Evolutionary insights into the host–specific adaptation and pathogenesis of group B Streptococcus

Alexandre Miguel Santos Almeida

Université Pierre et Marie Curie (Paris 6)
Ecole Doctorale: Complexité du Vivant

Thèse présentée par

Alexandre Miguel Santos Almeida

pour obtenir le grade de Docteur

Evolutionary insights into the host-specific adaptation and pathogenesis of group B *Streptococcus*

Soutenance prévue le 31 mars 2017.

Jury composé de:

Prof. Philippe Lopez	Université Paris VI	Président
Dr Immaculada Margarit	GSK Vaccines	Rapporteur
Prof. Ivan Matic	Université Paris V	Rapporteur
Dr Christine Citti	CNRS, INRA	Examinateur
Dr Olivier Tenaillon	CNRS, IAME, Université Paris VII	Examinateur
Dr Philippe Glaser	Institut Pasteur	Directeur de these

Université Pierre et Marie Curie (Paris 6)
Doctoral School: Complexité du Vivant

Thesis presented by

Alexandre Miguel Santos Almeida

for the degree of Doctoral of Philosophy

Evolutionary insights into the host-specific adaptation and pathogenesis of group B *Streptococcus*

Defence planned for March 31, 2017.

Jury composition:

Prof. Philippe Lopez	Université Paris VI	President
Dr Immaculada Margarit	GSK Vaccines	Reviewer
Prof. Ivan Matic	Université Paris V	Reviewer
Dr Christine Citti	CNRS, INRA	Examiner
Dr Olivier Tenaillon	CNRS, IAME, Université Paris VII	Examiner
Dr Philippe Glaser	Institut Pasteur	Thesis supervisor

"*There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.*"

— Charles Darwin

Acknowledgements

The collaborations and relationships developed with so many people during my PhD were crucial for accomplishing this work. First of all I would like to thank my supervisor Philippe Glaser for having given me the opportunity to work in his team and for being an exemplary PhD mentor. With his relentless curiosity and creative mind, Philippe constantly challenged me with new ideas to drive the project forward. Most importantly, he was a good friend throughout these years, caring and understanding when needed, always ready to give helpful advice.

I would also like to give a special thanks to Isabelle Rosinski-Chupin, the other member of our team who followed me closely through these years. Her hard working attitude and critical thinking inspired me to become a better scientist by having a more open mind and being more cautious and critical about my results.

I have to give a shout-out to all the other former and present members of the unit Ecologie et Evolution de la Résistance aux Antibiotiques (EERA), especially Alexandre Lecomte, Anita Annamalé, Dimitri Desvillechabrol, Elisabeth Sauvage, Nicolas Cabanel, Olivia Dowding, Pierre-Emmanuel Douarre, Rafael Patiño-Navarrete and Romain Guérillot. In some way or another they have all helped me during my stay and created a welcoming and friendly environment to work in.

A special thank you to all our collaborators, especially the teams of Claire Poyart, Fernando Tavares and Ilda Sanches, who were instrumental in providing the right resources and expertise to carry out this work. I would like to thank in particular Claire Poyart's team at the Institut Cochin and the Centre National de Référence des Streptocoques (CNR-Strep) for insightful scientific discussions and for providing access to a valuable set of clinical samples that were crucial to my PhD project.

Thank you also to Adrien Villain for getting me started in the world of bioinformatics and to Christiane Bouchier and Laurence Ma for their invaluable service at the Genopole sequencing platform. I also thank the teams of Carmen Buchrieser and Patrick Trieu-Cuot, for the useful feedback, advice and help provided.

I thank the Pasteur-Paris International PhD Program and the ANR LaBex IBEID for the funding and opportunity to develop my PhD project at the Institut Pasteur.

On a more personal level, I have to thank my partner in crime and the most important person in my life, Ana Silva. I thank her for her love and care, for being with me during the highs and lows and for having taken the courage to come with me on this journey.

Thank you to all my colleagues from the Pasteur PhD program and to the Portuguese community at the Institut Pasteur. A special thanks to Jovana Mihajlovic, Madhura Mukhopadhyay and Marisa Oliveira for all the fun and relaxing times we had to balance the hardships of everyday life.

And last, but not least, to all my friends and family back in Portugal who followed me on this endeavour from afar. I have to give a very special thanks to my parents and sister who supported me most closely on this personal venture.

Table of contents

List of abbreviations

AIC	Akaike Information Criterion
ABC	ATP-binding cassette
BHI	Brain heart infusion
BIC	Bayesian Information Criterion
CAMP	Christie Atkins and Munch-Petersen
CAS	CRISPR-associated
CC	Clonal complex
CDS	Coding sequence
CNV	Copy number variation
CovRS	Control of virulence regulator and sensor
CPS	Capsule polysaccharide
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	CRISPR RNA
CSF	Cerebrospinal fluid
dsDNA	Double-stranded DNA
EOD	Early-onset disease
FDR	False discovery rate
fMet	N-formylmethionine
GBS	Group B *Streptococcus*
GBC	Group B-specific antigen
GWAS	Genome-wide association study
HGT	Horizontal gene transfer
Hfr	High frequency of recombination
ICE	Integrative and conjugative element
Indel	Insertion or deletion
LOD	Late-onset disease
LTA	Lipoteichoic acid
MCMC	Markov chain Monte Carlo
MGE	Mobile genetic element
ML	Maximum likelihood

MLE	Maximum likelihood estimate
MLS	Macrolide-lincosamide-streptogramin B
MLST	Multi-locus Sequence Typing
MP	Maximum parsimony
MRCA	Most recent common ancestor
ncRNA	Non-coding RNA
NJ	Neighbour-joining
PCR	Polymerase chain reaction
PFGE	Pulse-field gel electrophoresis
PI	Pilus genomic island
PTS	Phosphotransferase system
RBS	Ribosomal binding site
SRR	Serine-rich repeat
ST	Sequence type
STM	Signature-tagged transposon mutagenesis
SNP	Single nucleotide polymorphism
TH	Todd Hewitt
tRNA	Transfer RNA
UTR	Untranslated region

List of figures

List of tables

Abstract

Streptococcus agalactiae (group B *Streptococcus*, GBS) is a commensal of the intestinal and genitourinary tracts in the human population, while also a leading cause of neonatal infections. Likewise, GBS remains a serious concern in many countries as frequently responsible for bovine mastitis. Therefore, the purpose of my PhD project was to use state-of-the-art whole-genome approaches to decipher the host-specific adaptation and pathogenesis of GBS in both humans and bovines. Under this framework, I examined the evolution of human-derived lineages through complementary evolutionary perspectives. Furthermore, I analysed the persistence of a bovine GBS clone in Portugal to bring about new insights into the evolutionary processes behind its adaptation to the bovine ecosystem.

By comparing the genomic profile of strains from infected newborns and their mothers we showed that the transmission of GBS from mother to child is accompanied in particular instances by the acquisition of specific pathoadaptive mutations. From this comparative analysis, we also argue that transmission between mother and child is not strictly one-directional. Moreover, from the study of the evolutionary forces acting on the human-specific and hypervirulent clonal complex (CC) 17, we reveal that various systems can evolve to improve the ability of GBS to survive in the human host. Functions related to metabolism, cell adhesion, regulation and immune evasion were among the most preferentially affected in GBS strains from human origin. Conversely, colonization of Portuguese dairy farms by one single CC61 clone for over 20 years highlighted that the specific regulation of iron/manganese uptake is a recurrent adaptive strategy in the bovine environment. Differential patterns of pseudogenization among human and bovine populations provided further insights into the evolutionary history of GBS, leading us to propose a model for the emergence of host-specific lineages and the risk of cross-species transmission.

We have successfully leveraged the use of whole-genome sequencing in the study of GBS evolution and of its host-specific lifestyle. The results we present improve our

understanding of adaptation among host generalist species, bringing useful insights that may specifically aid in the control and treatment of GBS infections worldwide.

Résumé

Streptococcus agalactiae (streptocoque du groupe B, SGB) est un commensal fréquent des voies intestinale et génito-urinaire dans la population humaine mais constitue une des causes principales d'infections néonatales. Dans le même temps, SGB est connu comme pathogène vétérinaire, responsable de mastites bovines à l'origine de pertes économiques importantes dans plusieurs pays comme le Portugal. L'objectif de ma thèse était d'analyser au niveau génomique les bases de l'adaptation spécifique de SGB à ses hôtes humains et bovins et de l'établissement des lignées plus pathogènes. Dans ce cadre, j'ai examiné l'évolution des lignages humains à deux échelles de temps différentes. J'ai aussi analysé l'adaptation et la persistance d'un clone de SGB bovin en Portugal pour caractériser les processus évolutifs contribuant à son adaptation à l'écosystème bovin.

Dans un premier temps, la comparaison des profils génomiques des souches isolées de nouveau-nés infectés et de leurs mères nous a permis de montrer que la transmission de SGB de mère à enfant est accompagnée dans certains cas par l'acquisition de mutations pathoadaptives spécifiques. Nous avons également proposé que la transmission entre la mère et l'enfant n'est pas strictement unidirectionnelle. Par ailleurs, l'analyse des séquences génomiques de plus de 600 souches de SGB appartenant au complexe clonal (CC) 17, hypervirulent et spécifique à l'hôté humain, nous a permis de caractériser les forces évolutives agissant sur ce complexe. Nous avons montré que divers systèmes ont évolué pour améliorer la capacité de SGB CC17 à survivre dans l'hôte humain. Les fonctions liées au métabolisme, à l'adhésion cellulaire, à la régulation de l'expression et à l'évasion immunitaire sont les plus souvent trouvées comme modifiées lors de cette adaptation. Inversement, l'étude de la colonisation des fermes laitières portugaises par un seul clone CC61 depuis plus de 20 ans a mis en évidence que la régulation spécifique de l'import du fer/manganèse est une stratégie d'adaptation récurrente dans l'environnement bovin. Les profils différentiels de pseudogénéisation chez les populations de souches d'origine humaine et bovine ont permis de mieux comprendre l'histoire évolutive de SGB, nous permettant de proposer un modèle pour l'apparition des lignages spécifiques et le risque de transmission inter-espèces.

