., 2004. Links between Human and Animal Health. *Emerg. Infect. Dis*. 10, 2065–2066.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J., Bryant, S.H., 2003. CDD: A curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31, 383–387.

Markussen, T., Marvig, R.L., Gómez-Lozano, M., Aanæs, K., Burleigh, A.E., Høiby, N., Johansen, H.K., Molin, S., Jelsbak, L., 2014. Environmental heterogeneity drives within-host diversification and evolution of *Pseudomonas aeruginosa. MBio* 5, e01592–14.

Marston, B.J., Lipman, H.B., Breiman, R.F., 1994. Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. *Arch. Intern. Med.* 154, 2417–22.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10.

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D., Corander, J., 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* 40, 1–12.

Marvig, R.L., Johansen, H.K., Molin, S., Jelsbak, L., 2013. Genome Analysis of a Transmissible Lineage of Pseudomonas aeruginosa Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. *PLoS Genet*. 9, e1003741.

Marvig, R.L., Sommer, L.M., Molin, S., Johansen, H.K., 2015. Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat Genet* 47, 57–65.

Mather, A.E., Matthews, L., Mellor, D.J., Reeve, R., Denwood, M.J., Boerlin, P., Reid-Smith, R.J., Brown, D.J., Coia, J.E., Browning, L.M., Haydon, D.T., Reid, S.W., 2012. An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations. *Proc Biol Sci* 279, 1630–1639.

Mather, A.E., Reid, S.W., Maskell, D.J., Parkhill, J., Fookes, M.C., Harris, S.R., Brown, D.J., Coia, J.E., Mulvey, M.R., Gilmour, M.W., Petrovska, L., de Pinna, E., Kuroda, M., Akiba, M., Izumiya, H., Connor, T.R., Suchard, M.A., Lemey, P., Mellor, D.J., Haydon, D.T., Thomson, N.R., 2013. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. *Science* (80). 341, 1514–1517.

Mathers, A.J., Stoesser, N., Sheppard, A.E., Pankhurst, L., Giess, A., Yeh, A.J., Didelot, X., Turner, S.D., Sebra, R., Kasarskis, A., Peto, T., Crook, D., Sifri, C.D., 2015. *Klebsiella pneumoniae* Carbapenemase (KPC)-Producing *K. pneumoniae* at a Single Institution: Insights into Endemicity from Whole-Genome Sequencing. *Antimicrob. Agents Chemother.* 59, 1656–1663.

Matthews, T. D., Edwards, R., & Maloy, S. (2010). Chromosomal Rearrangements Formed by rrn Recombination Do Not Improve Replichore Balance in Host-Specific *Salmonella enterica* Serovars. *PLoS ONE*, *5*(10), e13503.

McAdam, P.R., Templeton, K.E., Edwards, G.F., Holden, M.T.G., Feil, E.J., Aanensen, D.M., Bargawi, H.J.A., Spratt, B.G., Bentley, S.D., Parkhill, J., Enright, M.C., Holmes, A., Girvan, E.K., Godfrey, P.A., Feldgarden, M., Kearns, A.M., Rambaut, A., Robinson, D.A., Fitzgerald, J.R., 2012. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. *Proc. Natl. Acad. Sci*. 109, 9107–12.

McAdam, P.R., Vander Broek, C.W., Lindsay, D., Ward, M.J., Hanson, M.F., Gillies, M., Watson, M., Stevens, J.M., Edwards, G.F., Fitzgerald, J., 2014. Gene flow in environmental *Legionella pneumophila* leads to genetic and pathogenic heterogeneity within a Legionnaires' disease outbreak. *Genome Biol.* 15, 504.

McCarthy, A.J., Loeffler, A., Witney, A.A., Gould, K.A., Lloyd, D.H., Lindsay, J.A., 2014. Extensive horizontal gene transfer during *Staphylococcus aureus* co-colonization in vivo. *Genome Biol. Evol*. 6, 2697–2708.

McCarthy, A.J., van Wamel, W., Vandendriessche, S., Larsen, J., Denis, O., Garcia-Graells, C., Uhlemann, A.C., Lowy, F.D., Skov, R., Lindsay, J.A., 2012. *Staphylococcus aureus* CC398 clade associated with human-to-human transmission. *Appl. Environ. Microbiol.* 78, 8845–8848.

McCarthy, A.J., Witney, A.A., Gould, K.A., Moodley, A., Guardabassi, L., Voss, A., Denis, O., Broens, E.M., Hinds, J., Lindsay, J.A., 2011. The Distribution of Mobile Genetic Elements (MGEs) in MRSA CC398 Is Associated with Both Host and Country. *Genome Biol. Evol.* 3, 1164–1174.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20, 1297–1303.

McKinney, R.M., Porschen, R.K., Edelstein, P.H., Bissett, M.L., Harris, P.P., Bondell, S.P., Steigerwalt, A.G., Weaver, R.E., Ein, M.E., Lindquist, D.S., Kops, R.S., Brenner, D.J., 1981. *Legionella longbeachae* species nova, another etiologic agent of human pneumonia. *Ann. Intern. Med*. 94, 739–743.

McNamee, P.T., Smyth, J. a, 2000. Bacterial chondronecrosis with osteomyelitis ('femoral

head necrosis') of broiler chickens: a review. *Avian Pathol.* 29, 477–495.

McVicker, G., Prajsnar, T.K., Williams, A., Wagner, N.L., Boots, M., Renshaw, S.A., Foster, S.J., 2014. Clonal Expansion during *Staphylococcus aureus* Infection Dynamics Reveals the Effect of Antibiotic Intervention. *PLoS Pathog.* 10, e1003959.

Meaden, S., Koskella, B., 2017. Adaptation of the pathogen, *Pseudomonas syringae*, during experimental evolution on a native vs. alternative host plant. *Mol. Ecol.* 26, 1790–1801.

Mee-marquet, N. Van Der, Domelier, A., Arnault, L., Bloc, D., Laudat, P., Quentin, R., Hartemann, P., 2006. *Legionella anis*a , a Possible Indicator of Water Contamination by *Legionella pneumophila. Journ. Clin. Microbiology* 44, 56–59.

Melchior, M.B., Vaarkamp, H., Fink-Gremmels, J., 2006. Biofilms: A role in recurrent mastitis infections? *Vet. J.* 171, 398–407.

Mentula, S., Ruotsalainen, E., Perola, O., Pentikäinen, J., 2014. *Legionella longbeachae* infection in a persistent hand-wound after a gardening accident. *JMM Case Reports* 1, 2014–2016.

Menzies, P.I., Ramanoon, S.Z., 2001. Mastitis of sheep and goats. Vet. Clin. North *Am. Food Anim. Pract.* 17, 333–58, vii.

Mercante, J.W., Winchell, J.M., 2015. Current and emerging legionella diagnostics for laboratory and outbreak investigations. *Clin. Microbiol. Rev.* 28, 95–133.

Merhej, V., Royer-Carenzi, M., Pontarotti, P., Raoult, D. 2009. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biology Direct,* 4(1), 13.

Messenger, A.M., Barnes, A.N., Gray, G.C., 2014. Reverse zoonotic disease transmission (Zooanthroponosis): A systematic review of seldom-documented human biological threats to animals. *PLoS One* 9, e89055.

Messina, J.A., Thaden, J.T., Sharma-Kuinkel, B.K., Fowler, V.G., 2016. Impact of Bacterial and Human Genetic Variation on *Staphylococcus aureus* Infections. *PLoS Pathog.* 12, e1005330.

Michael, C.A., Dominey-Howes, D., Labbate, M., 2014. The Antimicrobial Resistance Crisis: Causes, Consequences, and Management. *Front. Public Heal.* 2, 145.

Miller, R.R.M., Price, J.R.R., Batty, E.M.M., Didelot, X., Wyllie, D., Golubchik, T., Crook, D.W.W., Paul, J., Peto, T.E.A.E.A., Wilson, D.J.J., Cule, M., Ip, C.L.C.L.C., Day, N.P.J.P.J., Moore, C.E.E., Bowden, R., Llewelyn, M.J.J., 2014. Healthcare-associated outbreak of meticillin-resistant *Staphylococcus aureus* bacteraemia: Role of a cryptic variant of an epidemic clone. *J. Hosp. Infect.* 86, 83–89.

Monecke, S., Coombs, G., Shore, A.C., Coleman, D.C., Akpaka, P., Borg, M., Chow, H., Ip, M., Jatzwauk, L., Jonas, D., Kadlec, K., Kearns, A., Laurent, F., O'Brien, F.G., Pearson, J., Ruppelt, A., Schwarz, S., Scicluna, E., Slickers, P., Tan, H.L., Weber, S., Ehricht, R., 2011. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. *PLoS One* 6, e17936.

Monecke, S., Gavier-Widén, D., Hotzel, H., Peters, M., Guenther, S., Lazaris, A., Loncaric, I., Müller, E., Reissig, A., Ruppelt-Lorz, A., Shore, A.C., Walter, B., Coleman, D.C., Ehricht, R., 2016. Diversity of *Staphylococcus aureus* isolates in european wildlife. *PLoS One* 11, e0168433.

Moran, G.J., Amii, R.N., Abrahamian, F.M., Talan, D.A., 2005. Methicillin-resistant *Staphylococcus aureus* in Community-acquired Skin Infections. *Emerg. Infect. Dis.* 11, 928–930.

Morand, S., McIntyre, K.M., Baylis, M., 2014. Domesticated animals and human infectious diseases of zoonotic origins: Domestication time matters. *Infect. Genet. Evol.* 24, 76–81.

Morikawa, K., Inose, Y., Okamura, H., Maruyama, A., Hayashi, H., Takeyasu, K., Ohta, T., 2003. A new staphylococcal sigma factor in the conserved gene cassette: Functional significance and implication for the evolutionary processes. *Genes to Cells* 8, 699–712.

Morse, S.S., 1995. Factors in the emergence of infectious diseases. *Emerg. Infect. Dis.* 1, 7–15.

Mostowy, R., Croucher, N.J., Andam, C.P., Corander, J., Hanage, W.P., Marttinen, P., 2017. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Mol. Biol. Evol. 3*4, 1167–1182.

Mothershed, E.A., Whitney, A.M., 2006. Nucleic acid-based methods for the detection of bacterial pathogens: Present and future considerations for the clinical laboratory. *Clin. Chim. Acta* 363, 206–220.

Murray, S., Pascoe, B., Méric, G., Mageiros, L., Yahara, K., Hitchings, M.D., Friedmann, Y., Wilkinson, T.S., Gormley, F.J., Mack, D., Bray, J.E., Lamble, S., Bowden, R., Jolley, K.A., Maiden, M.C.J., Wendlandt, S., Schwarz, S., Corander, J., Fitzgerald, J.R., Sheppard, S.K., 2017. Recombination-Mediated Host Adaptation by Avian *Staphylococcus aureus*. *Genome Biol. Evol.* 9, 830–842.

Nadalin, F., Vezzi, F., Policriti, A., 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics* 13, S8.

Naimi, T.S., Ledell, K.H., Como-sabetti, K., Borchardt, S.M., Boxrud, D.J., Johnson, S.K., Fridkin, S., Boyle, C.O., Danila, R.N., 2003. and Health Care – Associated *Staphylococcus aureus Infection* 290.

Narra, H.P., Ochman, H., 2006. Of What Use Is Sex to Bacteria? *Curr. Biol.* 16, R705–R710.

Nielsen, R., 2005. Molecular Signatures of Natural Selection. *Annu. Rev. Genet*. 39, 197–218.

Nilsson, P., Ripa, T., 2006. *Staphylococcus aureus* throat colonization is more frequent than colonization in the anterior nares. *J. Clin. Microbiol.* 44, 3334–3339.

Novick, R.P., Subedi, A., 2007. The SaPIs: Mobile pathogenicity islands of *Staphylococcus*. *Chem. Immunol. Allergy* 93, 42–57.

Nübel, U., Dordel, J., Kurt, K., Strommenger, B., Westh, H., Shukla, S.K., Zemlickova, H., Leblois, R., Wirth, T., Jombart, T., Balloux, F., Witte, W., 2010. A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog.* 6, e1000855.

Nübel, U., Nachtnebel, M., Falkenhorst, G., Benzler, J., Hecht, J., Kube, M., Bröcker, F., Moelling, K., Bührer, C., Gastmeier, P., Piening, B., Behnke, M., Dehnert, M., Layer, F., Witte, W., Eckmanns, T., 2013. MRSA Transmission on a Neonatal Intensive Care Unit: Epidemiological and Genome-Based Phylogenetic Analyses. *PLoS One* 8, e54898.

Nuermberger, E.L., Bishai, W.R., 2004. Antibiotic resistance in *Streptococcus pneumoniae*: what does the future hold? *Clin. Infect. Dis.* 38 Suppl 4, S363–S371.

O'Brien, T., 1995. WHONET: An Information System for Monitoring Antimicrobial Resistance. *Emerg. Infect. Dis.* 1, 66–66.

O'Connor, B.A., Carman, J., Eckert, K., Tucker, G., Givney, R., Cameron, S., 2007. Does using potting mix make you sick? Results from a *Legionella longbeachae* case-control study in South Australia. *Epidemiol. Infect.* 135, 34–39.

O'Driscoll, T., Crank, C.W., 2015. Vancomycin-resistant enterococcal infections:

epidemiology, clinical manifestations, and optimal management. *Infect. Drug Resist*. 8, 217–30.

O'Riordan, K., Lee, J.C., 2004. *Staphylococcus aureus* capsular polysaccharides. *Clin. Microbiol. Rev.* 17, 218–34.

Ochman, H., Lawrence, J.G., Groisman, E. a, 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304.

Odds, F.C., Jacobsen, M.D., 2008. Multilocus sequence typing of pathogenic *Candida* species. *Eukaryot. Cell* 7, 1075–1084.

Omoe, K., Hu, D.L., Takahashi-Omoe, H., Nakane, A., Shinagawa, K., 2003. Identification and characterization of a new staphylococcal enterotoxin-related putative toxin encoded by two kinds of plasmids. *Infect. Immun.* 71, 6088–6094.

Orsi, R.H., Sun, Q., Wiedmann, M., 2008. Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol.* Biol. 8, 233.

Otter, J.A., French, G.L., 2010. Molecular epidemiology of community-associated meticillin-resistant *Staphylococcus aureus* in Europe. *Lancet Infect. Dis.* 10, 227–239.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693.

Palmer, S.R., Soulsby, E.J.L., Simpson, D.I.H. 1998. Zoonoses: biology clinical practice, and public health control. Oxford University Press, New York.

Paterson, G.K., Harrison, E.M., Murray, G.G.R., Welch, J.J., Warland, J.H., Holden, M.T.G., Morgan, F.J.E., Ba, X., Koop, G., Harris, S.R., Maskell, D.J., Peacock, S.J., Herrtage, M.E., Parkhill, J., Holmes, M.A., 2015. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. *Nat. Commun.* 6, 6560.

Peng, B., Amos, C.I., Kimmel, M., 2007. Forward-time simulations of human populations with complex diseases. PLoS Genet. 3, 0407–0420.

Peng, B., Kimmel, M., 2007. Simulations provide support for the common disease-common variant hypothesis. *Genetics* 175, 763–776.

Pérez-Losada, M., Browne, E.B., Madsen, A., Wirth, T., Viscidi, R.P., Crandall, K.A., 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect. Genet. Evol.* 6, 97–112.