En conclusion, nous avons profité avec succès du séquençage massif des génomes bactériens pour l'étude de l'évolution de SGB et de son mode de vie spécifique à ses hôtes. Les résultats que nous présentons améliorent notre compréhension de l'adaptation chez les espèces hôte-généralistes, en apportant des idées utiles qui pourront spécifiquement aider à améliorer le contrôle et le traitement des infections de SGB mondialement.

Introduction

With the advent of whole-genome sequencing, it is now possible to accurately infer how bacterial populations are able to disseminate and adapt at multiple evolutionary scales. Current genomic approaches are striving to discover the basis of many bacterial phenotypes by predicting the outcome of adaptive changes occurring during evolution of individual populations. Moreover, genomic data can be used to make assumptions about the environmental conditions and evolutionary pressures encountered in various bacterial reservoirs, as well as on the microbial ecology and diversity of different ecosystems. The relentless adaptation of human pathogens is epitomised by the emergence and proliferation of antibiotic resistant bacteria in modern civilization. Therefore, deciphering how these genomic changes occur has strong implications at a fundamental level, in better understanding the biology of bacterial pathogens, but also for epidemiology and clinical purposes in determining the most effective control and therapeutic strategies.

The opportunistic pathogen *Streptococcus agalactiae* (group B *Streptococcus*, GBS) causes disease in various hosts, most notably in humans and cattle. But, its evolutionary history and host specificities are not entirely understood. To address this, current genomic methods provide an in-depth approach of studying the host-microbe interaction of GBS in order to: deduce the most significant selective constraints; assess how GBS interacts with the surrounding microbial communities and identify the genes that are most frequently exchanged; determine the risk of inter-host transmission of GBS and help control its ongoing dissemination.

1. The host generalist pathogen *Streptococcus agalactiae*

Streptococcus is a genus that comprises more than 50 recognized species of gram-positive and spherical bacteria that are able to colonize a variety of animal hosts (Koehler 2007). While studying infections in cattle, Nocard and Mollereau identified a particular mastitis-causing *Streptococcus*, which they named *Streptococcus nocardi* (Nocard and Mollereau 1887). Later, this species was renamed *Streptococcus agalactiae* (from the latin meaning of "no milk"), alluding to its negative impact on bovine milk production.

Historically, streptococci were divided into two groups: β-haemolytic and nonhaemolytic, based on their ability to cause the lysis of red blood cells. Later, in 1933, Rebecca Lancefield further divided β-haemolytic streptococci into several categories, according to a carbohydrate found in the cell surface of the bacteria (Lancefield 1933). Hence, *S. agalactiae* is equally referred to as group B *Streptococcus* (GBS) as it is the only streptococcal species expressing the B antigen.

GBS is a facultative anaerobe that can be easily cultured in Todd Hewitt (TH) or brain heart infusion (BHI) medium. Like other streptococci, it forms spherical chains that can be observed under a microscope, with the diameter of each cell ranging between 0.6-1.2 um (Figure 1).

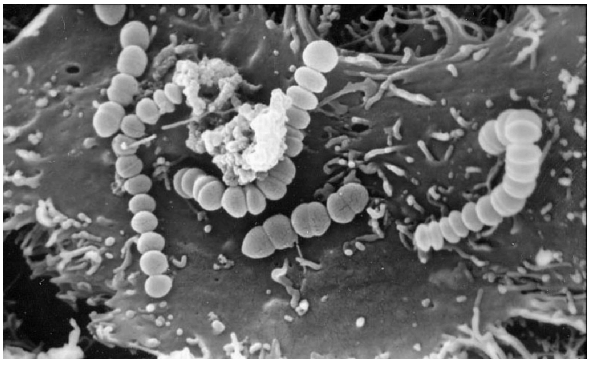


Figure 1. Scanning electron microscopy image of GBS (Brochet 2008).

In nature, this bacterium has a broad host range, but is most notably known for causing severe infections within humans, bovines and aquatic animals (Edmond et al 2012,

Wilkinso et al 1973, Wyder et al 2011). GBS inhabits the intestinal and genitourinary tract of 10-30% of humans, but emerged in the 1960s as a leading cause of neonatal infections. In the adult population, colonization is mostly asymptomatic, although opportunistic infections can also occur among the elderly or in immunocompromised individuals (van der Mee-Marquet et al 2008). In cattle, GBS adheres to the mammary epithelium causing bovine mastitis (Wyder et al 2011), while it was also found to be responsible for outbreaks of invasive disease in several fish farms (Evans et al 2002, Wilkinso et al 1973). Each one of these reservoirs represents distinct ecological niches, reflecting the adaptive potential of GBS to very diverse host environments.

1.1. Neonatal disease

Neonatal infections represent a significant cause of mortality and morbidity among infants. Particularly in regard to GBS, the incidence of neonatal infections saw a dramatic shift during the 20th century. This species was initially isolated from humans during the 1930s (Lancefield and Hare 1935), and until 1965 occasional reports of severe GBS disease continued to be detected (Eickhoff et al 1964, Kexel and Schoenbohm 1965). However, it was only during the 1970s that GBS was acknowledged as the leading cause of neonatal meningitis in the US and in Europe (Baker et al 1973, Barton et al 1973). At the time, one study estimated that the incidence rate of new infections was of two cases per 1000 live births, with a mortality rate of 50% (Franciosi et al 1973). Currently, the incidence rate of GBS disease in newborns varies considerably worldwide, but is estimated at an average of 0.53 per 1000 live births, with a mean case fatality ratio of 9.6% (Edmond et al 2012). The burden of GBS disease is substantially higher in African countries (Edmond et al 2012, Seale et al 2016), whereas it was reported to present the lowest risk in southeast Asia (Edmond et al 2012).

When infection occurs during pregnancy, complications arise that may lead to abortion or preterm labour (Allen et al 1999). Following birth, most GBS infections occur within the first week and are classified as “early-onset disease” (EOD). GBS and *Escherichia coli* are the most common agents involved in early cases of infection within the first days of the newborn’s life (Simonsen et al 2014). By contrast, if the infection occurs after the first week and up to three months of age, it is considered “late-onset disease” (LOD)

(Edwards and Baker 2005), in which the most frequent causes of infection are coagulase-negative staphylococci, *Staphylococcus aureus*, enterococci and GBS (Dong and Speer 2015). Instances of GBS infections have been detected after three months, which have led some to start adopting the term "ultra LOD" (Bisharat et al 2005, Teatero et al 2016).

In EOD, newborns are infected by GBS strains that are transmitted from their mothers during or just before birth. GBS-contaminated amniotic or vaginal fluid is transmitted from mother to baby through aspiration, resulting in the initial dissemination of the bacteria in the respiratory epithelium and in the child's bloodstream. In this case, clinical symptoms most often manifest in the form of pneumonia or sepsis (Doran and Nizet 2004, Edwards and Baker 2005) (Figure 2).

Late-onset disease, occurring later after the baby's birth, can also result in septicaemia (Edwards and Baker 2005), but quite frequently (at least 40-60% of cases) the bacterium is then able to trespass the blood-brain barrier and induce a clinical syndrome of meningitis (Doran and Nizet 2004, Johri et al 2006) (Figure 2). In LOD the actual contamination source is more difficult to track unambiguously because, in addition to the mother, GBS can also be transmitted from the community or the hospital environment.

The most effective strategy to prevent GBS infections during the perinatal period consists in *intra partum* antibiotic prophylaxis, which has been implemented since the 1990s in some high-income countries, including France and the US (Edmond et al 2012). For this purpose, penicillin is the most frequent treatment choice, as GBS strains are universally sensitive to β-lactam antibiotics (Daley and Garland 2004). The revised 2010 guidelines for the prevention of perinatal group B streptococcal disease, issued by the Centers for Disease Control and Prevention (CDC), state that if a patient has a history of penicillin allergy, macrolides such as erythromycin and clindamycin should be considered as primary alternatives for treatment. These preventive measures have only been shown to be effective at reducing the risk of EOD. Thus, even though there was a drastic decrease in the incidence of early infections following these treatment strategies, the level of LOD has remained largely the same (Poyart et al 2008).

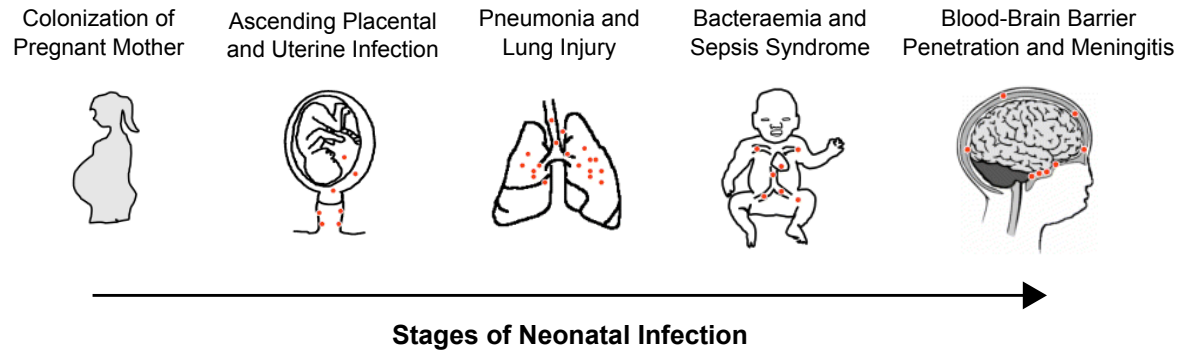


Figure 2. Infection pathways of GBS in neonatal disease. Adapted from Doran and Nizet 2004.

1.2. Bovine mastitis

Initially in 1887, Norcard and Mollereau identified GBS as a cause of bovine mastitis (Nocard and Mollereau 1887). Mastitis refers to an inflammatory disease in the mammary gland, which can be subdivided into clinical (symptomatic) or subclinical (asymptomatic) mastitis (McDonald 1979).

The most common cause of inflammation is the presence of pathogenic microorganisms in the udder. Although the teat skin cells are able to mostly protect the udder from these infectious agents, after the milking procedure the sphincter muscles remain dilated for 1-2 hours, causing the mammary gland to be at an increased risk of infection. During this period, pathogenic organisms may trespass the teat canal, multiplying and releasing toxins, enzymes and surface proteins. This in turn induces an inflammatory response from the host, increasing the number of polymorphonuclear neutrophils, phagocytes and other leukocytes. The degree of immune response depends on the causative agent, lactation stage, age and health status of the cow (Harmon 1994). Usually, a somatic cell count above 200 000 cells/ml is indicative of mastitis (Ruegg and Reinemann 2002). Due to this somatic cell increase, the composition and quality of the milk is significantly affected, diminishing its economic value. Therefore, mastitis is considered one of the costliest diseases of the dairy industry, with a financial burden ranging on average from 61€ to 97€ per infected cow per year, worldwide (Hogeveen et al 2011).

The bovine milk was long believed to be a bacteria-poor environment, but the advent of more sensitive culture-independent methods have revealed that the milk microbiota

consists in a complex community with many important biological roles (Addis et al 2016). Over 150 different species have been isolated from the bovine mammary gland, with most of them belonging to three major groups of organisms: *Staphylococcus*, *Streptococcus* and coliforms (Kuang et al 2009, Watts 1988). Mastitis-causing agents can either be contagious, if they are generally found solely in the udder, or environmental, if they also colonize the surrounding environment. Typically, contagious organisms are responsible for more persistent subclinical infections, since they may go unnoticed for extended periods of time. GBS is one of the most common contagious pathogens, alongside other species such as *S. aureus* and *Streptococcus dysgalactiae*. In contrast, environmental agents are more easily detected as they cause a more clinically severe case of mastitis. The most prominent environmental pathogens are *Streptococcus uberis*, *E. coli* and *Klebsiella pneumonia* (McDonald 1979). Recently, however, with the development of more precise diagnostic techniques, the classical distinction between contagious and environmental agents is fading. GBS in particular, once thought as a strictly contagious pathogen transmitted during milking, was recently claimed to possibly have an environmental route of transmission (Figure 3). A study on Norwegian dairy farms showed that GBS was able to transiently colonize the intestinal tract of cattle, in addition to the drinking water and the barn environment (Jorgensen et al 2016). However, given that in each farm studied in this work the GBS population found in the environment is quite homogeneous and similar to the one isolated from contaminated cattle, the presence of GBS in the surrounding environment could be indicative of cross-contamination from infected bovines rather than a true environmental colonization.

Until the second half of the 20th century, GBS predominated as a cause of intramammary infections, at which point several control measures were put in place to more efficiently treat infected cattle and to prevent further dissemination within a herd (Neave et al 1969). Consequently, a significant reduction in the number of GBS infections was registered throughout Europe. Yet, the prevalence of GBS mastitis remains elevated in certain countries, such as Portugal (Rato et al 2013) and Germany (Tenhagen et al 2006), while in others it has re-emerged (Mweu et al 2012).

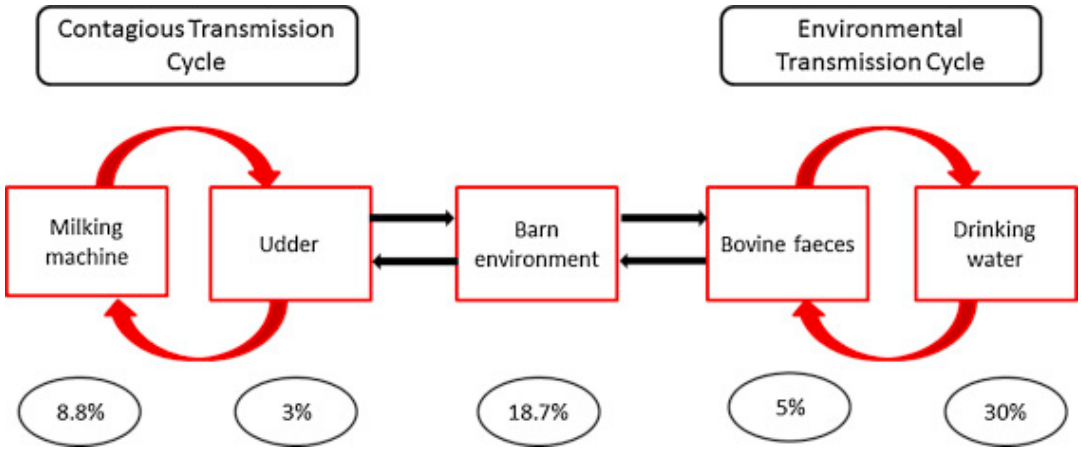


Figure 3. Transmission cycles of GBS in the bovine environment. Percentage values correspond to the proportion of samples testing positive in a longitudinal study of four Norwegian dairy herds during a 12-month period (Jorgensen et al 2016).

1.3. Disease in other hosts

Although rarely isolated from aquatic species until the 1970s, GBS has been responsible for epidemic outbreaks of invasive disease in several fish farms, with a mortality rate of up to 30%. Its infectivity is not restricted to a single species, as GBS has been isolated from tilapia, sea bream, bass, trout, as well as from poikilotherms (frogs) and aquatic mammals (Eldar et al 1994, Evans et al 2002, Wilkinso et al 1973).

In addition to humans, bovines and fish, GBS has also been found to cause disease in a wider range of animals. In the Horn of Africa, it has been frequently obtained from camels diagnosed with mastitis and with wound infections (Fischer et al 2013, Tibary et al 2006). Other hosts, such as cats, dogs, horses, monkeys and mice can also be colonized by GBS (Lammler et al 1998, Yildirim et al 2002a, Yildirim et al 2002b). However, the pathogenic potential and infection mechanism of GBS within these hosts is poorly understood. Most intriguing is that with such a wide host range the original ancestral reservoir of GBS remains unknown.

1.4. Colonization and virulence factors

Many genes and their corresponding products have been implicated in the ability of GBS to colonize and infect its hosts (Figure 4). But, functional studies have primarily focused on their role in the context of human infection. Through signature-tagged transposon mutagenesis (STM), several genomic loci responsible for a variety of bacterial processes were originally shown to be required in a neonatal rat sepsis infection model (Jones et al 2000). Since then, several studies have shown the role of specific factors for GBS infection and colonization. The pathogenesis of GBS relies principally on three mechanisms: (i) ability to colonize and cross tissue barriers within the host environment; (ii) ability to evade the host defence mechanisms; and (iii) expression of virulence factors that cause damage to the host (Mitchell 2003). With approximately two decades of research, there is extensive literature on the role of various genes and products implicated in the virulence of GBS. Therefore, in the following sections I will review the current understandings of functional traits most important and pertinent to the work here presented.

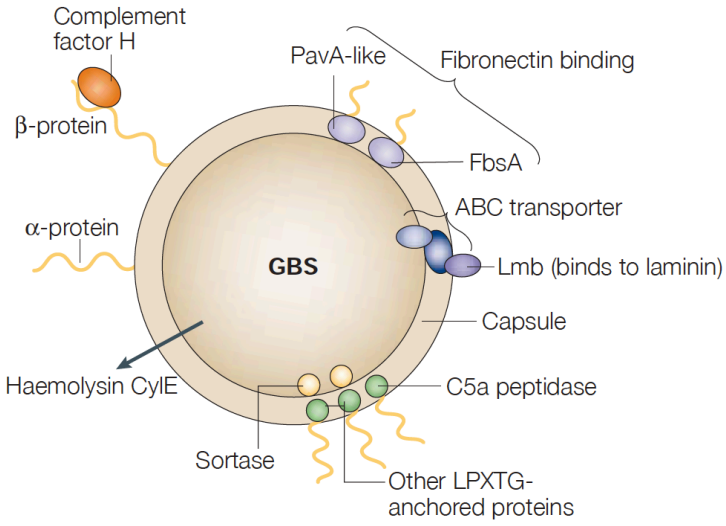


Figure 4. Main surface and secreted virulence-associated factors in GBS. Adapted from Mitchell 2003.

1.4.1. Surface polysaccharides

Common to all GBS strains is the group B-specific antigen (GBC), originally identified by Rebecca Lancefield (Lancefield 1934). This peptidoglycan-anchored antigen is

represented by a complex structure composed of rhamnose, galactose, N-acetylglucosamine and glucitol (Michon et al 1987). Although the biological role of GBC is not entirely known, a recent report showed that GBC-deficient mutant cells displayed several abnormalities related to morphology and growth (Caliot et al 2012), suggesting that GBC might consist in a structural component of the cell wall.