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., Crandall, K.A., 2013. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Genet. Evol.* 16, 38–53.

Peschel, A., Jack, R.W., Otto, M., Collins, L. V, Staubitz, P., Nicholson, G., Kalbacher, H., Nieuwenhuizen, W.F., Jung, G., Tarkowski, A., van Kessel, K.P., van Strijp, J.A., 2001. *Staphylococcus aureus* resistance to human defensins and evasion of neutrophil killing via the novel virulence factor MprF is based on modification of membrane lipids with l-lysine. *J. Exp. Med.* 193, 1067–76.

Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., Nielsen, R., 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* 17, 1336–1343.

Peton, V., Le Loir, Y., 2014. *Staphylococcus aureus* in veterinary medicine. Infect. *Genet. Evol.* 21, 602–615.

Potts, A., Donaghy, M., Marley, M., Othieno, R., Stevenson, J., Hyland, J., Pollock, K.G., Lindsay, D., Edwards, G., Hanson, M.F., Helgason, K.O., 2013. Cluster of Legionnaires' disease cases caused by *Legionella longbeachae* serogroup 1, Scotland, August to September 2013. *Eurosurveillance* 18, 20656.

Power, R.A., Parkhill, J., de Oliveira, T., 2017. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet. 1*8, 41–50.

Prasad, L.B., Newbould, F.H., 1968. Arylesterase activity of milk from normal and experimentally infected bovine mammary glands. *Can. Vet. J.* 9, 230–236.

Pravinkumar, S.J., Edwards, G., Lindsay, D., Redmond, S., Stirling, J., House, R., Kerr, J., Anderson, E., Breen, D., Blatchford, O., McDonald, E., Brown, A., 2010. A cluster of Legionnaires' disease caused by *Legionella longbeachae* linked to potting compost in Scotland, 2008-2009. *Eurosurveillance* 15, 1–3.

Price, E.P., Sarovich, D.S., Mayo, M., 2013. Within-Host Evolution of *Burkholderia pseudomallei* over a Twelve-Year Chronic Carriage Infection. *MBio* 4(4): e00388-13

Price, J.R., Golubchik, T., Cole, K., Wilson, D.J., Crook, D.W., Thwaites, G.E., Bowden, R., Walker, A.S., Peto, T.E.A., Paul, J., Llewelyn, M.J., 2014. Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of *Staphylococcus aureus* in an intensive care unit. *Clin. Infect. Dis.* 58, 609–618.

Price, L.B., Stegger, M., Hasman, H., Aziz, M., Larsen, J., Andersen, P.S., Pearson, T., Waters, A.E., Foster, J.T., Schupp, J., Gillece, J., Driebe, E., Liu, C.M., Springer, B., Zdovc, I., Battisti, A., Franco, A., Zmudzki, J., Schwarz, S., Butaye, P., Jouy, E., Pomba, C., Porrero, M.C., Ruimy, R., Smith, T.C., Robinson, D.A., Weese, J.S., Arriola, C.S., Yu, F., Laurent, F., Keim, P., Skov, R., Aarestrup, F.M., 2012. *Staphylococcus aureus* CC398: Host adaptation and emergence of methicillin resistance in livestock. *MBio* 3, 1–6.

Prugnolle, F., de Meeûs, T., 2011. Clonal Evolution. In: Genetics and Evoltuion of Infectious Diseases. pp. 133–146.

Purves, J., Blades, M., Arafat, Y., Malik, S.A., Bayliss, C.D., Morrissey, J.A., 2012. Variation in the genomic locations and sequence conservation of STAR elements among staphylococcal species provides insight into DNA repeat evolution. *BMC Genomics* 13.

Quebbemann, A.J., Rennick, B.R., 1968. Catechol transport by the renal tubule in the chicken. *Am. J. Physiol.* 214, 1201–4.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T.J., Hawkey, P., Loman, N.J., 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16, 114.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

Rao, C., Benhabib, H., Ensminger, A.W., 2013. Phylogenetic Reconstruction of the *Legionella pneumophila* Philadelphia-1 Laboratory Strains through Comparative Genomics. *PLoS One* 8, e64129.

Raphael, B.H., Baker, D.J., Nazarian, E., Lapierre, P., Bopp, D., Kozak-Muiznieks, N.A., Morrison, S.S., Lucas, C.E., Mercante, J.W., Musser, K.A., Winchell, J.M., 2016. Genomic resolution of outbreak-associated *Legionella pneumophila* serogroup 1 isolates from New York State. *Appl. Environ. Microbiol.* 82, 3582–3590.

Rasmussen, D.A., Volz, E.M., Koelle, K., 2014. Phylodynamic Inference for Structured

Epidemiological Models. *PLoS Comput. Biol.* 10, e1003570.

Ratcliff, R.M., Lanser, J.A., Manning, P.A., Heuzenroeder, M.W., 1998. Sequence-based classification scheme for the genus *Legionella* targeting the mip gene. J. *Clin. Microbiol.* 36, 1560–1567.

Reingold, A., 1998. Outbreak Investigations—A Perspective. *Emerg. Infect. Dis.* 4, 21–27.

Reller, L.B., Weinstein, M.P., Petti, C.A., 2007. Detection and Identification of Microorganisms by Gene Amplification and Sequencing. *Clin. Infect. Dis.* 44, 1108–1114.

Resch, G., François, P., Morisset, D., Stojanov, M., Bonetti, E.J., Schrenzel, J., Sakwinska, O., Moreillon, P., 2013. Human-to-Bovine Jump of *Staphylococcus aureus* CC8 Is Associated with the Loss of a β-Hemolysin Converting Prophage and the Acquisition of a New Staphylococcal Cassette Chromosome. *PLoS One* 8, e58187.

Reuter, S., Estee Torok, M., Holden, M.T.G., Reynolds, R., Raven, K.E., Blane, B., Donker, T., Bentley, S.D., Aanensen, D.M., Grundmann, H., Feil, E.J., Spratt, B.G., Parkhill, J., Peacock, S.J., 2016. Building a genomic framework for prospective MRSA surveillance in the United Kingdom and the Republic of Ireland. *Genome Res.* 26, 263–270.

Reuter, S., Harrison, T.G., Köser, C.U., Ellington, M.J., Smith, G.P., Parkhill, J., Peacock, S.J., Bentley, S.D., Török, M.E., 2013. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. *BMJ Open* 3, e002175.

Richardson, J.F., Reith, S., 1993. Characterization of a strain of methicillin-resistant *Staphylococcus aureus* (EMRSA-15) by conventional and molecular methods. *J. Hosp. Infect.* 25, 45–52.

Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* 106, 19126–31.

Robinson, E.R., Walker, T.M., Pallen, M.J., 2013. Genomics and outbreak investigation: from sequence to consequence. *Genome Med.* 5, 36.

Rohmer, L., Hocquet, D., Miller, S.I., 2011. Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol.* 19, 341–348.

Rooijakkers, S.H., Ruyken, M., Roos, a, Daha, M.R., Presanis, J.S., Sim, R.B., van Wamel, W.J., van Kessel, K.P., van Strijp, J. a, 2005. Immune evasion by a staphylococcal complement inhibitor that acts on C3 convertases. *Nat.Immunol.* 6, 920–927.

Rosenthal, K.L., Shofer, F.S., Rankin, S.C., 2009. Evaluation of mucosal and seborrheic sites for staphylococci in two populations of captive psittacines. *J. Am. Vet. Med. Assoc.* 234, 901–905.

Sakwinska, O., Giddey, M., Moreillon, M., Morisset, D., Waldvogel, A., Moreillon, P., 2011. *Staphylococcus aureus* host range and human-bovine host shift. A*ppl. Environ. Microbiol.* 77, 5908–5915.

Sánchez-Busó, L., Comas, I., Jorques, G., González-Candelas, F., 2014. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. *Nat. Genet.* 46, 1205–1211.

Sasaki, T., Tsubakishita, S., Tanaka, Y., Ohtsuka, M., Hongo, I., Fukata, T., Kabeya, H., Maruyama, S., Hiramatsu, K., 2012. Population genetic structures of *Staphylococcus aureus* isolates from cats and dogs in Japan. J. *Clin. Microbiol.* 50, 2152–2155.

Sass, P., Berscheid, A., Jansen, A., Oedenkoven, M., Szekat, C., Strittmatter, A., Gottschalk, G., Bierbaum, G., 2012. Genome sequence of *Staphylococcus aureus* VC40, a

vancomycin- and daptomycin-resistant strain, to study the genetics of development of resistance to currently applied last-resort antibiotics. *J. Bacteriol.* 194, 2107–2108.

Schaumburg, F., Mugisha, L., Peck, B., Becker, K., Gillespie, T.R., Peters, G., Leendertz, F.H., 2012. Drug-Resistant Human *Staphylococcus aureus* in Sanctuary Apes Pose a Threat to Endangered Wild Ape Populations. *Am. J. Primatol.* 74, 1071–1075.

Seemann, T., 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069.

Segal, G., Purcell, M., Shuman, H.A., 1998. Host cell killing and bacterial conjugation require overlapping sets of genes within a 22-kb region of the *Legionella pneumophila* genome. *Proc. Natl. Acad. Sci.* 95, 1669–1674.

Seidl, K., Stucki, M., Ruegg, M., Goerke, C., Wolz, C., Harris, L., Berger-Bächi, B., Bischoff, M., 2006. *Staphylococcus aureus* CcpA affects virulence determinant production and antibiotic resistance. Antimicrob. *Agents Chemother.* 50, 1183–1194.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., Whittam, T.S., 1986. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol*. 51, 873–884.

Sellers, T.A., Weaver, T.W., Phillips, B., Altmann, M., Rich, S.S., 1998. Environmental factors can confound identification of a major gene effect: Results from a segregation analysis of a simulated population of lung cancer families. *Genet. Epidemiol*. 15, 251–262.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B.B., Ideker, T., Owen Ozier, 2, Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B.B., Ideker, T., 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504.

Shearer, J.E.S., Wireman, J., Hostetler, J., Forberger, H., Borman, J., Gill, J., Sanchez, S., Mankin, A., Lamarre, J., Lindsay, J.A., Bayles, K., Nicholson, A., O'Brien, F., Jensen, S.O., Firth, N., Skurray, R.A., Summers, A.O., 2011. Major families of multiresistant plasmids from geographically and epidemiologically diverse staphylococci. *G3*. 1, 581–91.

Shepheard, M.A., Fleming, V.M., Connor, T.R., Corander, J., Feil, E.J., Fraser, C., Hanage, W.P., 2013. Historical Zoonoses and Other Changes in Host Tropism of *Staphylococcus aureus*, Identified by Phylogenetic Analysis of a Population Dataset. *PLoS One* 8, e62369.

Sheppard, S.K., Didelot, X., Meric, G., Torralbo, A., Jolley, K. a, Kelly, D.J., Bentley, S.D., Maiden, M.C.J., Parkhill, J., Falush, D., 2013. Genome-wide association study identi fi es vitamin B 5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl. Acad. Sci*. 110, 1–5.

Sheppard, S.K., McCarthy, N.D., Falush, D., Maiden, M.C.J., 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* 320, 237–239.

Shin, S.S., Pasechnikov, A.D., Gelmanova, I.Y., Peremitin, G.G., Strelis, A.K., Andreev, Y.G., Golubchikova, V.T., Tonkel, T.P., Yanova, G. V., Nikiforov, M., Yedilbayev, A., Mukherjee, J.S., Furin, J.J., Barry, D.J., Farmer, P.E., Rich, M.L., Keshavjee, S., 2006. Treatment outcomes in an integrated civilian and prison MDR-TB treatment program in Russia. *Int. J. Tuberc. Lung Dis*. 10, 402–408.

Siboo, I.R., Chaffin, D.O., Rubens, C.E., Sullam, P.M., 2008. Characterization of the accessory sec system of *Staphylococcus aureus*. *J. Bacteriol.* 190, 6188–6196.

Singhal, N., Kumar, M., Kanaujia, P.K., Virdi, J.S., 2015. MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. *Front. Microbiol*. 6, 791.

Sitkiewicz, I., Green, N.M., Guo, N., Mereghetti, L., Musser, J.M., 2011. Lateral gene transfer of streptococcal ICE element RD2 (region of difference 2) encoding secreted proteins. *BMC Microbiol.* 11, 65.

Smith, E.E., Smith, E.E., Buckley, D.G., Buckley, D.G., Wu, Z., Wu, Z., Saenphimmachak, C., Saenphimmachak, C., Hoffman, L.R., Hoffman, L.R., D'Argenio, D.A., D'Argenio, D.A., Miller, S.I., Miller, S.I., Ramsey, B.W., Ramsey, B.W., Speert, D.P., Speert, D.P., Moskowitz, S.M., Moskowitz, S.M., Burns, J.L., Burns, J.L., Kaul, R., Kaul, R., Olson, M. V, Olson, M. V, 2006. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci.* 103, 8487–92.

Smith, E.M., Green, L.E., Medley, G.F., Bird, H.E., Fox, L.K., Schukken, Y.H., Kruze, J. V., Bradley, A.J., Zadoks, R.N., Dowson, C.G., 2005. Multilocus sequence typing of intercontinental bovine *Staphylococcus aureus* isolates. *J. Clin. Microbiol.* 43, 4737–4743.

Smith, J.M., Smith, N.H., O'Rourke, M., Spratt, B.G., 1993. How clonal are bacteria? *Proc.Natl.Acad.Sci.* 90, 4384–4388.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T., 2011. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* 27, 431–432.

Smyth, D.S., Feil, E.J., Meaney, W.J., Hartigan, P.J., Tollersrud, T., Fitzgerald, J.R., Enright, M.C., Smyth, C.J., 2009. Molecular genetic typing reveals further insights into the diversity of animal-associated *Staphylococcus aureus*. *J. Med. Microbiol*. 58, 1343–1353.

Smyth, D.S., Robinson, D.A., 2009. Integrative and sequence characteristics of a novel genetic element, ICE6013, in *Staphylococcus aureus*. *J. Bacteriol*. 191, 5964–5975.

Sniegowski, P.D., Gerrish, P.J., 2010. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365, 1255–1263.

Snitkin, E.S., Zelazny, A., Thomas, P., Stock, F., 2012. Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Sci. Transl. Med*. 4, 148ra116.

Søndberg, E., Jelsbak, L., 2016. *Salmonella* Typhimurium undergoes distinct genetic adaption during chronic infections of mice. *BMC Microbiol*. 16, 30.

Song, J.H., Hsueh, P.R., Chung, D.R., Ko, K.S., Kang, C.I., Peck, K.R., Yeom, J.S., Kim, S.W., Chang, H.H., Kim, Y.S., Jung, S.I., Son, J.S., Man-Kit So, T., Lalitha, M.K., Yang, Y., Huang, S.G., Wang, H., Lu, Q., Carlos, C.C., Perera, J.A., Chiu, C.H., Liu, J.W., Chongthaleong, A., Thamlikitkul, V., Hung Van, P., 2011. Spread of methicillin-resistant *Staphylococcus aureus* between the community and the hospitals in Asian countries: An ANSORP study. *J. Antimicrob. Chemother*. 66, 1061–1069.

Soyer, Y., Orsi, R.H., Rodriguez-Rivera, L.D., Sun, Q., Wiedmann, M., 2009. Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes. *BMC Evol. Biol.* 9, 264.

Spoor, L.E., McAdam, P.R., Weinert, L.A., Rambaut, A., Hasman, H., Aarestrup, F.M., Kearns, A.M., Larsen, A.R., Skov, R.L., Ross Fitzgerald, J., 2013. Livestock origin for a human pandemic clone of community-associated methicillin-resistant *Staphylococcus aureus*. *MBio* 4, 1–6.

Spoor, L.E., Richardson, E., Richards, A.C., Wilson, G.J., Mendonca, C., Gupta, R.K., McAdam, P.R., Nutbeam-Tuffs, S., Black, N.S., O'Gara, J.P., Lee, C.Y., Corander, J., Fitzgerald, J.R., 2015. Recombination-mediated remodelling of host-pathogen interactions during *Staphylococcus aureus* niche adaptation. *Microb. genomics* 1, e000036.

Spratt, B.G., 2004. Exploring the Concept of Clonality in Bacteria. In: Genomics, Proteomics, and Clinical Bacteriology. Humana Press, Totowa, NJ, pp. 323–352.

Spratt, B.G., Maiden, M.C., 1999. Bacterial population genetics, evolution and epidemiology. P*hilos. Trans. R. Soc. Lond. B. Biol. Sci*. 354, 701–10.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690.

Steele, T.W., Lanser, J., Sangster, N., 1990. Isolation of *Legionella longbeachae* serogroup 1 from potting mixes. Appl. *Environ. Microbiol.* 56, 49–53.

Stefani, S., Chung, D.R., Lindsay, J.A., Friedrich, A.W., Kearns, A.M., Westh, H., MacKenzie, F.M., 2012. Meticillin-resistant *Staphylococcus aureus* (MRSA): Global epidemiology and harmonisation of typing methods. *Int. J. Antimicrob. Agents* 39, 273–282.

Stefani, S., Goglio, A., 2010. Methicillin-resistant *Staphylococcus aureus*: related infections and antibiotic resistance. *Int. J. Infect. Dis.* 14, S17–S20.

Stoesser, N., Sheppard, A.E., Shakya, M., Sthapit, B., Thorson, S., Giess, A., Kelly, D., Pollard, A.J., Peto, T.E.A., Walker, A.S., Crook, D.W., 2014. 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.

Strachan, N. J. C., Rotariu, O., Lopes, B., Macrae, M., Fairley, S., Laing, C., Gannon, V., Allison, L.J., Hanson, M.F., Dallman, T., Ashton, P., Franz, E., van Hoek, A.H.A.M., French, N.P., George, T., Biggs, P.J., Forbes, K. J., 2015. Whole Genome Sequencing demonstrates that Geographic Variation of *Escherichia coli* O157 Genotypes Dominates Host Association. *Scientific Reports*. 5, 14145.

Stranger, B.E., Stahl, E.A., Raj, T., 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 367–383.

Stryjewski, M.E., Corey, G.R., 2014. Methicillin-resistant *Staphylococcus aureus*: An evolving pathogen. *Clin. Infect. Dis.* 58, S10–9.

Su, F., Ou, H., Tao, F., Tang, H., Xu, P., 2013. PSP: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. *BMC Genomics.* 141 14, D448D454.

Sung, J.M.L., Lloyd, D.H., Lindsay, J.A., 2008. *Staphylococcus aureus* host specificity: Comparative genomics of human versus animal isolates by multi-strain microarray. *Microbiology* 154, 1949–1959.

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., 2011. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6, e21800.

Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 34, W609–12.

Svarrer, C.W., Uldum, S.A., 2012. The occurrence of *Legionella* species other than *Legionella pneumophila* in clinical and environmental samples in Denmark identified by *mip* gene

sequencing and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Clin. Microbiol. Infect.* 18, 1004–1009.

Takuno, S., Kado, T., Sugino, R.P., Nakhleh, L., Innan, H., 2012. Population genomics in bacteria: A case study of *Staphylococcus aureus*. *Mol. Biol. Evol.* 29, 797–809.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.* 24, 1596–1599.

Tang, P., Gardy, J.L., 2014. Stopping outbreaks with real-time genomic epidemiology. *Genome Med.* 6, 104.

Taylor, L.H., Latham, S.M., Woolhouse, M., 2001. Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*. 356, 983–9.

Tenover, F.C., Arbeit, R.D., Goering, R. V, 1997. How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Molecular Typing Working Group of the Society for Healthcare Epidemiology of America. Infect. Control *Hosp. Epidemiol.* 18, 426–439.

Thornton, T., McPeek, M.S., 2010. ROADTRIPS: Case-Control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. Am. *J. Hum. Genet*. 86, 172–184.

Tibayrenc, M., Ayala, F.J., 2012. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proc. Natl. Acad. Sci.* 109, E3305–13.

Toft, C., Andersson, S.G.E., 2010. Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet*. 11, 465–475.

Tognotti, E., 2013. Lessons from the History of Quarantine, from Plague to Influenza A. *Emerg. Infect. Dis*. 19, 254–259.

Tomley, F.M., Shirley, M.W., 2009. Livestock infectious diseases and zoonoses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 364, 2637–42.

Tong, S.Y.C., Davis, J.S., Eichenberger, E., Holland, T.L., Fowler, V.G., 2015. *Staphylococcus aureus* infections: Epidemiology, pathophysiology, clinical manifestations, and management. *Clin. Microbiol. Rev.* 28, 603–661.

Tong, S.Y.C., Schaumburg, F., Ellington, M.J., Corander, J., Pichon, B., Leendertz, F., Bentley, S.D., Parkhill, J., Holt, D.C., Peters, G., Giffard, P.M., 2015. Novel staphylococcal species that form part of a *Staphylococcus aureus*-related complex: the non-pigmented *Staphylococcus argenteus sp. nov.* and the non-human primate-associated *Staphylococcus schweitzeri sp. nov. Int. J. Syst. Evol. Microbiol.* 65, 15–22.

Toprak, E., Veres, A., Michel, J.B., Chait, R., Hartl, D.L., Kishony, R., 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet*. 44, 101–105.

Török, M.E., Harris, S.R., Cartwright, E.J.P., Raven, K.E., Brown, N.M., Allison, M.E.D., Greaves, D., Quail, M.A., Limmathurotsakul, D., Holden, M.T.G., Parkhill, J., Peacock, S.J., 2014. Zero tolerance for healthcare-associated MRSA bacteraemia: is it realistic? J. *Antimicrob. Chemother.* 69, 2238–45.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111.

Treangen, T.J., Rocha, E.P.C., 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* 7, e1001284.

Tsai, M., Ohniwa, R.L., Kato, Y., Takeshita, S.L., Ohta, T., Saito, S., Hayashi, H., Morikawa, K., 2011. *Staphylococcus aureus* requires cardiolipin for survival under conditions of high salinity. *BMC Microbiol*. 11, 13.

Tuite, A.R., Tien, J., Eisenberg, M., Earn, D.J.D., Ma, J., Fisman, D.N., 2011. Cholera epidemic in Haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. *Ann. Intern. Med*. 154, 593–601.

Tung, H. s, Guss, B., Hellman, U., Persson, L., Rubin, K., Rydén, C., Ryden, C., 2000. A bone sialoprotein-binding protein from *Staphylococcus aureus*: a member of the staphylococcal Sdr family. *Biochem.J*. 345 Pt 3, 611–619.

Turabelidze, D., Kotetishvili, M., Kreger, A., Morris, J.G., Sulakvelidze, A., 2000. Improved pulsed-field gel electrophoresis for typing vancomycin-resistant enterococci. *J. Clin. Microbiol.* 38, 4242–4245.

Turner, K.M.E., Feil, E.J., 2007. The secret life of the multilocus sequence type. Int. J. Antimicrob. Agents 29, 129–135.

Ubeda, C., Barry, P., Penadés, J.R., Novick, R.P., 2007. A pathogenicity island replicon in *Staphylococcus aureus* replicates as an unstable plasmid. *Proc. Natl. Acad. Sci.* 104, 14182–8.

Úbeda, C., Tormo, M.Á., Cucarella, C., Trotonda, P., Foster, T.J., Lasa, Í., Penadés, J.R., 2003. Sip, an integrase protein with excision, circularization and integration activities, defines a new family of mobile *Staphylococcus aureus* pathogenicity islands. *Mol. Microbiol*. 49, 193–210.

Uhlemann, A.C., Porcella, S.F., Trivedi, S., Sullivan, S.B., Hafer, C., Kennedy, A.D., Barbian, K.D., Mccarthy, A.J., Street, C., Hirschberg, D.L., Lipkin, W.I., Lindsay, J.A., Deleo, F.R., Lowy, F.D., 2012. Identification of a highly transmissible animal-independent *Staphylococcus aureus* ST398 clone with distinct genomic and cell adhesion properties. *MBio* 3, e00027–12.

Underwood, A.P., Jones, G., Mentasti, M., Fry, N.K., Harrison, T.G., 2013. Comparison of the Legionella pneumophila population structure as determined by sequence-based typing and whole genome sequencing. *BMC Microbiol.* 13, 302.

Van Cleef, B.A.G.L., Graveland, H., Haenen, A.P.J., Van De Giessen, A.W., Heederik, D., Wagenaar, J.A., Kluytmans, J.A.J.W., 2011. Persistence of livestock-associated methicillin-resistant *Staphylococcus aureus* in field workers after short-term occupational exposure to pigs and veal calves. *J. Clin. Microbiol.* 49, 1030–1033.

Van Duijkeren, E., Jansen, M.D., Flemming, S.C., De Neeling, H., Wagenaar, J.A., Schoormans, A.H.W., Van Nes, A., Fluit, A.C., 2007. Methicillin-resistant *Staphylococcus aureus* in pigs with exudative epidermitis. *Emerg. Infect. Dis.* 13, 1408–1410.

van der Woude, M. W. 2011. Phase variation: how to create and coordinate population diversity. *Curr. Opin. Microbiol.*, *14*(2), 205–211.

van Hal S. J., Ip C. L. C., Ansari A M., Wilson D. J., Espedido B. A., Jensen S. O., and B.R., 2015. Evolutionary dynamics of *Enterococcus faecium* reveals complex genomic relationships between isolates with independent emergence of vancomycin resistance. *Microb. Genomics* 2, 1–20.

van Wamel, W.J.B.B., Rooijakkers, S.H.M.M., Ruyken, M., Van Kessel, K.P.M.M., van Strijp, J.A.G.G., 2006. The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of *Staphylococcus aureus* are located on β-

hemolysin-converting bacteriophages. *J. Bacteriol.* 188, 1310–1315.

Vancraeynest, D., Haesebrouck, F., Deplano, A., Denis, O., Godard, C., Wildemauwe, C., Hermans, K., 2006. International dissemination of a high virulence rabbit *Staphylococcus aureus* clone. *J. Vet. Med. Ser. B Infect. Dis. Vet. Public Heal.* 53, 418–422.

Vannucci, F. A., Kelley, M. R., & Gebhart, C. J. 2013. Comparative genome sequencing identifies a prophage-associated genomic island linked to host adaptation of *Lawsonia intracellularis* infections. *Veterinary Research*, *44*(1), 49.

Velonakis, E.N., Kiousi, I.M., Koutis, C., Papadogiannakis, E., Babatsikou, F., Vatopoulos, A., 2010. First isolation of *Legionella* species, including *L. pneumophila* serogroup 1, in Greek potting soils: Possible importance for public health. *Clin. Microbiol. Infect. 1*6, 763–766.

Versalovic, J., Lupski, J.R., 2002. Molecular detection and genotyping of pathogens: More accurate and rapid answers. *Trends Microbiol.* 10.

Viana, D., Blanco, J., Tormo-Más, M.Á., Selva, L., Guinane, C.M., Baselga, R., Corpa, J.M., Lasa, Í., Novick, R.P., Fitzgerald, J.R., Penadés, J.R., 2010. Adaptation of *Staphylococcus aureus* to ruminant and equine hosts involves SaPI-carried variants of von Willebrand factor-binding protein. *Mol. Microbiol.* 77, 1583–1594.

Viana, D., Comos, M., McAdam, P.R., Ward, M.J., Selva, L., Guinane, C.M., González-Muñoz, B.M., Tristan, A., Foster, S.J., Fitzgerald, J.R., Penadés, J.R., 2015. A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet.* 47, 361–6.

Vinatzer, B.A., Monteil, C.L., Clarke, C.R., 2014. Harnessing Population Genomics to Understand How Bacterial Pathogens Emerge, Adapt to Crop Hosts, and Disseminate. A*nnu. Rev. Phytopathol.* 52, 19–43.

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24.

Visschers, V.H.M., Iten, D.M., Riklin, A., Hartmann, S., Sidler, X., Siegrist, M., 2014. Swiss pig farmers′ perception and usage of antibiotics during the fattening period. *Livest. Sci*. 162, 223–232.

Visser, J.A.G., Lenski, R.E., 2002. Long-term experimental evolution in *Escherichia coli* . XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evol. Biol.* 2, 1.

Volz, E.M., Kosakovsky Pond, S.L., Ward, M.J., Leigh Brown, A.J., Frost, S.D.W., 2009. Phylodynamics of infectious disease epidemics. *Genetics* 183, 1421–1430.

von Mentzer, A., Connor, T.R., Wieler, L.H., Semmler, T., Iguchi, A., Thomson, N.R., Rasko, D. a, Joffre, E., Corander, J., Pickard, D., Wiklund, G., Svennerholm, A.-M., Sjöling, A., Dougan, G., 2014. Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat. Genet*. 46, 1321–1326.

Voordeckers, K., Kominek, J., Das, A., Espinosa-Cant??, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., Yang, Y., Zhu, B., Marchal, K., DeLuna, A., Van Noort, V., Jelier, R., Verstrepen, K.J., 2015. Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLoS Genet.* 11, e1005635.

Voss, A., Loeffen, F., Bakker, J., Klaassen, C., Wulf, M., 2005. Methicillin-resistant *Staphylococcus aureus* in Pig Farming. *Emerg. Infect. Dis.* 11, 1965–1966.

Votintseva, A.A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., Chatterjee, A., Smith, E.G., Sanderson, N., Walker, T.M., Morgan, M.R., Wyllie, D.H., Walker, A.S., Peto, T.E.A., Crook, D.W., Iqbal, Z., 2017. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J. Clin. Microbio*l. 55, 1285–1298.

Voyich, J.M., Otto, M., Mathema, B., Braughton, K.R., Whitney, A.R., Welty, D., Long, R.D., Dorward, D.W., Gardner, D.J., Lina, G., Kreiswirth, B.N., DeLeo, F.R., 2006. Is Panton-Valentine Leukocidin the Major Virulence Determinant in Community-Associated Methicillin-Resistant *Staphylococcus aureus* Disease? *J. Infect. Dis*. 194, 1761–1770.

Walker, T.M., Ip, C.L.C., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., Eyre, D.W., Wilson, D.J., Hawkey, P.M., Crook, D.W., Parkhill, J., Harris, D., Walker, a S., Bowden, R., Monk, P., Smith, E.G., Peto, T.E.A., 2013. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* 13, 137–46.

Weese, J.S., 2010. Methicillin-resistant *Staphylococcus aureus* in animals. *Ilar* j 51, 233–244.

Weinert, L. a., Welch, J.J., Suchard, M. a., Lemey, P., Rambaut, a., Fitzgerald, J.R., 2012. Molecular dating of human-to-bovid host jumps by *Staphylococcus aureus* reveals an association with the spread of domestication. *Biol. Lett.* 8, 829–832.