Surface antigens are major contributors to the pathogenic potential of GBS, and one of the most important ones is the capsule polysaccharide (CPS) (Mitchell 2003). Anticapsular antibodies were shown to confer protective immunity in an animal model, so special interest has been placed on the study of the capsule and of its potential as a target for vaccination against GBS (Baker et al 1988, Edwards and Baker 2005). The capsule operon comprises four conserved genes belonging to the *neu* operon (*neuA*, *neuB*, *neuC* and *neuD*), six genes conserved across all capsular types (*cpsA*, *cpsB*, *cpsC*, *cpsD*, *cpsE* and *cpsL*) and an additional variable set of six that determine the capsular type of each strain (*cpsF*, *cpsG*, *cpsH*, *cpsI*, *cpsJ*, *cpsK*, *cpsM*) (Cieslewicz et al 2005). The *neu* genes are responsible for the synthesis of sialic acid and subsequent sialylation of the capsule. This process provides a critical function to the polysaccharide capsule, as it allows the bacteria to resist opsonophagocytosis by avoiding activation of the alternative pathway of the complement system (Marques et al 1992). Despite its important role, the capsule polysaccharide is not expressed by a significant, but variable, proportion of GBS isolates. In the human-derived population, they normally account for 5-20% of the isolates (Ippolito et al 2010), while in bovines 30-77% of GBS strains are non-capsulated (Kong et al 2008). Recently, the molecular basis for capsule loss has been further investigated, suggesting that distinct mutations predominantly within *cpsE* were responsible for inactivation of capsule synthesis (Rosini et al 2015). This raises concerns for future implementation of an anti-CPS vaccine that is currently under development to prevent GBS colonization (Kobayashi et al 2016).

1.4.2. Surface and secreted proteins

Cell adhesion plays an essential role in the survival and invasion of the host environment, and there are various proteins located on the surface of the bacteria involved in this process. One of the major groups of surface proteins with adhesive

properties is characterized by the presence of a LPXTG motif that promotes covalent attachment to the cell wall peptidoglycan. The hypervirulent GBS adhesin (HvgA) is an example of a LPXTG protein functioning as an important virulence factor in GBS. It includes two main variants: *hvgA* and *bibA* (Santi et al 2007, Tazi et al 2010). Both consist in a variable core with 50-60% sequence identify, flanked by conserved 5' and 3' ends. Importantly, it was shown that HvgA is overexpressed *in vivo* and efficiently adheres to intestinal epithetial cells. Further experiments showed that HvgA contributes to the translocation of the intestinal barrier and to the crossing of the blood-brain barrier. Altogether, these observations demonstrated that HvgA is a major virulence factor in GBS, associated with the onset of meningitis and LOD (Tazi et al 2010).

Two types of pilus-like structures have been identified in gram-positive species: one characterized by short thin rods, while another represented by a longer and more flexible pilus (McNab et al 1999, Mora et al 2005). In GBS, the composition of the pilus comprises three structural proteins, a backbone protein, two ancillary proteins and two class C sortase enzymes (Rosini et al 2006). These pilin-specific enzymes promote pilus polymerization, while the housekeeping transpeptidase SrtA is responsible for termination of pilus assembly and for its covalent attachment to the cell wall peptidoglycan through the LPXTG motif (Dramsi et al 2006). Each component of the pilus locus was shown to play an important role in GBS colonization and disease progression. The ancillary proteins help initiate adherence to the host tissue (Maisey et al 2007), whereas the backbone protein further contributes to invasion and translocation within the host cells (Pezzicoli et al 2008). Specifically in GBS, three distinct genomic islands have been found: Pilus island 1 (PI-1), Pilus island 2a (PI-2a) and Pilus island 2b (PI-2b) (Lauer et al 2005). For PI-2, the two genomic variants PI-2a and PI-2b are mutually exclusive within GBS, as they share the same insertion site (Brochet et al 2006). However, functional differences between the two variants have been described. Of the two PI-2 islands, PI-2a was especially implicated in biofilm formation (Rinaudo et al 2010), whereas the Spb1 protein within PI-2b was revealed to contribute to an enhanced survival inside macrophages (Chattopadhyay et al 2011). Given the highly immunogenic properties of proteins comprising the pilus, a pilus-based vaccine has been shown to confer protective immunity in an *in vivo* mouse model (Margarit et al 2009).

Surface proteins derived from the alpha component of the antigen C protein, termed the alpha-like family, are of especial importance in GBS. They have been implicated in virulence by promoting invasion, adherence and translocation of GBS across epithelial cells (Baron et al 2007, Bolduc and Madoff 2007). The alpha protein is distinguished from the beta protein, the other component of the C antigen, by its ability to resist trypsin (Ferrieri 1988). So far, six members of the alpha-like family have been identified: alpha, Rib, Alp1 (epsilon), Alp2, Alp3 (R28) and Alp4 (Kong et al 2002, Wastfelt et al 1996). A particular characteristic common to this family of proteins is the presence of a series of identical, tandem repeating motifs that follow a unique N-terminal part. Genes encoding the alpha family proteins are mostly conserved across different strains, with variation limited to the number of repeat units (Kong et al 2002, Lindahl et al 2005). Because of their immunogenicity, alpha-like proteins have been studied as potential vaccine candidates. Indeed, *in vivo* studies evidenced that vaccine immunization with alpha and Rib conferred protective immunity against GBS infection (Kling et al 1997, Stalhammarcarlemalm et al 1993). Interestingly, the immunogenicity of the purified Alp proteins was shown to be inversely correlated with the number of repeats (Gravekamp et al 1997, Madoff et al 1996).

Another virulence-associated surface protein is the C5a peptidase ScpB. This serine protease is responsible for inactivating the human C5a protein produced by the complement system (Wexler et al 1985). This in turn may interfere with neutrophil recruitment and reduce the host inflammatory response (Rubens et al 1991). Additionally, it has also been shown that the C5a peptidase may facilitate binding to fibronectin while also functioning as an invasin of human epithelial cells (Beckmann et al 2002, Cheng et al 2002). The gene encoding this protein, *scpB,* is located in a putative composite transposon shared with the gene coding for the surface, laminin-binding protein Lmb (Franken et al 2001). ScpB is also almost identical to ScpA of *Streptococcus pyogenes* (> 95% residue identity), suggesting interspecies horizontal exchange (Bohnsack et al 2000).

Serine-rich repeat (SRR) glycoproteins are important mediators of bacterial attachment to human platelets. In GBS, two mutually exclusive SRR proteins have been described, Srr1 and Srr2 (Brochet et al 2006, Seifert et al 2006). They have been shown to promote

colonization of the vaginal tract (Sheen et al 2011) and contribute to the pathogenic potential of GBS (Seo et al 2012, van Sorge et al 2009). Controlling the translocation of these proteins across the cytoplasmic membrane is the Sec2 system, encoded by the *secA2-Y2* locus (Mistou et al 2009). The *srr1-secA2* locus, widely distributed across GBS, comprises at least 15 genes: eight glycosyltransferases, six proteins belonging to the SecA2 secretion system, and the transcriptional regulator *rga* previously shown to control the expression of *srr1* (Mistou et al 2009).

Fibrinogen binding proteins are another class of antigens with an important function in GBS. Three different proteins, encoded by *fbsA*, *fbsB* and *fbsC* have been identified. The surface-located FbsA was shown to promote adherence of GBS to epithelial cells, but not sufficient on its own for host invasion (Schubert et al 2004). By contrast, FbsB is a secreted protein determined to be most important for the invasion process. In fact, its absence was revealed to not impair the actual binding to fibrinogen (Gutekunst et al 2004). A third variant (FbsC) presenting both adhesive and invasive properties was more recently discovered (Buscetta et al 2014).

Secreted proteins play an important role at various stages of colonization and infection, while also associated with particular idiosyncratic traits of GBS. For instance, β-haemolysis was first described in GBS in 1934 (Todd 1934) as a characteristic zone of lysis around colonies grown in blood agar. The 12-gene operon *cyl*, suggested to be involved in the biosynthesis of fatty acids, was later found to be responsible for the production of the β-haemolysin/cytolysin of GBS (Pritzlaff et al 2001). Indeed, nonhaemolytic strains, representing approximately 1-5% of human GBS isolates, were found to carry various mutations within *cyl* (Brimil et al 2006, Merritt and Jacobs 1976, Spellerberg et al 1999). Given that the β-haemolytic activity is a distinctive trait of GBS, nonhaemolytic strains are a significant concern for GBS diagnostics. Anecdotally, the expression of *cyl* was also long associated with the production of an orange carotenoid pigment (Tapsall 1987), and it was recently discovered that the GBS haemolysin and the ornithine-rhamnolipid pigment are in fact the same molecule (Rosa-Fraile et al 2014, Whidbey et al 2013). This secreted toxin is associated with GBS virulence as it induces proinflammatory responses, causes damage to human host tissues and contributes to disease progression (Doran et al 2003, Ring et al 2002). Additionally, the haemolytic

pigment is involved in the penetration of the amniotic barrier and fetal injury by GBS (Whidbey et al 2013, Whidbey et al 2015).

The CAMP factor, short for the names of the three researchers that discovered it (Christie, Atkins and Munch-Petersen), is encoded by the *cfb* gene (Schneewind et al 1988) and is an extracellular protein considered a virulence factor for its pore-forming and lytic activity. It was also found to exhibit a protein A-like binding to the Fc fragments of immunoglobulins, earning it the designation of protein B (Jürgens et al 1987). When interacting with the sphingomyelinase from *S. aureus*, the CAMP factor induces a reaction that creates a distinctive zone of haemolysis on blood agar plates. Since its discovery in 1944 (Christie et al 1944), the CAMP reaction has been used in diagnostic microbiology to identify the presence of GBS.

1.4.3. Gene regulation

Transcriptional regulation is an important mechanism of adaptation in many bacterial species (Yang et al 2011). A recent transcriptome analysis of GBS revealed the existence of many regulatory mechanisms, some of which were differentially expressed in distinct environmental conditions (Rosinski-Chupin et al 2015). By rapidly modifying the expression of numerous genes in response to environmental stimuli, two-component regulatory systems (TCS) are especially implicated in the ability of GBS to colonize, infect and persist in the host environment. Moreover, given their widespread impact, small modifications in the function or structure of these regulators are enough to have dramatic phenotypic effects, making them ideal targets for rapid evolution. Two-component systems comprise genes coding for a sensor histidine kinase (HK) and an associated response regulator (RR). Once the HK senses an external stimulus, conformational changes phosphorylate and activate the C-terminal transmitter domain. This in turn activates the RR through phosphorylation, which will respond by modifying the gene expression of its target genes. More than 20 TCS have been identified in GBS (Faralla et al 2014) and I will review below the two most pertinent to this work.

The first, and one of most well studied regulators, is the “Control of virulence regulator and sensor” (CovRS) system (Jiang et al 2005, Lamy et al 2004). With an orthologous

variant in *S. pyogenes* (Heath et al 1999), CovR is a DNA-binding protein shown to regulate the expression of approximately 7-27% of the genetic repertoire in GBS, which includes many virulence determinants (Jiang et al 2008, Lamy et al 2004). Its regulatory functions are influenced by fluctuations in environmental conditions, such as high glucose and low pH conditions (Di Palo et al 2013, Park et al 2012). The role of CovR as a repressor is well documented, as it binds to the promoter of major virulence-related genes, such as *bibA*/*hvgA* and the *cyl* operon, blocking their expression. Therefore, reduced activity of this system derepresses its virulence-associated targets, and is frequently linked to an increased pathogenic potential of GBS (Jiang et al 2005, Jiang et al 2008). In spite of functioning mainly as a repressor, CovR is essential for GBS virulence (Lamy et al 2004). Nevertheless, how CovR increases the expression of certain genes, such as *cfb*, is still unclear. The serine/threonine kinase Stk1 and the Abi-domain protein Abx1 have both been shown to inhibit activity of the CovRS system through distinct mechanisms (Firon et al 2013, Lin et al 2009). Stk1 was shown to inactivate CovR through phosphorylation of its T65 residue, while Abx1 blocks the activity of CovR by directly interacting with CovS (Figure 5).

The other TCS with important implications for virulence and adaptation of GBS to its environment is DltRS. This system is involved in the regulation of the *dlt* operon, which is responsible for D-alanylation of lipoteichoic acids (LTAs) (Poyart et al 2001). LTAs consist in amphiphilic polymers of polyphosphoglycerol or polyphosphoribitol anchored to the cytoplasmic membrane by a glycolipid. The incorporation of D-alanine decreases the net anionic charge of LTAs, playing a role in many important functions, such as biofilm formation (Fabretti et al 2006), adherence to mammalian cells (Abachin et al 2002) and tolerance to low pH (Boyd et al 2000). Most notably, D-alanylation has been shown to confer resistance to cationic antimicrobial peptides, which are innate immune factors produced by the human host or by the intestinal microflora (Peschel et al 1999). Each of the four genes of the *dlt* operon is essential for the D-alanylation process: *dltA* codes for the protein that ligates D-alanine to the D-alanyl carrier protein encoded by *dltC*; *dltB* codes for a D-alanyl transport protein; and *dltD* encodes a membrane protein that catalyzes the ligation of D-alanine to LTA (Reichmann et al 2013).

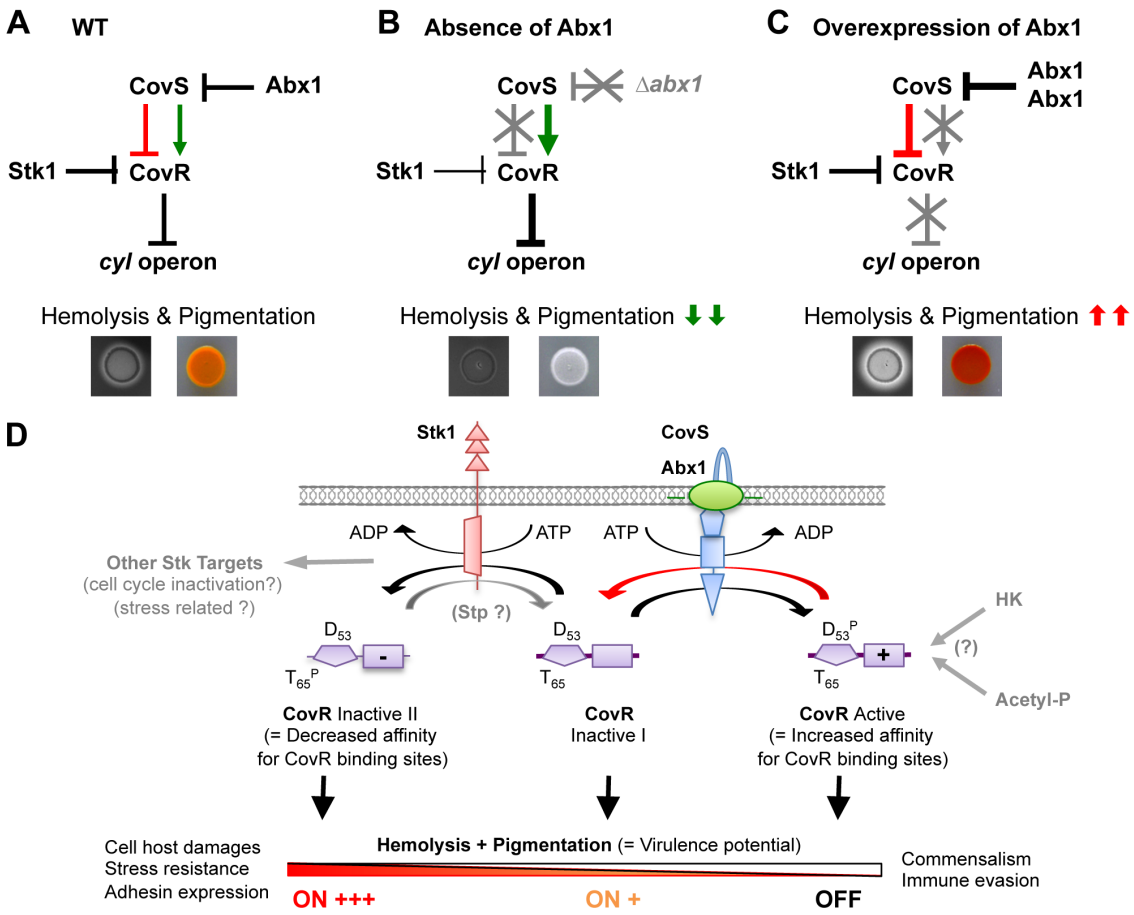


Figure 5. Model of the regulatory network of the CovRS system in GBS.
(A) Diagram depicting the balance maintained by the kinase (green arrow) and the phosphatase activity (red lines) of CovS. (B) Without Abx1, CovS remains in its kinase-competent form that activates CovR and represses the *cyl* operon. (C) With an excess of Abx1, the phosphatase-competent form of CovS is preferred. Stk1 may further inactivate CovR, derepressing *cyl* and other CovR targets. (D) Representation of CovRS signaling affected by CovS (blue), Abx1 (green), Stk1 (red) and CovR (purple) (Firon et al 2013).

1.4.4. Transporters and antibiotic resistance

Transporters play an important role in many aspects of bacterial physiology (Davidson et al 2008). Most correspond to a class of ATP-binding cassette (ABC) transporters that comprise a highly conserved ATP-hydrolyzing domain contributing to the activity of an associated transmembrane subunit. They can be primarily classified into importers or exporters (Davidson et al 2008). Importers mediate the uptake of various substrates, including mono and oligosaccharides, amino acids, vitamins and metals. Exporters, on the other hand, participate in the secretion of molecules such as lipids, drugs and toxins, in addition to peptides, polysaccharides and metals (Davidson et al 2008).

In GBS, many transporters provide functions that are essential for survival, metabolism, virulence and antibiotic resistance. CylAB is an ABC-type transporter shown to promote the export of multi drug resistance substrates (Gottschalk et al 2006). MtsABC, first described in *S. pyogenes*, encompasses a lipoprotein, a hydrophobic integral membrane protein and an ATP-binding protein with affinity to metal nutrients, such as iron, zinc and manganese (Janulczyk et al 2003). Another important class of transporters corresponds to the phosphotransferase systems (PTS), which are commonly associated with sugar uptake (Davidson et al 2008). In GBS, one of the most prominent examples is contained within the Lac.2 operon, shared with *S. dysgalactiae* (Richards et al 2011). The PTS subunit IIABC within this region has been shown to contribute to the uptake and metabolism of lactose, which is one of the key functions for the adaptation of GBS to the bovine udder (Richards et al 2013).

In terms of resistance, ABC transporters are frequently found in GBS as contributing to a reduced susceptibility towards bacteriocins, a class of antimicrobial peptides produced by bacteria. Nisin, a broadly used bacteriocin belonging to the lantibiotic group, is encoded by the *nis* operon, which comprises genes for synthesis, production and self-immunity, the latter conferred by an ABC exporter (Richards et al 2011, Wirawan et al 2006) (Table 1). The antimicrobial peptides macedocin and salivaricin are also targeted by ABC transporters encoded by the *mcd* and *sal* operons, respectively (Papadelli et al 2007, Ross et al 1993). An example of antibiotic resistance mediated by ABC transporters is the *mef* family of genes, encoding efflux pumps that reduce the susceptibility to macrolides (Cai et al 2007) (Table 1) .