Weinert, L.A., Chaudhuri, R.R., Wang, J., Peters, S.E., Corander, J., Jombart, T., Baig, A., Howell, K.J., Vehkala, M., Välimäki, N., Harris, D., Chieu, T.T.B., Van Vinh Chau, N., Campbell, J., Schultsz, C., Parkhill, J., Bentley, S.D., Langford, P.R., Rycroft, A.N., Wren, B.W., Farrar, J., Baker, S., Hoa, N.T., Holden, M.T.G., Tucker, A.W., Maskell, D.J., 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus sui*s. *Nat. Commun.* 6, 6740.

Weis, A.M., Storey, D.B., Taff, C.C., Townsend, A.K., Huang, B.C., Kong, N.T., Clothier, K.A., Spinner, A., Byrn, B.A., Weimer, B.C., 2016. Genomic comparison of *Campylobacter spp.* and their potential for zoonotic transmission between birds, primates, and livestock. *Appl. Environ. Microbiol*. 82, 7165–7175.

Wertheim, H.F.L., Melles, D.C., Vos, M.C., Van Leeuwen, W., Van Belkum, A., Verbrugh, H.A., Nouwen, J.L., 2005. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect. Dis.* 5, 751–762.

Whiley, H., Bentham, R., 2011. *Legionella longbeachae* and legionellosis. *Emerg. Infect. Dis*. 17, 579–583.

Whitaker, R.J., Banfield, J.F., 2006. Population genomics in natural microbial communities. *Trends Ecol. Evol.* 21, 508–516.

Wiethoelter, A.K., Beltrán-Alcrudo, D., Kock, R., Mor, S.M., 2015. Global trends in infectious diseases at the wildlife-livestock interface. *Proc. Natl. Acad. Sci*. 112, 1–6.

Wilson, D.J., 2012. Insights from Genomics into Bacterial Pathogen Populations. *PLoS Pathog.* 8, e1002874.

Wolfe, N.D., Dunavan, C.P., Diamond, J., 2007. Origins of major human infectious diseases. *Nature* 447, 279–283.

Woolhouse, M.E.J., 2001. Population Biology of Multihost Pathogens. *Science* (80). 292, 1109–1112.

Worby, C.J., Chang, H.H., Hanage, W.P., Lipsitch, M., 2014. The distribution of pairwise genetic distances: A tool for investigating disease transmission. *Genetics* 198, 1395–1404.

Worby, C.J., Lipsitch, M., Hanage, W.P., 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput. Biol.* 10, e1003549.

Worby, C.J., Read, T.D., 2015. "SEEDY" (Simulation of Evolutionary and Epidemiological Dynamics): An R package to follow accumulation of within-host mutation in pathogens. *PLoS One* 10, 1–14.

World Health Organization, 2017. World Health Statistics 2017 : Monitoring Health for The SDGs, World Health Organization.

Wozniak, M., Tiuryn, J., Wong, L., 2014. GWAMAR: Genome-wide assessment of mutations associated with drug resistance in bacteria. *BMC Genomics* 15 Suppl 1, S10.

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C.Y., Wei, L., 2011. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322.

Xu, Y., Zhu, Y., Wang, Y., Chang, Y.-F., Zhang, Y., Jiang, X., … Wang, J. (2016). Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic *Leptospira*. *Sci. Rep.*, *6*(1), 20020.

Xu, Z., Chen, H., Zhou, R., 2011. Genome-wide evidence for positive selection and recombination in *Actinobacillus pleuropneumoniae. BMC Evol. Biol*. 11, 203.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., Bork, P., 2011. IPath2.0: Interactive pathway explorer. *Nucleic Acids Res.* 39, W412–5.

Yang, L., Jelsbak, L., Marvig, R.L., Damkiær, S., Workman, C.T., Rau, M.H., Hansen, S.K., Folkesson, A., Johansen, H.K., Ciofu, O., Høiby, N., Sommer, M.O.A., Molin, S., 2011. Evolutionary dynamics of bacteria in a human host environment. *Proc. Natl. Acad. Sci.* 108, 7481–7486.

Yang, Z., 2006. Computational molecular evolution, Oxford: Oxford University Press. Oxford University Press.

Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 24, 1586–1591.

Yang, Z., Wong, W.S.W., Nielsen, R., 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118.

Yarwood, J.M., McCormick, J.K., Paustian, M.L., Orwin, P.M., Kapur, V., Schlievert, P.M., 2002. Characterization and expression analysis of *Staphylococcus aureus* pathogenicity island 3. Implications for the evolution of staphylococcal pathogenicity islands. *J. Biol. Chem.* 277, 13138–13147.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z., 2009. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871.

Young, B.C., Golubchik, T., Batty, E.M., Fung, R., Larner-Svensson, H., Votintseva, A.A., Miller, R.R., Godwin, H., Knox, K., Everitt, R.G., Iqbal, Z., Rimmer, A.J., Cule, M., Ip, C.L., Didelot, X., Harding, R.M., Donnelly, P., Peto, T.E., Crook, D.W., Bowden, R., Wilson, D.J., 2012. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc. Natl. Acad. Sci*. 109, 4550–4555.

Ypma, R.J.F.F., van Ballegooijen, W.M., Wallinga, J., 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195, 1055–1062.

Yu, V.L., Plouffe, J.F., Pastoris, M.C., Stout, J.E., Schousboe, M., Widmer, A., Summersgill,

J., File, T., Heath, C.M., Paterson, D.L., Chereshsky, A., 2002. Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. *J. Infect. Dis.* 186, 127–8.

Zdziarski, J., Brzuszkiewicz, E., Wullt, B., Liesegang, H., Biran, D., Voigt, B., Gronberg-Hernandez, J., Ragnarsdottir, B., Hecker, M., Ron, E.Z., Daniel, R., Gottschalk, G., Hacker, J., Svanborg, C., Dobrindt, U., 2010. Host imprints on bacterial genomes-rapid, divergent evolution in individual patients. *PLoS Pathog.* 6, 95–96.

Zerbino, D.R., 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. Curr. Protoc. Bioinforma. Chapter 11, Unit 11.5.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–9.

Zou, D., Kaneko, J., Narita, S., Kamio, Y., 2000. Prophage, phi PV83-pro, carrying panton-valentine leukocidin genes, on the *Staphylococcus aureus* p83 chromosome: Comparative analysis of the genome structures of phi PV83-pro, phi PVL, phi 11, and other phages. *Biosci. Biotechnol. Biochem*. 64, 2631–2643.

Appendix 1

Supplementary Table 1. Details of *Staphylococcus spp*. isolates used in this study.

This table can be found at: http://dx.doi.org/10.7488/ds/2330.

Citation: Bacigalupe, Rodrigo; Fitzgerald, Ross. (2018). Details of *Staphylococcus spp*. isolates used in a *Staphylococcus aureus* Host-Adaptation Study, [dataset]. University of Edinburgh. The Roslin Institute. http://dx.doi.org/10.7488/ds/2330.

Supplementary Table 2. Isolates selected for positive selection analysis.

CC45 Humans 1	CC45 Humans 2	CC45 Humans 3	CC59 Humans 1
434061	423817	423811	0864N0012
438615	423818	433081	0864N0014
4414311	438626	434051	0864N0015
441433	438674	438612	61223
4414510	441426	4395311	7229531
441458	441454	441466	7474263
613335	443011	441528	7474292
7068520	446554	7474282	M013
7229649	6133112	8113456	SA40
7396285	623614	CA347	SA957
ST30 Humans 1	**ST30 Humans 2**	**ST30 Humans 3**	**CC59 Humans 2**
435023	438625	433025	0864N0012
4386610	441413	4386212	0864N0014
438662	441421	441411	0864N0015
438675	441461	4414511	7068522
441424	441529	4414610	7229531
441453	458473	613311	7474263
443015	613318	613316	7474292
613315	7396260	7065830	M013
613334	8113481	8113425	SA40
MRSA252	8728565	8113466	SA957
CC5 Humans 1	**CC5 Humans 2**	**CC5 Humans 3**	**ST239 Humans 1**
433027	16035	18583	434063
434054	435018	433024	434064
4350112	4350211	433026	6133212
435011	438617	433083	613332
438687	4386510	435027	7065864
4395211	438673	441415	Bmb9393
441412	439527	613314	JKD6008
441428	6133110	6236111	T0131
441457	613319	7229516	TW20
4415311	JH1	JH9	Z172
CC15 Humans 1	**CC15 Humans 2**	**CC15 Humans 3**	**ST239 Humans 2**
435075	4350210	4350111	434063
441435	435024	438618	434064
4414612	435075	438683	438669
441469	441436	439523	612111
441522	441451	441431	6133212
443017	443012	441469	613332
443028	6133211	441522	7065864
623616	613331	441526	Bmb9393
7229330	7229329	613321	T0131
7229695	7229695	7068523	TW20

CC22 Humans 1	CC22 Humans 2	CC22 Humans 3	ST239 Humans 3
438658	4340512	4386710	434063
438672	4386811	4386711	434064
441423	439521	4386712	6133212
441524	439525	4395210	613332
441537	441455	441523	7065864
443019	4415212	441531	Bmb9393
443025	441533	441534	JKD6008
6236110	4430111	441535	T0131
7065844	458471	443022	TW20
8113592	7068517	613313	Z172
ST8 Humans 1	**ST8 Humans 2**	**ST8 Humans 3**	**CC12 Humans**
423816	433023	4238111	0864N0092
4330811	435019	423814	441418
433089	4386211	434066	441452
434062	4386512	443029	443024
438656	438657	4465510	623619
446552	4386611	789385	7229312
7229483	438665	8113414	7229488
7396136	7229692	8113435	7229536
8113414	7893810	8113529	8113449
Newman	COL	8113575	8113487
CC151 Cows 1	**CC151 Cows 2**	**973N0083**	**CC133 Ruminants 1**
7068527	7068527	CC151 Cows 3	61233
7068529	7068529	10900259	61240
7068538	7068531	7068529	61243
973N0058	7068538	7068531	61248
973N0061	973N0061	7068538	61249
973N0062	973N0062	973N0061	61252
973N0068	973N0068	973N0062	61258
973N0082	973N0076	973N0068	973N0009
973N0083	973N0082	973N0076	973N0048
RF122	973N0083	973N0082	ED133
CC97 Cows 1	**CC97 Cows 2**	**CC97 Cows 3**	**CC133 Ruminants 2**
10900246	10900246	52701	61234
52704	52705	80382	61235
52710	80379	80386	61237
7068528	80381	80388	61241
80382	80383	9119280	61243
80386	80386	973N0053	61244
80387	80388	973N0075	61249
973N0057	CHILE5	CHILE5	61258
973N0075	CTH54	CTH54	973N0010
LMA1166B	RC7	LMA1166B	ED133

CC398 Pigs 1	CC398 Pigs 2	CC398 Pigs 3	CC133 Ruminants 3
973N0044	973N0044	973N0044	61233
SRR445028	SRR445028	SRR445028	61234
SRR445034	SRR445034	SRR445034	61235
SRR445230	SRR445060	SRR445035	61237
SRR445239	SRR445236	SRR445060	61240
SRR445265	SRR445239	SRR445236	61243
SRR445266	SRR445266	SRR445237	973N0009
SRR445281	SRR445276	SRR445266	973N0010
SRR445286	SRR445281	SRR445286	973N0054
SRR445291	SRR445291	SRR445291	ED133
CC5 Birds 1	**CC5 Birds 2**	**CC5 Birds 3**	**CC385 Birds**
61267	61267	61267	61263
61270	61270	61270	61269
61271	61272	61274	61282
61274	61280	61280	61283
61280	61281	61281	61284
61281	61287	61288	61286
61287	61288	61289	61290
61288	61289	61291	8014464
61289	61291	973N0091	973N0056
973N0091	973N0091	ED98	CC385 Birds

Isolates selected for the lineages (CCs or STs) associated with specific hosts. For most of them triplicates were selected, sampling without replacement when possible. If the number of genomes available for the evolutionary lineage was not enough, two replicates without replacement (CC59 Humans) or only 10 isolates (CC385 Birds) had to be selected.

Supplementary Table 3. Details on the gene products enriched in different host-species using the k-mer approach in the panGWAS analysis.

Host	Type of element	ID	Product
Birds	Pathogenicity island	04494	transposon-related protein
Birds	Pathogenicity island	04495	hypothetical protein
Birds	Pathogenicity island	04496	hypothetical protein
Birds	Pathogenicity island	04497	transposon-related protein
Birds	Pathogenicity island	04498	transposon-related protein
Birds	Pathogenicity island	04499	transposon-related protein
Birds	Pathogenicity island	04500	pathogenicity island protein
Birds	Pathogenicity island	04501	phage-like protein
Birds	Pathogenicity island	04502	hypothetical protein
Birds	Pathogenicity island	04503	hypothetical protein
Birds	Pathogenicity island	04504	hypothetical protein
Birds	Pathogenicity island	04505	putative lipoprotein
Birds	Pathogenicity island	04506	transposon-related protein
Birds	Pathogenicity island	04507	hypothetical protein
Birds	Pathogenicity island	04508	hypothetical protein
Birds	Pathogenicity island	04509	hypothetical protein
Birds	Pathogenicity island	04510	transposon-related protein
Birds	Pathogenicity island	04511	transposon-related protein
Birds	Pathogenicity island	04512	transposon-related protein
Birds	Pathogenicity island	04513	ornithine cyclodeaminase
Birds	Pathogenicity island	04514	caax amino protease family protein; putative membrane
Birds	Pathogenicity island	04515	hypothetical protein
Birds	Pathogenicity island	04516	Helix-turn-helix domain
Birds	Pathogenicity island	04517	hypothetical protein
Birds	Pathogenicity island	04518	putative phage primase
Birds	Pathogenicity island	04519	putative phage DNA polymerase
Birds	Pathogenicity island	04520	hypothetical protein
Birds	Pathogenicity island	04522	Zn-dependent hydrolase
Birds	Pathogenicity island	04523	phi PVL orf 39-like protein
Birds	Pathogenicity island	04524	hypothetical protein
Birds	Pathogenicity island	04525	hypothetical protein
Birds	Pathogenicity island	04526	hypothetical protein
Birds	Pathogenicity island	04527	hypothetical protein
Birds	Pathogenicity island	04528	hypothetical protein
Birds	Pathogenicity island	04529	hypothetical protein
Birds	Pathogenicity island	04530	phi77 ORF017-like protein
Birds	Pathogenicity island	04531	hypothetical protein
Birds	Pathogenicity island	04532	hypothetical protein
Birds	Pathogenicity island	04533	hypothetical protein
Birds	Pathogenicity island	04534	hypothetical protein
Birds	Pathogenicity island	04535	primase C 1 family protein
Birds	Pathogenicity island	04536	antitoxin HipB