However, resistance is not solely dependant on the action of exporter detoxification systems. For instance, more than 20 genes associated with tetracycline resistance code for energy-dependent efflux proteins, ribosomal modification proteins or inactivating enzymes, depicting various alternative means of antibiotic resistance (Roberts 2005). In GBS, tetracycline resistance is essentially due to the action of ribosomal modification systems encoded by *tetM* or *tetO* (Da Cunha et al 2014, Poyart et al 2003) (Table 1). Another group of antibiotic resistance genes commonly found in GBS belongs to the *erm* class (Heelan et al 2004, Poyart et al 2003), coding for resistance to the macrolide-lincosamide-streptogramin B (MLS) antibiotic class (Table 1). Among the *erm* family,

ermA, *ermB* and *ermTR* are the most frequent determinants found, encoding rRNA methyltransferases that confer resistance through modification of the ribosomal site targeted by the antibiotic (Dogan et al 2005). The molecular basis of cross-resistance to lincosamides, streptogramin A and pleuromutilins conferred by the *lsa* genes has been a matter of debate (Douarre et al 2015, Malbruny et al 2011), but recent evidence suggests they might be related to a mechanism of ribosomal protection (Sharkey et al 2016). The enzymatic activity of nucleotidyltransferases encoded by the *lnu* genes also provides resistance to lincosamides (Gravey et al 2013) (Table 1). Finally, resistance to kanamycin and streptomycin can additionally be found in GBS, linked to the activity of aminoglycoside modifying enzymes encoded by the *aphA3* and *aad6*-related genes, respectively (Ounissi et al 1990, Poyart et al 2003) (Table 1).

Table 1. Antibiotic/bacteriocin resistance genes most frequently identified in GBS.

Gene	Antibiotic/Bacteriocin targeted	Resistance mechanism
aadE	Streptomycin	Aminoglycoside modification
aphA3	Kanamycin	Aminoglycoside modification
ermA	Macrolides, lincosamides and streptogramins B	Ribosomal target modification
ermB	Macrolides, lincosamides and streptogramins B	Ribosomal target modification
ermTR	Macrolides, lincosamides and streptogramins B	Ribosomal target modification
lnuB	Lincosamides	Lincosamide nucleotidylation
lsaC	Lincosamides and streptogramins A	Ribosomal protection
mefE	Macrolides	Efflux pump
nis	Nisin	Efflux pump
tetM	Tetracycline	Ribosomal target modification
tetO	Tetracycline	Ribosomal target modification

1.5. Population structure and host specificities

1.5.1. From capsular to whole-genome typing

The development of molecular biology tools for diagnostic and epidemiological purposes has improved the detection and treatment of GBS infections. Besides the CAMP reaction, culture-based methods such as the API 20 and the VITEK system are able to

positively identify the presence of GBS (Ligozzi et al 2002, Poutrel and Ryniewicz 1984). Moreover, using a selective and differential culture medium, termed Granada, it is also possible to distinguish GBS from other microorganisms based on the detection of a GBS-specific orange pigment (Rosa-Fraile et al 1999). Since all these techniques rely on a strictly biochemical approach, they can be replaced by PCR-based platforms in order to achieve a more sensitive, specific and cost-effective detection, with many recommended protocols having been published thus far (Almeida et al 2013, Shome et al 2011).

After positive detection of GBS, further classification of the specific strain can be achieved with several approaches. Before, serotyping was one of the only methods available to quickly determine the relatedness between strains. This technique is able to identify one of the ten capsular types (Ia, Ib, II, III, IV, V, VI, VII, VIII and IX), through an antibody-antigen agglutination reaction of the capsular polysaccharides present in the surface of GBS (Elliott et al 2004). But, because of inactivating mutations within the genes responsible for capsule production, a small percentage of strains do not cross-react with antisera and are thus non-typeable (Bisharat et al 2004, Rosini et al 2015). Therefore, a few PCR schemes have been published for a more reliable identification of the *cps* genes and corresponding capsular types (Imperi et al 2010, Poyart et al 2007). However, given the recombinogenic potential of GBS, inferences based on the capsular profile can be misleading, as strains with the same serotype may in fact have very distinct genomic backgrounds and evolutionary histories (Jones et al 2003, Luan et al 2005).

Pulse-fied gel electrophoresis (PFGE) has been frequently used as a more discriminatory approach for epidemiological purposes (Fasola et al 1993, Oliveira et al 2005, Rato et al 2013). In it, DNA molecules from each sample are first digested using specific restriction endonucleases, such as *Sma*I. The resulting fragments are then separated in an agarose gel electrophoresis by regularly switching the voltage between three different directions. Although the discriminatory potential can be greater than with traditional serotyping, PFGE is time-consuming, lacks consistency, and has also been shown to sometimes present insufficient resolution to accurately differentiate epidemiologically unrelated strains (Pillai et al 2009) .

Alongside PFGE, multi-locus sequence typing (MLST) is among the most popular genotyping methods and is still widely used worldwide in spite of its limited resolution (Jones et al 2003, Luan et al 2005). This technique consists in the PCR amplification of seven genes of GBS (*adhP*, *pheS*, *atr*, *glnA*, *sdhA*, *glcK* and *tkt*) corresponding to constitutive regions with housekeeping functions. Subsequent sequencing of each of the genes allows the user to assign a code to each allele and a final sequence type (ST) based on the combination of the different alleles identified. Compared to the traditional serotyping, MLST is a more discriminative tool and was the basis for classifying the population structure of GBS into different clonal complexes (CC) (Jones et al 2003). For consistency purposes, a clonal complex is henceforth considered as a group of STs with no more than two allele differences from their "founder ST".

After the widespread implementation of MLST, a relation between strain capsular type and ST has been frequently observed (Bisharat et al 2004). For instance, most CC17 and CC19 strains are serotype III, CC61/67 are frequently serotype II, CC1 are associated with serotype V and CC23 with serotype Ia. However, even if the majority of CCs are represented by one predominant serotype, additional strains within each lineage have been shown to exhibit distinct capsular types (Bisharat et al 2004, Da Cunha et al 2014).

The locus coding for the "Clustered regularly interspaced short palindromic repeats" (CRISPR) system has also been found to be useful as an epidemiological marker (Lopez-Sanchez et al 2012). It comprises a series of DNA repeats separated by non-repetitive segments called spacers (Jansen et al 2002). Because of their highly dynamic nature, the organization of the spacer sequences was shown to closely mirror the population structure of GBS, with a strong association between ST and the CRISPR profile (Lopez-Sanchez et al 2012).

1.5.2. Clinical and host association

Several studies have attempted to establish a correlation between the CPS type and the pathogenic potential of GBS. The diversity of GBS strains infecting neonates in the case of EOD is very similar to that detected in carriage, as the GBS population colonizing the newborn is most often the result of mother-to-child transmission. This means that the

frequency of each serotype detected is quite variable, but there is a particular over-representation of serotypes Ia and III (Edmond et al 2012). In particular, there are many reports detailing a strong association between serotype III and the incidence of meningitis and LOD (Bidet et al 2003, Jones et al 2003, Luan et al 2005). Among CPS type III, CC17 is considered "hypervirulent" as strains belonging to this lineage are most predominantly found in neonatal sepsis and meningitis in the context of LOD (Bidet et al 2003, Jones et al 2003, Luan et al 2005). This lineage is also frequently detected in EOD cases, but to a lesser extent. In fact, strains isolated from carriage or EOD are associated with CC1, CC10, CC17, CC19 and CC23 at varying proportions across different studies (Manning et al 2009, Martins et al 2007).

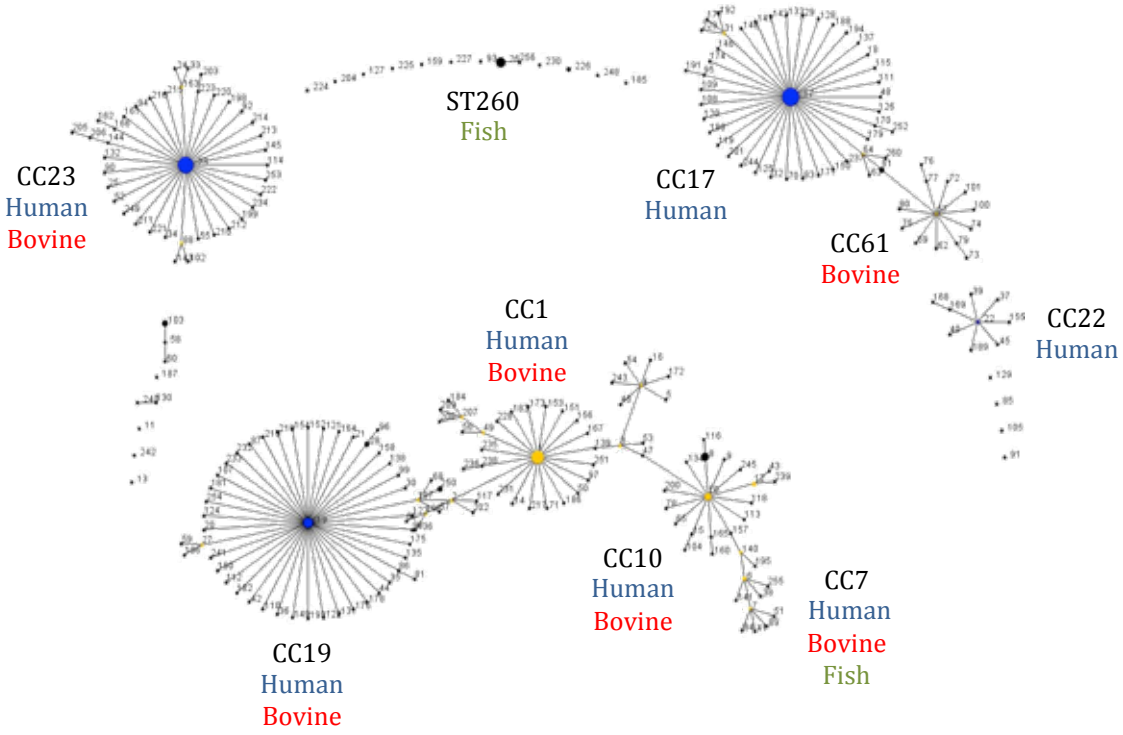


Figure 6. GBS clonal complexes defined by MLST, with their associated host origin.