Host	Type of element	ID	Product
Birds	Pathogenicity island	04537	hypothetical protein
Birds	Pathogenicity island	04538	staphostatin
Birds	Pathogenicity island	04539	hypothetical protein
Birds	Pathogenicity island	04540	plasmid replication protein
Birds	Pathogenicity island	04541	chromosome partitioning ATPase
Birds	Pathogenicity island	04542	replication protein
Birds	Pathogenicity island	04543	hypothetical protein
Birds	Pathogenicity island	04544	hypothetical protein
Birds	Pathogenicity island	04545	caax amino protease family
Birds	Pathogenicity island	04546	pemK-like family protein
Birds	Pathogenicity island	04547	addiction module antidote
Birds	Pathogenicity island	04548	lysophospholipase
Birds	Pathogenicity island	04549	caax amino protease family
Birds	Bacteriophage	05178	prophage L54a, HNH endonuclease
Birds	Bacteriophage	05179	terminase small subunit
Birds	Bacteriophage	05180	terminase large subunit
Birds	Bacteriophage	05181	prophage L54a,
Birds	Bacteriophage	05182	prophage L54a, Clp protease
Birds	Bacteriophage	05183	bacteriophage capsid protein
Birds	Bacteriophage	05184	prophage L54a, DNA packaging protein
Birds	Bacteriophage	05185	SLT orf 110-like protein
Birds	Bacteriophage	05186	SLT orf 123-like protein
Birds	Bacteriophage	05187	Bacteriophage
Birds	Bacteriophage	05188	prophage L54a, major tail protein
Birds	Bacteriophage	05189	prophage L54a, major tail protein
Birds	Bacteriophage	05190	phiSLT ORF116b-like protein
Birds	Bacteriophage	05191	hypothetical protein
Birds	Bacteriophage	05192	prophage L54a, tail tape measure protein
Birds	Bacteriophage	05193	holin-like protein
Birds	Bacteriophage	05194	SLT orf 527-like protein
Birds	Bacteriophage	05195	phiSLT ORF96-like protein
Birds	Bacteriophage	05196	SLT orf 636-like protein
Birds	Bacteriophage	05197	phiSLT ORF488-like protein
Birds	Bacteriophage	05198	SLT orf 129-like protein
Birds	Bacteriophage	05199	phage-like protein
Birds	Bacteriophage	05200	phi SLT orf 99-like protein
Birds	Bacteriophage	05201	prophage L54a, holing
Humans	Bacteriophage	03599	beta-hemolysin
Humans	Bacteriophage	03600	complement inhibitor SCIN
Humans	Bacteriophage	03601	chemotaxis-inhibiting protein CHIPS
Humans	Bacteriophage	03602	truncated amidase
Humans	Bacteriophage	03603	staphylokinase precursor
Humans	Bacteriophage	03604	CHAP domain-containing protein
Humans	Bacteriophage	03605	phage phi LC3 family holin
Humans	Bacteriophage	03606	enterotoxin P
Humans	Bacteriophage	03607	hypothetical protein

Host	Type of element	ID	Product
Humans	Bacteriophage	03608	Bacteriophage
Humans	Bacteriophage	03609	phi PVL ORF 22-like protein
Humans	Bacteriophage	03610	phage protein
Humans	Bacteriophage	03611	phage minor structural protein
Humans	Bacteriophage	03612	phage protein
Humans	Bacteriophage	03613	TP901 family phage tail tape measure protein
Humans	Bacteriophage	03614	phi77 ORF100-like protein
Humans	Bacteriophage	03615	phage protein
Humans	Bacteriophage	03616	phi13 family phage major tail protein
Humans	Bacteriophage	03617	phage protein
Humans	Bacteriophage	03618	phage protein
Humans	Bacteriophage	03619	phage head-tail adaptor
Humans	Bacteriophage	03620	phage protein
Humans	Bacteriophage	03621	phage transcriptional terminator
Humans	Bacteriophage	03622	phi77 ORF006-like protein, capsid protein
Humans	Bacteriophage	03623	peptidase S14 ClpP
Humans	Bacteriophage	03624	Portal protein
Humans	Bacteriophage	03625	phage terminase
Humans	Bacteriophage	03626	phage protein
Humans	Bacteriophage	03627	phage-associated homing endonuclease
Humans	Bacteriophage	03628	phage transcriptional regulator, RinA
Humans	Bacteriophage	03629	phi PVL ORF 60-like protein
Humans	Bacteriophage	03630	transcriptional activator RinB
Humans	Bacteriophage	03631	phage-like protein
Humans	Bacteriophage	03632	phage protein
Humans	Bacteriophage	03633	dUTPase
Humans	Bacteriophage	03634	phi PVL ORF 52-like protein
Humans	Bacteriophage	03635	phage conserved Open Reading Frame 51
Humans	Bacteriophage	03636	phage-like protein
Humans	Bacteriophage	03637	PVL ORF-50 family protein
Humans	Bacteriophage	03638	endodeoxyribonuclease
Humans	Bacteriophage	03639	phage protein
Humans	Bacteriophage	03640	primosome subunit DnaD
Humans	Bacteriophage	03641	single-strand binding protein
Humans	Bacteriophage	03642	phage protein
Humans	Bacteriophage	03643	RecT protein
Humans	Bacteriophage	03644	phiPVL ORF41-like protein
Humans	Bacteriophage	03645	phage protein
Humans	Bacteriophage	03646	phage protein
Ruminants	Superantigens region	02135	putative membrane protein
Ruminants	Superantigens region	02136	FAD/NAD(P)-binding Rossmann fold superfamily protein
Ruminants	Superantigens region	02137	superantigen-like protein
Ruminants	Superantigens region	02138	superantigen-like protein
Ruminants	Superantigens region	02139	superantigen-like protein
Ruminants	Superantigens region	02140	superantigen-like protein

Host	Type of element	ID	Product
Ruminants	Superantigens region	02141	superantigen-like protein 5
Ruminants	Superantigens region	02142	superantigen-like protein
Ruminants	Superantigens region	02143	superantigen-like protein
Ruminants	Superantigens region	02144	superantigen-like protein
Ruminants	Superantigens region	02145	superantigen-like protein
Ruminants	Superantigens region	02146	type I restriction-modification system, M subunit
Ruminants	Superantigens region	02147	type I restriction-modification system S subunit
Ruminants	Superantigens region	02148	superantigen-like protein
Ruminants	Superantigens region	02149	hypothetical protein
Ruminants	Bacteriophage	04980	prophage L54a, antirepressor
Ruminants	Bacteriophage	04981	hypothetical protein
Ruminants	Bacteriophage	04982	hypothetical protein
Ruminants	Bacteriophage	04983	phi PV83 orf 27-like protein
Ruminants	Bacteriophage	04984	hypothetical protein
Ruminants	Bacteriophage	04985	hypothetical protein
Ruminants	Bacteriophage	04986	hypothetical protein
Ruminants	Bacteriophage	04987	RinA family phage transcriptional regulator
Ruminants	Bacteriophage	04988	small terminase
Ruminants	Bacteriophage	04989	phage terminase large subunit
Ruminants	Bacteriophage	04990	phage-like protein
Ruminants	Bacteriophage	04991	SPP1 family phage head morphogenesis protein
Ruminants	Bacteriophage	04992	hypothetical protein
Ruminants	Bacteriophage	04993	phage-like protein
Ruminants	Bacteriophage	04994	phage-related head protein
Ruminants	Bacteriophage	04995	phi Mu50B-like protein
Ruminants	Bacteriophage	04996	phi Mu50B-like protein
Ruminants	Bacteriophage	04997	phage-like protein
Ruminants	Bacteriophage	04998	HK97 family phage protein
Ruminants	Bacteriophage	04999	phage-like protein
Ruminants	Bacteriophage	05000	TP901-1 family phage major tail protein
Ruminants	Bacteriophage	05001	phage-like protein
Ruminants	Bacteriophage	05002	phage-like protein
Ruminants	Bacteriophage	05003	phage tape measure protein
Ruminants	Bacteriophage	05004	phi ETA orf 54-like protein
Ruminants	Bacteriophage	05005	phage minor structural protein
Ruminants	Bacteriophage	05006	phiETA ORF57-like protein
Ruminants	Bacteriophage	05007	SLT orf 129-like protein
Ruminants	Bacteriophage	05008	phage-related cell wall hydrolase
Ruminants	Bacteriophage	05009	tail fiber
Ruminants	Bacteriophage	05010	phi ETA orf 63-like protein
Ruminants	Bacteriophage	05011	phage phi LC3 family holin
Ruminants	Bacteriophage	05012	N-acetylmuramoyl-L-alanine amidase

Pangenome identifier (ID).

Supplementary Table 4. Details of *Staphylococcus aureus* isolates used in the experimental evolution study.

Strain	Isolate	Clonal lineage	Clone	Prev. Sheep	Sheep	d passaged	d in sheep	Description
OV1	OV2093	N315	1			0		original strain
OV2	OV2096	CH3657	1			0		original strain
CH3657	OV1903	C1	1	45210	88596	395	74	Human strain in sheep
CH3657	OV1904	C1	2	45210	88596	395	74	Human strain in sheep
CH3657	OV1905	C1	3	45210	88596	395	74	Human strain in sheep
CH3657	OV1918	C211	1	3111	18864	340	74	Human strain in sheep
CH3657	OV1919	C211	2	3111	18864	340	74	Human strain in sheep
CH3657	OV1920	C211	3	3111	18864	340	74	Human strain in sheep
CH3657	OV1933	C212	1	57544	20112	323	74	Human strain in sheep
CH3657	OV1934	C212	2	57544	20112	323	74	Human strain in sheep
CH3657	OV1935	C212	3	57544	20112	323	74	Human strain in sheep
CH3657	OV1923	C221	1	9912	12042	364	74	Human strain in sheep
CH3657	OV1924	C221	2	9912	12042	364	74	Human strain in sheep
CH3657	OV1925	C221	3	9912	12042	364	74	Human strain in sheep
CH3657	OV1928	C221	1	9912	19136	364	74	Human strain in sheep
CH3657	OV1929	C221	2	9912	19136	364	74	Human strain in sheep
CH3657	OV1930	C221	3	9912	19136	364	74	Human strain in sheep
CH3657	OV1938	C3	1	6639	6115	261	74	Human strain in sheep
CH3657	OV1939	C3	2	6639	6115	261	74	Human strain in sheep
CH3657	OV1940	C3	3	6639	6115	261	74	Human strain in sheep
CH3657	OV1948	C410	1	2025	18760	338	74	Human strain in sheep
CH3657	OV1949	C410	2	2025	18760	338	74	Human strain in sheep
CH3657	OV1950	C410	3	2025	18760	338	74	Human strain in sheep
CH3657	OV1953	C421	1	57993	19292	338	74	Human strain in sheep
CH3657	OV1956	C421	4	57993	19292	338	74	Human strain in sheep
CH3657	OV1955	C421	3	57993	19292	338	74	Human strain in sheep

Strain	Isolate	Clonal lineage	Clone	Prev. Sheep	Sheep	d passaged	d in sheep	Description
CH3657	OV1958	C422	1	57993	65789	338	74	Human strain in sheep
CH3657	OV1959	C422	2	57993	65789	338	74	Human strain in sheep
CH3657	OV1960	C422	3	57993	65789	338	74	Human strain in sheep
N315	OV1963	N111	1	3179	92675	289	74	Human strain in sheep
N315	OV1964	N111	2	3179	92675	289	74	Human strain in sheep
N315	OV1965	N111	3	3179	92675	289	74	Human strain in sheep
N315	OV1978	N1121	1	3685	92697	331	39	Human strain in sheep
N315	OV1979	N1121	2	3685	92697	331	39	Human strain in sheep
N315	OV1980	N1121	3	3685	92697	331	39	Human strain in sheep
N315	OV1983	N1122	1	7259	40792	378	66	Human strain in sheep
N315	OV1984	N1122	2	7259	40792	378	66	Human strain in sheep
N315	OV1985	N1122	3	7259	40792	378	66	Human strain in sheep
N315	OV1988	N121	1	98821	270	402	66	Human strain in sheep
N315	OV1989	N121	2	98821	270	402	66	Human strain in sheep
N315	OV1990	N121	3	98821	270	402	66	Human strain in sheep
N315	OV2045	N122	1	5798	0.2752	370	54	Human strain in sheep
N315	OV2046	N122	2	5798	0.2752	370	54	Human strain in sheep
N315	OV2047	N122	3	5798	0.2752	370	54	Human strain in sheep
N315	OV1993	N21	1	0.5797	23901	333	66	Human strain in sheep
N315	OV1994	N21	2	0.5797	23901	333	66	Human strain in sheep
N315	OV1995	N21	3	0.5797	23901	333	66	Human strain in sheep
N315	OV1998	N222	1	7265	19275	370	66	Human strain in sheep
N315	OV1999	N222	2	7265	19275	370	66	Human strain in sheep
N315	OV2000	N222	3	7265	19275	370	66	Human strain in sheep
CH3657	OV873	C200	1	2479	4965	57	57	Human strain in sheep
CH3657	OV874	C200	2	2479	4965	57	57	Human strain in sheep
CH3657	OV875	C200	3	2479	4965	57	57	Human strain in sheep

Strain	Isolate	Clonal lineage	Clone	Prev. Sheep	Sheep	d passaged	d in sheep	Description
CH3657	OV2003	C220	1	4965	38619	158	70	Human strain in sheep
CH3657	OV2004	C220	2	4965	38619	158	70	Human strain in sheep
CH3657	OV2005	C220	3	4965	38619	158	70	Human strain in sheep
CH3657	OV2008	C220b	1	38619	3126	228	41	Human strain in sheep
CH3657	OV2009	C220b	2	38619	3126	228	41	Human strain in sheep
CH3657	OV2010	C220b	3	38619	3126	228	41	Human strain in sheep
CH3657	OV858	C400	1	9115	4912	57	57	Human strain in sheep
CH3657	OV859	C400	2	9115	4912	57	57	Human strain in sheep
CH3657	OV860	C400	3	9115	4912	57	57	Human strain in sheep
CH3657	OV2040	C400b	1	4912	38616	158	70	Human strain in sheep
CH3657	OV2041	C400b	2	4912	38616	158	70	Human strain in sheep
CH3657	OV2042	C400b	3	4912	38616	158	70	Human strain in sheep
CH3657	OV2013	C420	1	38616	57993	228	77	Human strain in sheep
CH3657	OV2014	C420	2	38616	57993	228	77	Human strain in sheep
CH3657	OV2015	C420	3	38616	57993	228	77	Human strain in sheep
N315	OV2065	N221	1	68096	0.4503	251	9	Human strain in sheep
N315	OV2066	N221	2	68096	0.4503	251	9	Human strain in sheep
N315	OV2067	N221	3	68096	0.4503	251	9	Human strain in sheep
N315	OV2070	N22	1	0.838	68096	242	84	Human strain in sheep
N315	OV2071	N22	2	0.838	68096	242	84	Human strain in sheep
N315	OV2072	N22	3	0.838	68096	242	84	Human strain in sheep
N315	OV2077		1	Tube 2		120		Lab Growing
N315	OV2078		2	Tube 2		120		Lab Growing
N315	OV2079		1	Tube 3		126		Lab Growing
N315	OV2080		2	Tube 3		126		Lab Growing
CH3657	OV2081		1	Tube 1		121		Lab Growing
CH3657	OV2082		2	Tube 1		121		Lab Growing

Strain	Isolate	Clonal lineage	Clone	Prev. Sheep	Sheep	d passaged	d in sheep	Description
CH3657	OV2083		1	Tube 2		126		Lab Growing
CH3657	OV2084		2	Tube 2		126		Lab Growing
CH3657	OV2085		1	Tube 3		120		Lab Growing
CH3657	OV2086		2	Tube 3		120		Lab Growing

Days (d), previous (prev.)

Supplementary Table 5. Details on the infected versus inoculated sheep in every passage for the infection experiments with *S. aureus* strains NCTC 8325 (top) and N315 (bottom).

***S. aureus* NCTC 8325**

1	2	3	4	5	6
5/8	0/2	-	-	-	-
5/8	1/2	3/3	1/4	0/2	-
5/8	1/2	3/3	1/2	1/1	2/2
5/8	1/4	3/4	0/2	-	-
5/8	1/4	3/4	2/2	2/2	-
5/8	1/4	3/4	2/2	1/2	-
5/8	1/4	3/4	1/2	2/2	2/2
5/8	1/2	2/4	0/2	1/4	-
5/8	1/2	2/4	2/3	1/2	0/2
5/8	1/2	2/3	1/2	0/2	-
5/8	1/2	2/3	2/2	1/2	-
5/8	1/2	2/3	2/2	2/2	-

***S. aureus* N315**

1	2	3	4	5	6	7
3/10	1/2	0/3	-	-	-	-
3/10	1/2	3/3	1/3	2/7	2/2	1/2
3/10	1/2	3/3	1/3	2/7	2/2	1/2
3/10	1/2	3/3	2/3	1/3	1/3	-
3/10	1/2	3/3	2/3	2/4	1/1	1/2
3/10	1/2	2/3	1/2	2/2	1/3	-
3/10	1/2	2/3	2/3	1/2	0/2	-
3/10	1/2	2/3	2/3	2/2	1/2	-
3/10	1/2	2/3	2/3	2/2	0/2	-

Number of infected sheep (left), number of inoculated sheep (right).

Supplementary Table 6. Details of *Legionella spp*. isolates used in this study.

Identifier	Species/serogroup	Date*	Source	Country	Linked to†
02.4755	*L. anisa*	24/09/2002	Hot water	Scotland	-
03.5252	*L. anisa*	12/11/2003	Patient	Scotland	-
04.2845	*L. longbeachae* Sg1	07/06/2004	Patient	Scotland	-
06.3325	*L. anisa*	18/07/2006	Patient	Scotland	-
08.1921	*L. longbeachae* Sg1	01/04/2008	Patient	Scotland	08.2077,08.2078
08.2077	*L. longbeachae* Sg1	17/04/2008	Compost	Scotland	08.1921
08.2078	*L. longbeachae* Sg1	31/07/2008	Compost	Scotland	08.1921
09.5279	*L. longbeachae* Sg1	20/05/2009	Patient	Scotland	09.5470-4
09.5470-4	*L. longbeachae* Sg2	02/06/2009	Compost	Scotland	09.5279
09.6863	*L. longbeachae* Sg1	19/11/2009	Compost	Scotland	Patient negative
10.4571	*L. longbeachae* Sg1	19/03/2010	Patient	Scotland	Compost negative
11.3483(3)	*L. longbeachae* Sg1	13/05/2011	Compost	Scotland	Dundee
11.3484(1)	*L. longbeachae* Sg1	13/05/2011	Compost	Scotland	Dundee
12.4709	*L. longbeachae* Sg1	18/06/2012	Patient	Scotland	No compost
13.8641	*L. longbeachae* Sg1	18/06/2012	Compost	Scotland	Strathclyde
13.8642	New species	18/06/2012	Compost	Scotland	Strathclyde
13.8643	*L. longbeachae* Sg1	18/06/2012	Compost	Scotland	Strathclyde
13.8644	*L. longbeachae* Sg1	18/06/2012	Compost	Scotland	Strathclyde
13.8645	*L. longbeachae* Sg1	18/06/2012	Compost	Scotland	Strathclyde
13.8646	*L. longbeachae* Sg1	18/06/2012	Compost	Scotland	Strathclyde
13.4628	*L. longbeachae* Sg1	26/06/2013	Compost	Scotland	Dundee-Unsure
13.4630	*L. longbeachae* Sg1	26/06/2013	Compost	Scotland	Dundee-Unsure
13.5970	*L. longbeachae* Sg1	23/08/2013	Patient	Scotland	13.6038/39
13.59701	*L. longbeachae* Sg1	23/08/2013	Patient	Scotland	13.6038/39
13.59702	*L. longbeachae* Sg1	23/08/2013	Patient	Scotland	13.6038/39
13.59703	*L. longbeachae* Sg1	23/08/2013	Patient	Scotland	13.6038/39
13.59704	*L. longbeachae* Sg1	23/08/2013	Patient	Scotland	13.6038/39
13.6038	*L. longbeachae* Sg1	27/08/2013	Top soil	Scotland	13.5970
13.6121	*L. longbeachae* Sg1	30/08/2013	Patient	Scotland	13.6619/27
13.61211	*L. longbeachae* Sg1	30/08/2013	Patient	Scotland	13.6619/27
13.61212	*L. longbeachae* Sg1	30/08/2013	Patient	Scotland	13.6619/27
13.61214	*L. longbeachae* Sg1	30/08/2013	Patient	Scotland	13.6619/27
13.6310	*L. longbeachae* Sg1	06/09/2013	Patient	Scotland	13.6762
13.6314	*L. longbeachae* Sg1	06/09/2013	Patient	Scotland	13.6619
13.6472	*L. longbeachae* Sg1	13/09/2013	Patient	Scotland	No compost
13.64721	*L. longbeachae* Sg1	13/09/2013	Patient	Scotland	No compost
13.64722	*L. longbeachae* Sg1	13/09/2013	Patient	Scotland	No compost
13.64723	*L. longbeachae* Sg1	13/09/2013	Patient	Scotland	No compost
13.64724	*L. longbeachae* Sg1	13/09/2013	Patient	Scotland	No compost
13.6557	*L. longbeachae* Sg1	17/09/2013	Patient	Scotland	No compost
13.6614	*L. longbeachae* Sg1	19/09/2013	Compost	Scotland	-
13.6619	*L. longbeachae* Sg1	19/09/2013	Compost	Scotland	13.6121
13.6627	*L. longbeachae* Sg1	19/09/2013	Compost	Scotland	13.6121
13.6634	*L. longbeachae* Sg1	20/09/2013	Patient	Scotland	No compost

Identifier	Species/serogroup	Date*	Source	Country	Linked to†
13.6762	*L. longbeachae* Sg1	25/09/2013	Compost	Scotland	13.6310
13.6763	*L. longbeachae* Sg1	25/09/2013	Compost	Scotland	Dundee- Unsure
13.6764	*L. longbeachae* Sg1	25/09/2013	Compost	Scotland	Dundee- Unsure
13.6912	*L. longbeachae* Sg1	01/10/2013	Compost	Scotland	Dundee- Unsure
13.6914	*L. longbeachae* Sg1	01/10/2013	Compost	Scotland	Dundee- Unsure
8702918	*L. longbeachae* Sg1	29/05/2014	Patient	Scotland	870286x
8702860	*L. longbeachae* Sg1	07/05/2014	Soil	Scotland	8702918
8702861	*L. longbeachae* Sg1	08/05/2014	Soil	Scotland	8702918
8702862	*L. longbeachae* Sg1	09/05/2014	Soil	Scotland	8702918
8702863	*L. longbeachae* Sg1	10/05/2014	Soil	Scotland	8702918
13.8292	*L. anisa*	2010	Compost	New Zealand	13.8293
13.8293	*L. longbeachae* Sg1	2010	Patient	New Zealand	13.8292
13.8294	*L. longbeachae* Sg1	2004	Patient	New Zealand	-
13.8295	New species	2013	Patient	New Zealand	-
13.8296	*L. longbeachae* Sg2	2012	Sump drain	New Zealand	-
13.8297	*L. longbeachae* Sg2	2007	Compost	New Zealand	13.8301
13.8298	*L. longbeachae* Sg1	1996	Compost	New Zealand	-
13.8299	*L. longbeachae* Sg1	2003	Compost	New Zealand	-
13.8300	*L. longbeachae* Sg2	2011	Compost	New Zealand	-
13.8301	*L. longbeachae* Sg2	2007	Patient	New Zealand	13.8297
Sg2	*L. longbeachae* Sg2	-	-	-	-
NSW 150	*L. longbeachae* Sg1	-	Patient	Australia	-
C-4E7	*L. longbeachae* Sg2	-	Patient	Australia	-
D-4968	*L. longbeachae* Sg1	-	Patient	USA	-
ATCC39642	*L. longbeachae* Sg1	-	Patient	USA	-
98072	*L. longbeachae* Sg2	-	Patient	USA	-

*Date received in the reference laboratory. †Isolates that are linked to each other, as patient and cognate compost samples.

Appendix 2

Information on the manuscripts published and under preparation from this thesis:

Chapter 2. - Evolutionary history of a multi-host pathogen and signatures of host-adaptation.

Part of this work under review in *Nature Ecology and Evolution*.

Title: Gene exchange drives the ecological success of a multi-host bacterial pathogen

Authors: Emily J. Richardson*, **Rodrigo Bacigalupe***, Ewan M. Harrison*, Lucy A. Weinert*, Samantha Lycett, Matthew T.G. Holden, Edward J. Feil, Gavin K. Paterson, Steven Y.C. Tong, Adebayo Shittu , Willem van Wamel, David M. Aanensen, Julian Parkhill, Sharon J. Peacock, Jukka Corander, Mark Holmes and J. Ross Fitzgerald.

Chapter 3. - Host-adaptive evolution during experimental infection in the face of regular bottlenecks.

This work is in preparation for submission.

Chapter 4. - Population genomics of Legionella longbeachae and hidden complexities of infection source attribution.

This work has been published in *Emerging Infectious Diseases.* 2017;23(5):750-757. https://dx.doi.org/10.3201/eid2305.161165

Title: Population Genomics of *Legionella longbeachae* and Hidden Complexities of Infection Source Attribution

Authors: **Rodrigo Bacigalupe**, Diane Lindsay, Giles Edwards, and J. Ross Fitzgerald.

Population Genomics of *Legionella longbeachae* and Hidden Complexities of Infection Source Attribution

Rodrigo Bacigalupe, Diane Lindsay, Giles Edwards, J. Ross Fitzgerald

Legionella longbeachae is the primary cause of legionellosis in Australasia and Southeast Asia and an emerging pathogen in Europe and the United States; however, our understanding of the population diversity of *L. longbeachae* from patient and environmental sources is limited. We analyzed the genomes of 64 *L. longbeachae* isolates, of which 29 were from a cluster of legionellosis cases linked to commercial growing media in Scotland in 2013 and 35 were non–outbreak-associated isolates from Scotland and other countries. We identified extensive genetic diversity across the *L. longbeachae* species, associated with intraspecies and interspecies gene flow, and a wide geographic distribution of closely related genotypes. Of note, we observed a highly diverse pool of *L. longbeachae* genotypes within compost samples that precluded the genetic establishment of an infection source. These data represent a view of the genomic diversity of *L. longbeachae* that will inform strategies for investigating future outbreaks.

Legionellosis presents as 2 clinically distinct forms: an influenza-like illness called Pontiac fever and a severe pneumonia known as Legionnaires' disease (*1*). In Europe and the United States, most legionellosis cases are caused by *Legionella pneumophila* serogroup 1 (*1*,*2*); <5% of cases are caused by nonpneumophila *Legionella* spp. (*3*,*4*). In Australasia, New Zealand, and some countries in Asia, infections caused by *L. longbeachae* occur at comparable levels to infections caused by *L. pneumophila* (*5*–*7*). Unlike *L. pneumophila* infections, which are typically linked to artificial water systems, *L. longbeachae* infections are associated with exposure to soil, compost, and potting mixes (*8*).

The number of legionellosis cases caused by *L. longbeachae* is increasing worldwide (*7*), with a notable rise reported across Europe (*9*–*11*). Within the United Kingdom, most *L. longbeachae* infections have been identified in Scotland, where 6 cases were diagnosed during 2008–2012 (*12*) and another 6 were diagnosed in the summer of 2013 and represented a singular increased incidence or cluster with all patients requiring intensive care hospitalization (*11*). Epidemiologic investigation revealed that most patients from the 2013 cluster were avid gardeners, and *L. longbeachae* was isolated from respiratory secretions and from samples of the growing media they had used for gardening before becoming ill (*11*,*12*). However, an investigation into the provenance of the growing media did not reveal a single commercial or manufacturing source that would suggest a common origin for the *L. longbeachae* associated with the outbreak (*11*).

Molecular typing methods used to discriminate between *L. longbeachae* and other *Legionella* spp. and between the 2 *L. longbeachae* serogroups have limited efficacy, and although considerable evidence supports growing media as a source for *L. longbeachae* infections (*13*,*14*), there is still a lack of genetic evidence for an epidemiologic link. Furthermore, a population genomic study involving large numbers of *L. pneumophila* isolates has been conducted (*15*,*16*), but the same has not been done for *L. longbeachae*, so the diversity of environmental and pathogenic genotypes and the relationship between them remains unknown for *L. longbeachae*. To examine the etiology of the 2013 cluster of legionellosis cases in Scotland in the context of *L. longbeachae* species diversity, we analyzed the genomes of 70 *Legionella* spp. isolates from 4 countries over 18 years.

Materials and Methods

Bacterial Isolates

We sequenced 65 isolates that had previously been identified as *L. longbeachae*. These isolates were obtained during 1996–2014 from several patients, growing media samples (including compost and soil), and a hot water supply. Of these isolates, 55 were from Scotland (29 from the 2013 cluster of infections and 26 from other clinical and environmental samples) and 10 were from patients and

Author affiliations: The Roslin Institute, University of Edinburgh, Midlothian, Scotland, UK (R. Bacigalupe, J.R. Fitzgerald); Glasgow Royal Infirmary, Glasgow, Scotland, UK (D. Lindsay, G. Edwards)

DOI: http://dx.doi.org/10.3201/eid2305.161165

environmental compost samples in New Zealand (online Technical Appendix Table, https://wwwnc.cdc.gov/EID/article/23/5/16-1165-Techapp1.pdf).

In our analysis, we also included all publicly available genome sequences for *L. longbeachae*: *L. longbeachae* NSW150 (serogroup 1) and *L. longbeachae* C-4E7 (serogroup 2) isolated from patients in Australia; and *L. longbeachae* D-4968 (serogroup 1), *L. longbeachae* ATCC39642 (serogroup 1), and *L. longbeachae* 98072 (serogroup 2) isolated from patients in the United States (*17*–*19*). We sequenced multiple isolates (n = 2 to 5) for each of 3 patients and their linked growing media samples from the 2013 outbreak in Scotland and for 2 additional compost samples. The species of all isolates had been determined by serotyping or macrophage infectivity potentiator (mip) gene sequencing (*20*,*21*).

Bacterial Culture, Genomic DNA Isolation, and Whole-Genome Sequencing

We cultured *Legionella* spp. isolates in a microaerophilic and humid environment at 37°C on BCYE (buffered charcoal yeast extract) agar plates for 48 h. We then picked individual colonies from the plates and grew them in ACES-buffered yeast extract broth containing *Legionella* BCYE Growth Supplement (Oxoid Ltd., Basingstoke, UK) with shaking at 37°C for 24–48 h. We extracted genomic DNA from fresh cultures by using the QIAGEN DNeasy Blood and Tissue Kit (QIAGEN Benelux B.V., Venlo, the Netherlands).