In correlating the diversity of GBS to its host specificity, population studies have shown that GBS comprises several generalist lineages found across its different hosts, combined with a few host-restricted clones (Sorensen et al 2010) (Figure 6). Genomic studies so far have mostly confirmed the population structure previously established for GBS using MLST-based methods (Da Cunha et al 2014, Rosini et al 2015, Seale et al 2016). In this regard, it was shown that most human carriage and clinical isolates collected worldwide belong to one of five major clonal complexes (CC1, CC10, CC17,

CC19 and CC23) (Da Cunha et al 2014, Sorensen et al 2010). Some of these human-associated lineages have also been detected in cattle to some degree, as bovine-isolated strains frequently belong to CC1, CC7, CC10, CC23 and CC26 (Haenni et al 2010, Rato et al 2013, Sorensen et al 2010). However, the majority are part of the host-restricted CC61, a CC in which only one human strain has been retrieved. Regarding the other GBS hosts, two main clones have been specifically associated with fish infections: CC7 and ST260/261. The former is a host generalist group encompassing strains that have also been isolated from humans and bovines (Evans et al 2008). ST260/261, however, is unique to fish hosts (Delannoy et al 2013) and was originally thought to belong to a different streptococcal species, *Streptococcus difficile* (Eldar et al 1994).

Whole-genome sequencing revolutionized evolutionary studies of human bacterial pathogens, providing a more fine-tuned characterization of their adaptation to different environments. However, genome-wide studies of the evolution and diversity of microbial populations are still striving to reliably assess the functional outcome of genomic changes predicted to be under adaptive evolution, as well as their role in pathogenesis and survival within the host.

2. Genomics of bacterial evolution and adaptation

2.1. Phylogenetic and population genomic methods

2.1.1. General concepts of molecular evolution

Evolution is defined as a process of change that occurs from generation to generation. In all living organisms, the majority of spontaneous DNA mutations occur due to errors introduced during replication, which can cause varying degrees of phenotypic repercussions. The classical model and definition of a gene in bacterial genomes, first deduced from pioneer studies such as those from François Jacob and Jacques Monod (Jacob and Monod 1961), has been refined throughout the last decades. Generally, a gene can be described as a DNA region that can affect phenotypic traits through expression of a functional protein product. The promoter sequence of a gene is where the RNA polymerase, with an associated sigma factor, will bind to initiate transcription. Then, within the mRNA transcript, the ribosomal binding site, also known as the Shine-Dalgarno sequence, is the location where the ribosome will bind for protein translation to begin and is usually similar to the consensus sequence of AGGAGG. Lastly, the so-called coding sequence (CDS) encompasses a set of nucleotide triplets (codons) within the mRNA that will be translated into a protein, typically starting in bacteria from the amino acid N-formylmethionine (fMet) up to a nonsense codon that will stop the translation process. These fundamental tenets determining the flow of genetic information in bacterial organisms are observed in the majority of cases. However, non-canonical mechanisms of translation can also be found (de Groot et al 2014, Vockenhuber et al 2011), revealing for instance the existence of leaderless transcripts for which translation occurs immediately at the start of transcription. Additionally, studies on non-coding RNAs (ncRNAs) have debunked the classical presumption that all genetic information with biological implications is expressed as proteins. Indeed, ncRNAs have been increasingly shown to affect regulation of gene expression by interfering with RNA transcription, translation and stability (Conway et al 2014, Toledo-Arana et al 2009)

The genetic code is considered redundant, or degenerated, since all amino acids except tryptophan can be encoded by more than one triplet code. Nonetheless, small mutations affecting a single nucleotide in the encoded messages may have profound effects. Of the possible nucleotide substitutions, those that replace a purine (A or G) by a pyrimidine (T or C) or vice-versa, called transversions, are less likely to occur than transitions (purine by purine, or pyrimidine by pyrimidine) because of chemical and steric constraints. If a single nucleotide polymorphism (SNP) modifies the amino acid that will be coded, then the mutation is labelled as nonsynonymous. In general, nonsynonymous substitutions can also be referred to as missense, but when the amino acid change results in a premature stop codon and truncation of the protein product, it is considered a nonsense mutation. Synonymous substitutions, on the other hand, are point mutations that do not modify the amino acid sequence of the protein product. Until recently, synonymous mutations were also called "silent", since they were assumed to not influence protein function by conserving its structure. But, recent advances have suggested that synonymous SNPs could affect mRNA stability, protein expression and conformation, with important phenotypic implications (Bailey et al 2014).

When one or several new bases are inserted or deleted from the genome, they are collectively referred to as an indel. If multiples of three nucleotides are involved in the insertion/deletion process inside the CDS then the reading frame is unaffected, but the removal or addition of new amino acids changes the protein sequence. More significantly, however, is when an indel involves a set of nucleotides that is not a multiple of three. In these cases, a disruption to the reading frame causes a different protein to be coded, with different amino acids and length depending on where the indel occurs. This type of mutation is called a frameshift.

Mutations can also land within non-coding or intergenic regions. The 5' and 3' extremities within the mRNA transcript that are not translated into a protein are called the 5' or 3' untranslated regions (UTRs), respectively. Besides affecting UTRs, mutations might also be acquired within regulatory elements such as promoter regions, operators and terminators, which may result in many possible phenotypic effects.

After genetic variability develops in a population, mutations are either kept (fixed) or purged in subsequent generations. The term substitution or evolutionary rate denotes how frequently new mutations fix within a population. This differs from the mutation rate, which corresponds to the frequency of errors (mutations) occurring during DNA replication. Whether a new mutation becomes fixed or not depends mostly on the effects of genetic drift and natural selection (Figure 7). Drift corresponds to a change in allele frequency as a result of random sampling and fluctuations in population size (Charlesworth 2009). A drastic reduction in genetic diversity and population size due to external events is usually referred to as a population bottleneck. This is particularly evident during bacterial transmission, as only a fraction of the population from a donor is transferred to the recipient. But, it can also occur, for instance, due to compartmentalization of a population when travelling between different body parts within the host. Under the neutral theory of molecular evolution, the random fixation of genomic mutations contributes the most to the accumulation of nucleotide substitutions. However, evolution is not only a stochastic process, as it is also under the effects of selective pressures from the environment. Natural selection can act in one of two ways: diversifying (positive) selection increases the frequency of alleles underlying functional changes that provide a beneficial effect, whereas purifying (negative) selection acts to preserve genetic function by purging mutations that cause a deleterious impact on the fitness of the bacteria. The process of natural selection varies depending on the environment and on the evolutionary time scales. In longer evolutionary periods, selection has a more significant impact on the population diversity, as enough time has passed to either fix or purge individual variants from the population. Contrariwise, in shorter evolutionary scales, the effects of genetic drift are more pronounced, as selection has not yet been able to act on the genomic variation that is arising. This might be a contributing factor to the higher rate of mutational changes seen during transmission outbreaks (Ho et al 2007). Nonetheless, both the host environment and therapeutic measures, such as the use of antibiotics, represent strong selective pressures that may overshadow the effects of random drift.

Adaptive mutations are unevenly distributed along bacterial genomes, as the general observation is that the majority of microbial genes are under neutral or purifying selection, with positive selection acting only on particular genomic positions (Lieberman

et al 2014, Marvig et al 2013). In certain circumstances, a mutation sweeps through the population not due to its direct impact on the fitness of the bacteria, but rather because it is in linkage disequilibrium with an advantageous allele within the chromosome. This phenomenon is called genetic hitchhiking. In contrast to sexually reproducing eukaryotes — in which recombination is an integral part of the reproductive process — genetic hitchhiking in bacteria is not as dependent on the physical proximity of each allele within the chromosome.

With this in mind, several approaches can be taken to deduce important information from the evolutionary changes occurring in a population, including the phylogenetic diversity, the evolutionary rate and the direction of natural selection.