We prepared sequencing libraries by using the Nextera XT kit for MiSeq or HiSeq (all from Illumina, San Diego, CA, USA) sequencing at Edinburgh Genomics, University of Edinburgh (Edinburgh, Scotland, UK). For each isolate, one 2 × 250–bp or two 2 × 200–bp paired-end sequencing runs were carried out using the MiSeq and HiSeq technologies, respectively. Raw reads were quality checked using FastQC v0.10.1 (*22*), and primers were trimmed by using Cutadapt (*23*). We used wgsim software (*24*) to simulate sequence reads for publicly available, complete whole-genome sequences.

Bioinformatic Analysis and Data Deposition

A detailed description of the bioinformatic analyses is available in the online Technical Appendix. The sequence data for the 65 genomes of *Legionella* spp. sequenced in this study were deposited in the SRA database (accession no. PRJEB14754).

Results

Limitations of Current Typing Approaches for *Legionella* spp. Identification

We sequenced 65 isolates obtained from several patients and environmental samples over 18 years in different countries and previously identified as *L. longbeachae*. To confirm the species identity of the *Legionella* isolates, we constructed a phylogenetic tree that included all *Legionella* type strains for which cultures are available, based on the 16S rRNA gene sequence (*25*). We also built phylogenetic trees based on the whole-genome content and core-genome diversity. For each approach, 64 of the 70 isolates examined co-segregated within the *L. longbeachae*–specific clade, 4 isolates clustered with *Legionella anisa*, and 2 belonged to a separate clade that was distinct from all known *Legionella* spp. (Figure 1; online Technical Appendix Figures 1, 2). The species identities were further supported by determination of the average nucleotide identity values (online Technical Appendix Figure 3), a widely used method for bacterial species delineation based on genomic relatedness (*26*). Of note, *L. anisa* is the most common nonpneumophila *Legionella* spp. in Europe (*27*–*29*). In addition, *L. longbeachae* isolates 13.8642 (from a compost sample from Scotland) and 13.8295 (from a patient in New Zealand) belong to a putative novel *Legionella* spp. Overall, the data indicate that current serotyping methods and *mip* gene sequencing are limited in their capacity to identify *L. longbeachae* to the species level.

To investigate the genetic relatedness of *L. longbeachae* strains associated with the 2013 outbreak to temporally and geographically distinct isolates, we constructed a core genome–based neighbor-joining tree of the 64 confirmed *L. longbeachae* isolates obtained from 4 countries over 18 years (online Technical Appendix Figure 4). This phylogenetic tree presents a comet-like pattern, with 2 distinct clades separated by 9,911 single-nucleotide polymorphisms, representing the major serogroups (serogroups 1 and 2) previously identified for *L. longbeachae* (*20*), each containing isolates from patient and environmental samples from different years. In contrast with findings from a previous analysis of 2 isolates of *L. longbeachae* serogroup 1 (*20*), we observed a higher diversity among the 56 isolates within serogroup 1 (online Technical Appendix Figures 1, 4); this finding is not unexpected, given the difference in the number of genomes examined. Nevertheless, compared with isolates from the same serogroup in other *Legionella* spp., such as *L. pneumophila* serogroup 1 (2% polymorphism) (*20*), *L. longbeachae* serogroup 1 exhibits a lower diversity (<0.1% polymorphism). Although serogroup 1 and 2 clades contained isolates from Scotland, Australasia, and the United States, 96% of the isolates from Scotland (including all of the 2013 outbreak isolates) belonged to serogroup 1, suggesting that serogroup 1 may be more clinically relevant in Scotland than in some other countries where *L. longbeachae* is a more established cause of legionellosis. However, analysis of more isolates from different countries would be required to investigate this observation further.

Figure 1. 16S rRNA gene–based phylogenetic tree of the sequenced genomes and all the cultured and type *Legionella* spp. strains available in the ribosomal database project (http://rdp.cme.msu.edu/), as accessed in May 2015. Scale bar indicates the mean number of nucleotide substitutions per site.

Effect of Recombination on *L. longbeachae* Serogroup 1 Population Structure

It is established that recombination has played a key role in shaping the evolutionary history of *L. pneumophila,* but its effect on *L. longbeachae* population structure is unknown (*22*,*30*). This knowledge is critical because for highly recombinant bacteria, recombination networks may represent evolutionary relationships more explicitly than traditional phylogenetic trees. Therefore, we constructed a recombination network of all serogroup 1 isolates by using the neighbor-net algorithm of SplitsTree4 (*31*). The resultant network displayed a reticulate topology with an extensive reticulated background from which clusters of isolates emerge, supporting an evolutionary history involving recombination ($p< 0.01$ by ϕ test) (*32*), followed by clonal expansion and subsequent additional recombination events among some lineages (online Technical Appendix Figure 5). Using BratNextGen (*33*), we identified a total of 94 predicted recombination events affecting more than half of the core genome (1.74 Mb of 3.36 Mb) and representing recent and ancient recombination events of different sizes (range 1,350 bp–350 Kbp) distributed across the phylogeny (online Technical Appendix Figure 6). Given the reported limitation in sensitivity of BratNextGen for the identification of all recombination events (*34*), we also used ClonalFrameML (*35*), an algorithm that uses maximum likelihood inference to simultaneously detect recombination in bacterial genomes and account for it in phylogenetic reconstruction. The estimated average length of the recombined fragments was 8,047 bp, and the ratio of recombination to mutation was 1.42, indicating a greater role for recombination over mutation in the diversification

of *L. longbeachae*. This estimate is in accordance with early estimates for *L. pneumophila* based on multiple gene sequence data (*36*), but it is low compared with recent estimates based on whole-genome sequence data [recombination to mutation ratios of 16.8 (*30*) or 47.93 (*37*)]. Differences in the clonal diversity of *Legionella* spp. sequence datasets used to determine recombination rates could affect the estimates. Reconstruction of the phylogeny after removal of all predicted recombinant sequences resulted in a tree with largely similar clusters of isolates but with reduced branch lengths and variation in the position of nodes deep in the phylogeny (Figure 2).

Accessory Genome Analysis Indicates Extensive Interspecies and Intraspecies Gene Flow

The extent to which horizontal gene transfer occurs among *L. longbeachae* isolates and between *L. longbeachae* and other *Legionella* spp. is unknown. In our study, the pangenome of *L. longbeachae* represented by the 56 serogroup 1 isolates was 6,890 genes, including a core genome of 2,574 genes; the average gene content was 3,558 genes per strain. The accessory genome, which included only strain-dependent genes varied from 809 to 1,155 genes, depending on the strain. A parsimony clustering analysis based on the presence or absence of all genes classified the isolates in a manner distinct from that in a core genome–based maximum-likelihood tree, suggesting extensive horizontal gene transfer among *L. longbeachae* isolates (online Technical Appendix Figures 1, 2). BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis of all assembled contigs was used to filter for plasmid-related homologous sequences, revealed 2 major plasmids: pLLO, described previously in *L. longbeachae* NSW150 (*20*), and pLELO, originally identified

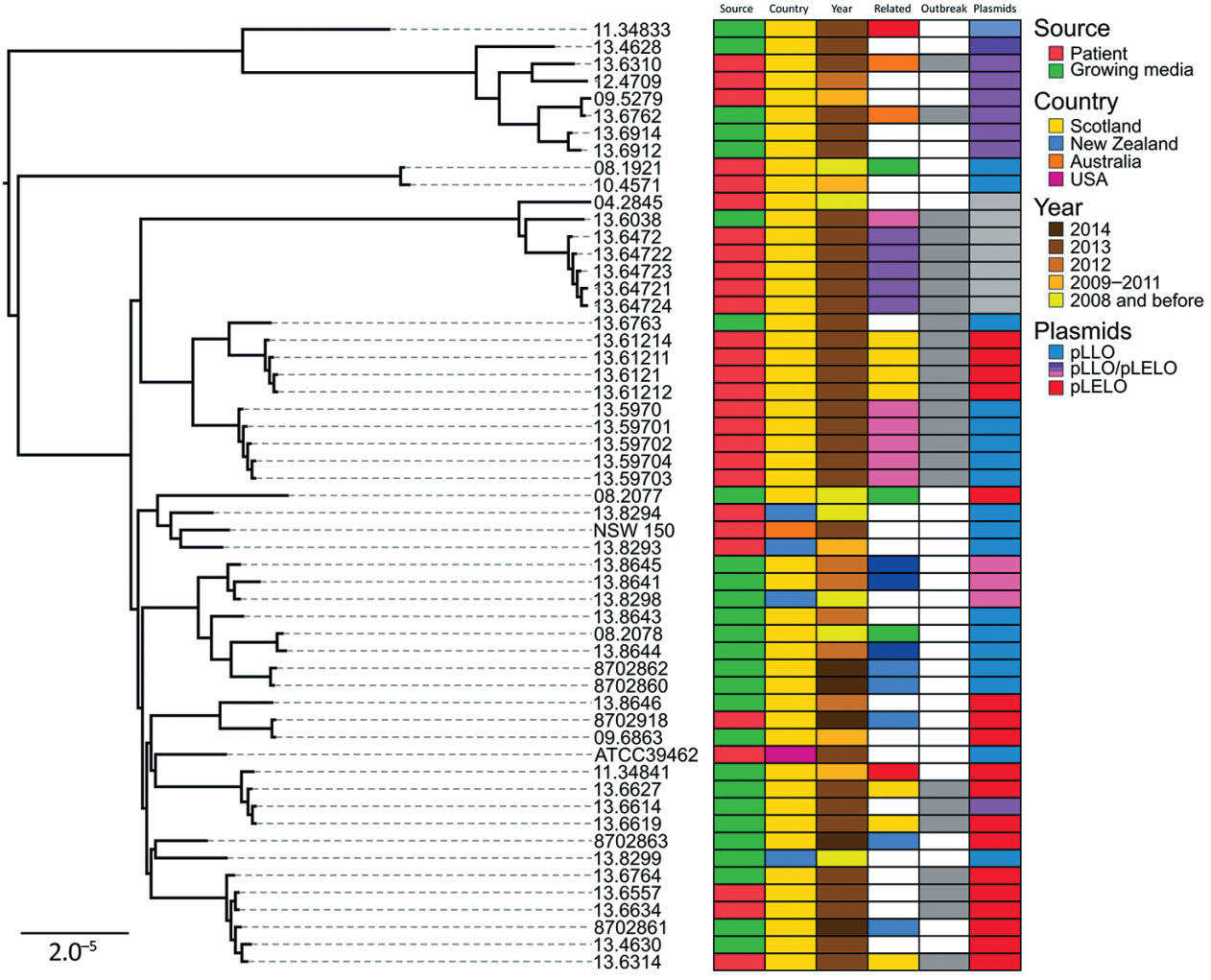


Figure 2. Core genome–based maximum-likelihood phylogeny of *Legionella longbeachae* serogroup 1 isolates corrected for recombination; source, country, year of isolation, relatedness and plasmid carriage are indicated. Related isolates are shown in the same color; those from the 2013 outbreak are indicated by gray. Isolates from the same patient are clustered together but do not co-segregate with cognate compost samples. Scale bar indicates the mean number of nucleotide substitutions per site.

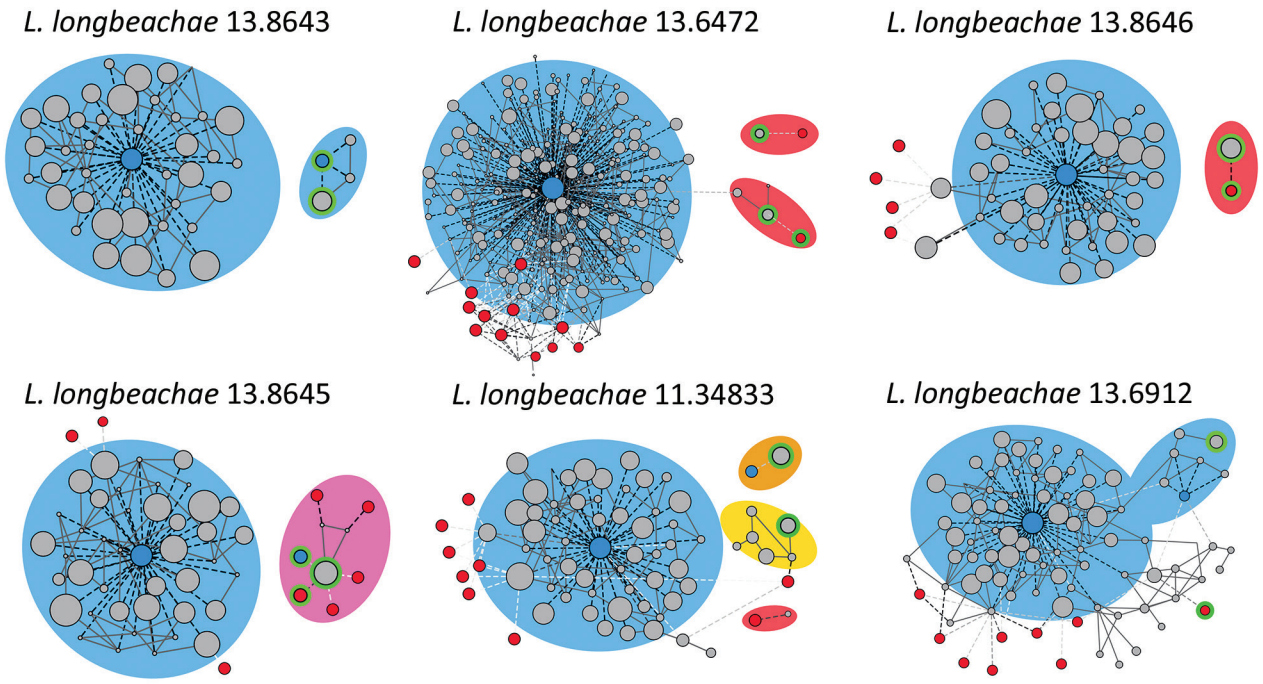


Figure 3. *Legionella longbeachae* plasmid analysis: contigs networks reconstructions for 6 representative *L. longbeachae* types of plasmid content. The networks of the contigs representing the main chromosome and plasmids comprising the genome obtained by using PLACNET (*38*), a program enabling reconstruction of plasmids from whole-genome sequence datasets. The sizes of the contig nodes (in gray) are proportional to their lengths; continuous lines correspond to scaffold links. Dashed lines represent BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) hits to the *L. longbeachae* (blue) or *L. pneumophila* (red) strains; intensity of the line is proportional to the hit (white indicates low, black indicates high). Green lines correspond to plasmid contigs. Background colors indicate species relatedness for the main chromosome and plasmids (blue for *L. longbeachae*, red for *L. pneumophila*, pink for a combination of both, and yellow for previously unidentified genomic content).

in *L. pneumophila subsp. pneumophila* (*22*). Of the 55 serogroup 1 isolates, 36 contained sequences for the pLLO and pLELO plasmids. Of note, the distribution of these plasmids among the *L. longbeachae* isolates correlated with the gene content–based clustering, whereas the distribution of plasmids in the core genome–based tree was independent of the phylogeny (Figure 2). In addition, 11 isolates appeared to contain plasmids with sequences homologous to those for pLLO and pLELO, which is indicative of recombinant forms of the plasmid. Further examination of plasmid diversity using a modified version of PLACNET (*38*), a program enabling reconstruction of plasmids from whole-genome sequence datasets, confirmed that some plasmids consisted of a mosaic of recombinant fragments homologous to pLELO, pLLO, or other unknown plasmids (Figure 3). Overall, these data indicate the high prevalence of specific plasmids among *L. longbeachae* isolates and reveal extensive recombination and horizontal gene transfer among different *Legionella* spp (*39*). The high prevalence of plasmids in *L. longbeachae* is notable, considering these elements may be less common in *L. pneumophila* (*30*).

To examine the possibility that clinical and environmental isolates of *L. longbeachae* contained genomic differences reflecting their distinct origins, we compared their accessory genome content. For isolates obtained from a single patient sample, the accessory genome was highly conserved compared with those for environmental isolates from a single compost sample or closely related environmental isolates from distinct compost samples (Figure 4, panel A). In addition, considering the average gene content of all sequenced isolates (28 clinical and 27 environmental), the gene content for *L. longbeachae* from growing media samples (3,586 genes) was significantly higher than that for isolates from patients (3,533 genes; 2-sample *t*-test, $t = 2.5213$; d.f. = 53; $p = 0.01474$) (Figure 4, panel B). The data imply that gene loss occurs during human infection or that *L. longbeachae* strains with reduced gene content have enhanced human infectivity. However, we did not identify a specific enriched gene or functional category in clinical or environmental samples (data not shown).

Source Attribution Confounded by Complex Serogroup 1 Populations within Environmental Samples

Having accounted for the influence of recombination on the phylogeny of *L. longbeachae*, we investigated the diversity of isolates associated with 5 patients and their linked compost samples obtained during 2008–2014, including

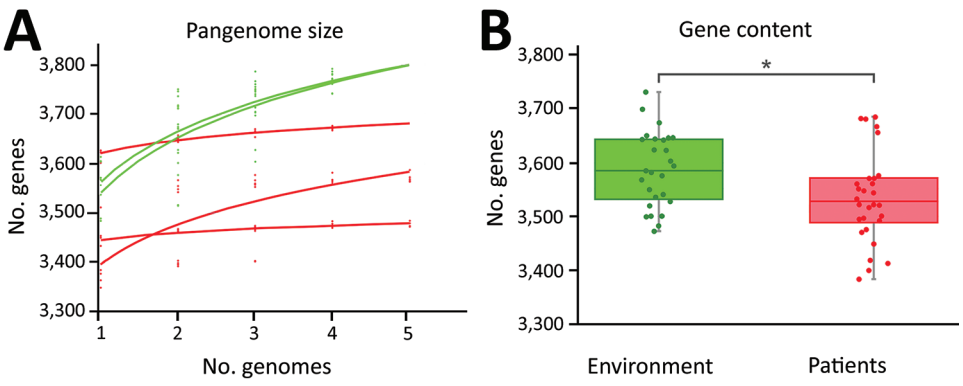


Figure 4. Variation in gene content between environmental and patient *Legionella longbeachae* samples. A) Increase in pangenome size with every addition of a *L. longbeachae* genome. Environmental isolates pangenomes (green) are larger and continue increasing after the addition of 5 genomes, consistent with an open pangenome, but the within-patient pangenome plateaus quickly, consistent with a more closed pangenome. B) Average gene content of environmental isolates is significantly higher than that of clinical isolates ($p = 0.01474$).

3 patients from the 2013 outbreak in Scotland. Of note, isolates from the 2013 outbreak were distributed across several subclades of the tree, indicating that the infections were caused by different strains (Figure 2). However, all isolates from a single patient clustered together, consistent with a monoclonal etiology of each infection. Of note, for all 5 patients, clinical isolates were not closely allied to the environmental isolates obtained from linked compost samples, and therefore a genetic link between patient and compost samples could not be established. Most subclades included isolates of diverse geographic origin, consistent with a wide distribution for *L. longbeachae* strains; however, 3 *L. longbeachae* isolates originating from Australasia (strains 13.8294, 13.8293, and NSW150) belonged to their own region-specific cluster (Figure 2).

We hypothesized that the lack of genetic relatedness between *L. longbeachae* isolates from patients and linked compost samples could be explained by a highly diverse population of *L. longbeachae* in growing media samples compounded by a sampling strategy consisting of a single sequenced isolate. All 5 compost samples for which we had >1 isolate contained isolates distributed across multiple clades in the phylogenetic tree. In particular, 5 isolates from the same growing media sample linked to a patient infected in Edinburgh in 2014 were distributed across 4 distinct clades, demonstrating that within a single environmental sample, considerable species diversity may be represented (Figure 2). Taken together, these data suggest that for future outbreak investigations, extensive sampling of environmental samples may be required to identify genotypes responsible for episodes of legionellosis infection, if indeed they are present.

Discussion

Our findings reveal the population genomic structure for *L. longbeachae*, an emerging pathogen in Europe and the United States, and includes a genome-scale investigation into an outbreak of *L. longbeachae* legionellosis. We provide evidence for extensive recombination and lateral gene transfer among *L. longbeachae*, including the presence of widely distributed mosaic plasmids that have likely recombined with plasmids from other *Legionella* spp., suggesting an ecologic overlap or shared habitat. Our analysis highlights the need to account for recombination events when determining the genetic relatedness of *L. longbeachae* isolates.

Our application of whole-genome sequencing for diagnostic purposes revealed the misidentification, using current serotyping methods, of several *L. anisa* isolates as *L. longbeachae* and led to the identification of a putative novel *Legionella* sp. linked to legionellosis. These findings highlight the limitations of current typing methods for differentiation of *Legionella* spp. and accurate identification of legionellosis etiology.

We used whole-genome sequencing to attempt to establish a genetic link between legionellosis infections and associated compost samples. Our inability to establish a link probably reflects the traditional strategy of single isolate sampling, which when applied to a highly diverse pool of *L. longbeachae* genotypes fails to detect the infecting genotype. We suggest that the approach to investigating the source of future legionellosis cases linked to growing media will require a radical revision of sampling protocols to maximize the chances of isolating the infecting strain, if present. Taken together, our findings provide a view of the population structure of *L. longbeachae* and highlight the complexities of tracing the origin of legionellosis associated with growing media. Overall, our findings demonstrate the resolution afforded by whole-genome sequencing for understanding the biology underpinning legionellosis and provide information that should be considered for future epidemiologic investigations.

Acknowledgment

We are grateful to Carmen Buchrieser for providing the original sequence reads for *L. longbeachae* strains ATCC39642, 98072, and C-4E7. We thank David Harte for supplying the New Zealand strains.

Funding was provided by the Chief Scientist's Office Scotland (grant ETM/421 to J.R.F.) and by the Biotechnology and Biological Sciences Research Council (ISP3 grant BB/J004227/1 to J.R.F.).

Mr. Bacigalupe is a PhD candidate at the Roslin Institute, University of Edinburgh. His primary research focuses on the evolution, adaptation, and outbreak dynamics of bacterial pathogens.

References

1. Fields BS, Benson RF, Besser RE. *Legionella* and Legionnaires' disease: 25 years of investigation. Clin Microbiol Rev. 2002; 15:506–26. http://dx.doi.org/10.1128/CMR.15.3.506-526.2002
2. European Centre for Disease Prevention and Control. Surveillance report. Legionnaires' disease in Europe, 2010. 2012 [cited 2016 Jul 9]. http://ecdc.europa.eu/en/publications/publications/sur-legionnaires-disease-surveillance-2010.pdf
3. Joseph CA, Ricketts KD; European Working Group for Legionella Infections. Legionnaires disease in Europe 2007-2008. Euro Surveill. 2010;15:19493.
4. Marston BJ, Lipman HB, Breiman RF. Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. Arch Intern Med. 1994;154:2417–22. http://dx.doi.org/10.1001/archinte.1994.00420210049006
5. Li JS, O'Brien ED, Guest C. A review of national legionellosis surveillance in Australia, 1991 to 2000. Commun Dis Intell Q Rep. 2002;26:461–8.
6. Cramp GJ, Harte D, Douglas NM, Graham F, Schousboe M, Sykes K. An outbreak of Pontiac fever due to *Legionella longbeachae* serogroup 2 found in potting mix in a horticultural nursery in New Zealand. Epidemiol Infect. 2010;138:15–20. http://dx.doi.org/10.1017/S0950268809990835
7. Whiley H, Bentham R. *Legionella longbeachae* and legionellosis. Emerg Infect Dis. 2011;17:579–83. http://dx.doi.org/10.3201/eid1704.100446
8. Yu VL, Plouffe JF, Pastoris MC, Stout JE, Schousboe M, Widmer A, et al. Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: an international collaborative survey. J Infect Dis. 2002;186:127–8. http://dx.doi.org/10.1086/341087
9. García C, Ugalde E, Campo AB, Miñambres E, Kovács N. Fatal case of community-acquired pneumonia caused by *Legionella longbeachae* in a patient with systemic lupus erythematosus. Eur J Clin Microbiol Infect Dis. 2004;23:116–8. http://dx.doi.org/10.1007/s10096-003-1071-7
10. den Boer JW, Yzerman EPF, Jansen R, Bruin JP, Verhoef LPB, Neve G, et al. Legionnaires' disease and gardening. Clin Microbiol Infect. 2007;13:88–91. http://dx.doi.org/10.1111/j.1469-0691.2006.01562.x
11. Potts A, Donaghy M, Marley M, Othieno R, Stevenson J, Hyland J, et al. Cluster of Legionnaires disease cases caused by *Legionella longbeachae* serogroup 1, Scotland, August to September 2013. Euro Surveill. 2013;18:20656. http://dx.doi.org/10.2807/1560-7917.ES2013.18.50.20656
12. Lindsay DSJ, Brown AW, Brown DJ, Pravinkumar SJ, Anderson E, Edwards GF. *Legionella longbeachae* serogroup 1 infections linked to potting compost. J Med Microbiol. 2012;61:218–22. http://dx.doi.org/10.1099/jmm.0.035857-0
13. Steele TW, Lanser J, Sangster N. Isolation of *Legionella longbeachae* serogroup 1 from potting mixes. Appl Environ Microbiol. 1990;56:49–53.
14. Koide M, Arakaki N, Saito A. Distribution of *Legionella longbeachae* and other legionellae in Japanese potting soils. J Infect Chemother. 2001;7:224–7. http://dx.doi.org/10.1007/s101560170017
15. Reuter S, Harrison TG, Köser CU, Ellington MJ, Smith GP, Parkhill J, et al. A pilot study of rapid whole-genome sequencing for the investigation of a *Legionella* outbreak. BMJ Open. 2013;3:e002175. http://dx.doi.org/10.1136/bmjopen-2012-002175
16. Rao C, Benhabib H, Ensminger AW. Phylogenetic reconstruction of the *Legionella pneumophila* Philadelphia-1 laboratory strains through comparative genomics. PLoS One. 2013;8:e64129. http://dx.doi.org/10.1371/journal.pone.0064129
17. Cazalet C, Gomez-Valero L, Rusniok C, Lomma M, Dervins-Ravault D, Newton HJ, et al. Analysis of the Legionella longbeachae genome and transcriptome uncovers unique strategies to cause Legionnaires' disease. PLoS Genet. 2010;6:e1000851. http://dx.doi.org/10.1371/journal.pgen.1000851
18. Kozak NA, Buss M, Lucas CE, Frace M, Govil D, Travis T, et al. Virulence factors encoded by Legionella longbeachae identified on the basis of the genome sequence analysis of clinical isolate D-4968. J Bacteriol. 2010;192:1030–44. http://dx.doi.org/10.1128/JB.01272-09
19. Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, et al. Extensive recombination events and horizontal gene transfer shaped the Legionella pneumophila genomes. BMC Genomics. 2011;12:536. http://dx.doi.org/10.1186/1471-2164-12-536
20. Ratcliff RM, Lanser JA, Manning PA, Heuzenroeder MW. Sequence-based classification scheme for the genus Legionella targeting the mip gene. J Clin Microbiol. 1998;36:1560–7.
21. Fallon RJ, Abraham WH. Experience with heat-killed antigens of L. longbeachae serogroups 1 and 2, and L. jordanis in the indirect fluorescence antibody test. Zentralbl Bakteriol Mikrobiol Hyg A. 1983;255:8–14.
22. Babraham Bioinformatics. FastQC. 2010 [cited 2016 Jul 9]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
23. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–12.
24. GitHub, Inc. wgsim Read Simulator [cited 2016 Jul 9]. https://github.com/lh3/wgsim
25. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42(D1):D633–42. http://dx.doi.org/10.1093/nar/gkt1244
26. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;64:346–51. http://dx.doi.org/10.1099/ijs.0.059774-0
27. Health Protection Scotland. Surveillance report: legionellosis in Scotland 2013–2014. 2015 Sep 1 [cited 2015 Aug 15]. http://www.hps.scot.nhs.uk/resp/wrdetail.aspx?id=65135&wrtype=6
28. van der Mee-Marquet N, Domelier AS, Arnault L, Bloc D, Laudat P, Hartemann P, et al. *Legionella anisa*, a possible indicator of water contamination by *Legionella pneumophila.* J Clin Microbiol. 2006;44:56–9. http://dx.doi.org/10.1128/JCM.44.1.56-59.2006
29. Svarrer CW, Uldum SA. The occurrence of *Legionella* species other than *Legionella pneumophila* in clinical and environmental samples in Denmark identified by *mip* gene sequencing and matrix-assisted laser desorption ionization time-of-flight mass

spectrometry. Clin Microbiol Infect. 2012;18:1004–9. http://dx.doi.org/10.1111/j.1469-0691.2011.03698.x
30. Underwood AP, Jones G, Mentasti M, Fry NK, Harrison TG. Comparison of the *Legionella pneumophila* population structure as determined by sequence-based typing and whole genome sequencing. BMC Microbiol. 2013;13:302. http://dx.doi.org/10.1186/1471-2180-13-302
31. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67. http://dx.doi.org/10.1093/molbev/msj030
32. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172:2665–81. http://dx.doi.org/10.1534/genetics.105.048975
33. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012;40:e6. http://dx.doi.org/10.1093/nar/gkr928
34. de Been M, van Schaik W, Cheng L, Corander J, Willems RJ. Recent recombination events in the core genome are associated with adaptive evolution in *Enterococcus faecium.* Genome Biol Evol. 2013;5:1524–35. http://dx.doi.org/10.1093/gbe/evt111
35. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLOS Comput Biol. 2015;11:e1004041. http://dx.doi.org/10.1371/journal.pcbi.1004041
36. Coscollá M, Comas I, González-Candelas F. Quantifying nonvertical inheritance in the evolution of *Legionella pneumophila.* Mol Biol Evol. 2011;28:985–1001. http://dx.doi.org/10.1093/molbev/msq278
37. Sánchez-Busó L, Comas I, Jorques G, González-Candelas F. Recombination drives genome evolution in outbreak-related *Legionella pneumophila* isolates. Nat Genet. 2014;46:1205–11. http://dx.doi.org/10.1038/ng.3114
38. Lanza VF, de Toro M, Garcillán-Barcia MP, Mora A, Blanco J, Coque TM, et al. Plasmid flux in *Escherichia coli* ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. PLoS Genet. 2014;10:e1004766. http://dx.doi.org/10.1371/journal.pgen.1004766
39. Cazalet C, Rusniok C, Brüggemann H, Zidane N, Magnier A, Ma L, et al. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nat Genet. 2004;36:1165–73. http://dx.doi.org/10.1038/ng1447

Address for correspondence: J. Ross Fitzgerald, The Roslin Institute, University of Edinburgh, Midlothian, EH259RG Scotland, UK; email: Ross.Fitzgerald@ed.ac.uk