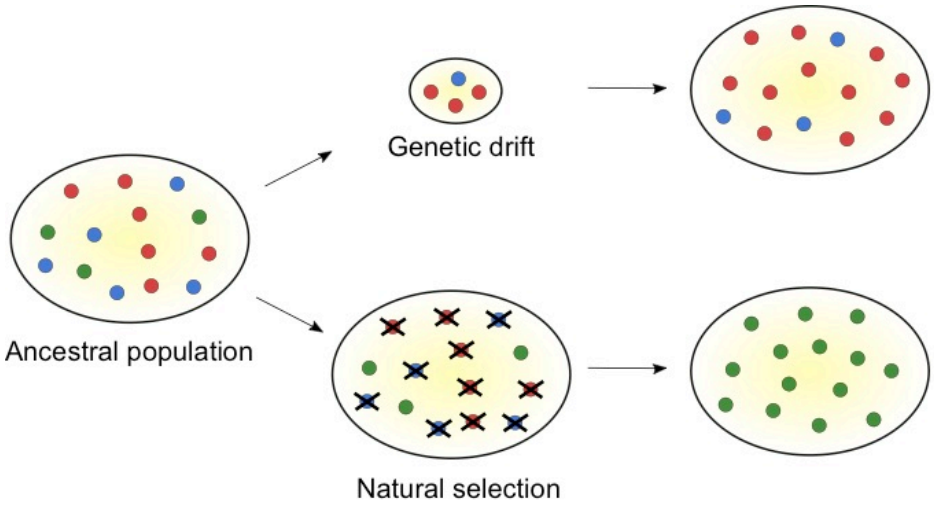


Figure 7. Evolutionary dynamics of genetic drift and natural selection. In the case of genetic drift, a reduction in population size results in a founding effect in which only a random portion of the original population remains. Under natural selection, specific variants are outcompeted and replaced by others present in the existing population due to selective pressures of the environmental conditions.

2.1.2. Phylogenetic inference

When two organisms evolve from a common ancestor, they diverge and independently acquire multiple mutations. Therefore, the genetic distance between two organisms reflects the level of divergence between them. The relationship between the different subjects can then be further established by building a phylogenetic tree. The genetic distance, and hence the changes in DNA sequence, can be modelled as a random event

around a specific substitution model. The simplest model, called the JC69 (Jukes and Cantor 1969), calculates genetic distances assuming that equilibrium frequencies of each nucleotide are 25% each and that any nucleotide substitution has the same probability of occurring. On the other extreme is the general time reversible (GTR) model (Tavaré 1986), which estimates the frequency of each nucleotide base and the rate at which every possible nucleotide substitution occurs, while assuming that symmetric substitutions (e.g. A to T and T to A) occur at the same rate. Over-parameterization can be a problem when using more complex models, as having to estimate more parameters separately will increase the overall level of variance and uncertainty in the analysis. Thus, the general recommendation is to assess how well each model fits with the tested dataset using methods such as Akaike Information Criterion (AIC) (Akaike 1974) or the Bayesian Information Criterion (BIC) (Schwarz 1978) to find the one that contains as much complexity as needed for a reliable representation.

Another factor to take into account when estimating the rate of nucleotide substitution is how it varies between different positions within the sequence. It is common knowledge that evolutionary rate differs between each codon position, with the first base evolving the slowest and the third one the fastest. This is a result of different evolutionary pressures acting at each position, as the chance of modifying the amino acid coded is the highest in the first codon base. Therefore, depending on the dataset, a model that takes into account rate variation among sites may be required to infer accurate genetic distances.

In bacteria, a major confounding factor of the evolutionary signal is homologous recombination. By replacing the genomic fragment of one cell by a homologous sequence present in a more distant organism, significant variation may be introduced to a recipient bacterium. Horizontal gene transfer (HGT) can also involve the acquisition or loss of novel, non-homologous genetic fragments through mobile genetic elements (MGEs). These genomic regions cause more significant evolutionary changes by potentially bringing new functions to the population that may help it thrive in its environment (Bennett 2004, Toussaint and Merlin 2002). Transfer can occur through transformation, conjugation or transduction (Figure 8). Transformation involves the

uptake of exogenous DNA in which specific proteins required for this process are encoded within the chromosome of the bacteria. Conjugation occurs through direct contact between two bacteria and may either involve the transfer of conjugative plasmids or of chromosomally integrated conjugative elements (ICEs). They encode their own mobilization proteins, promoting DNA exchange between closely related bacteria (Burrus and Waldor 2004). During the majority of their lifecycle, ICEs remain integrated into the chromosome without expressing any of the conjugation genes. However, under particular circumstances, ICE gene expression may be induced, resulting in the excision of the element to form a circular dsDNA molecule, akin to a plasmid. This process requires the use of large protein complexes, such as the type IV secretion system (T4SS), to mediate the transfer of DNA from donor to recipient. Genes coded within this translocation system are responsible for assembling a mating pore that allows the transfer of the ICE DNA. An ICE-encoded relaxase then nicks and attaches itself to one strand of the ICE dsDNA, forming a complex known as the transfer DNA. The mating machinery subsequently promotes transfer of this complex to a recipient cell, where it recircularizes and integrates into the chromosome through a site-specific recombinase (integrase). The specificity of insertion can vary, but ICEs are often inserted in tRNA genes (Johnson and Grossman 2015). When conjugation involves the transfer of additional chromosomal regions, it is referred to as a high frequency of recombination (Hfr) mechanism. Lastly, another method of HGT is transduction, a process mediated by bacterial viruses, known as bacteriophages, which may incorporate part of the host DNA into a new cell (Frost et al 2005). Small fragments of recombining DNA called transposons may exist within larger mobile elements. These fragments either solely encode the transposase enabling its transposition (insertion sequences) or include other accessory genes that will be transferred together (Bennett 2004).

Since phylogenetic inferences are based on vertically acquired mutations during cell division, HGT can significantly mislead evolutionary studies of bacterial populations. Thus, care must be taken to remove both MGEs and regions that underwent homologous recombination from the sequence alignment before proceeding with further phylogenetic analyses (Croucher et al 2015).

To infer ancestral relationships between taxa from a phylogenetic tree, an outgroup has to be defined. This will determine the position of the root and the hypothetical most recent common ancestor (MRCA). If no additional information is available and the evolutionary rate is assumed to not vary significantly between branches, the root can be set at the midpoint of the tree, the halfway position between the two most divergent taxa. However, if longer branches are due to an increased evolutionary rate or to undetected recombination, this approach may misrepresent the phylogeny. Thus, an outgroup should preferably be selected to place the root between it and the ingroup taxa. The outgroup corresponds to a taxon that must belong to a distinct lineage in relation to the other samples under analysis, but should not be too divergent for its sequence to still correctly align with the rest and provide a discriminative signal.

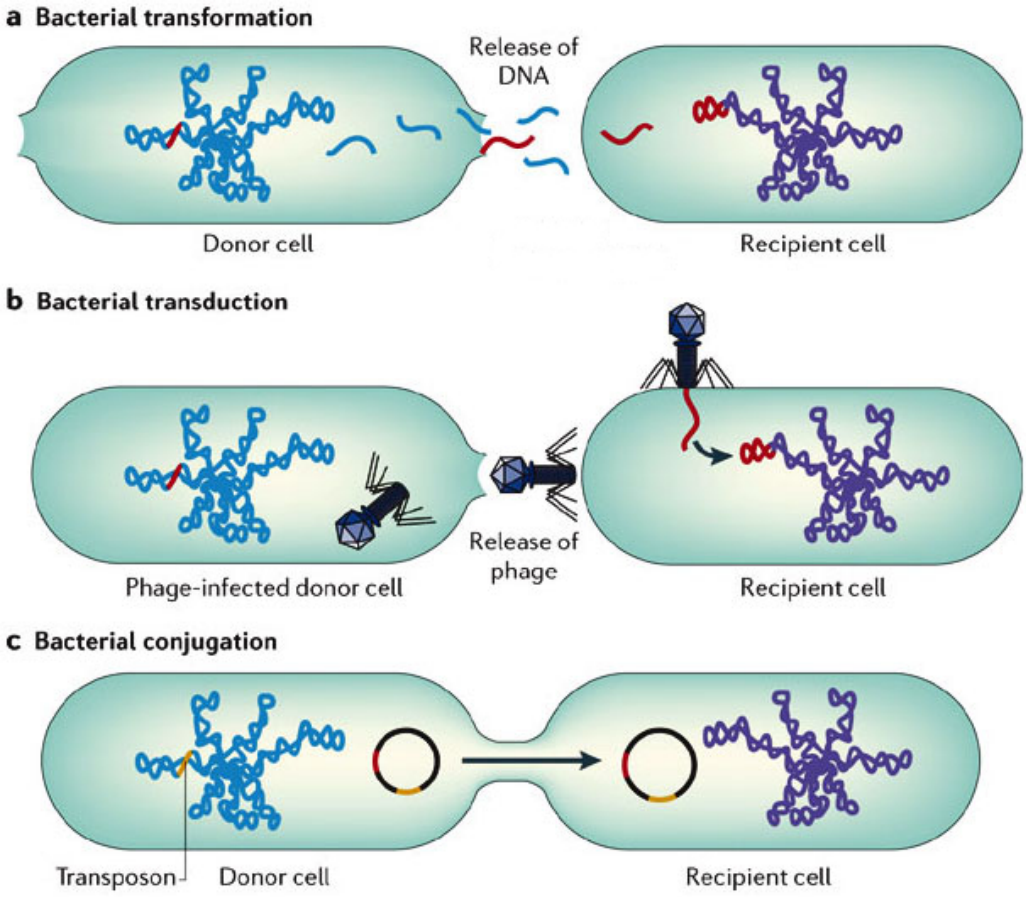


Figure 8. Mechanisms of horizontal gene transfer between bacteria (Furuya and Lowy 2006).

When constructing a phylogenetic tree there are several popular methods, such as maximum parsimony (MP), maximum likelihood (ML), Bayesian inference, or those based solely on distance methods. There is much, and inconclusive, debate on which is

accepted as the best, since each method presents different advantages and drawbacks (Kim 1996, Kolaczkowski and Thornton 2004). Regardless of the method used, a phylogeny is a prediction of the evolutionary history and relationship of a set of sequences, which may not correspond to the actual relatedness between each element in the dataset. To further assess the validity of the topology, a bootstrap analysis may be performed. This method is a way of estimating the reliability of each clade in the tree by resampling from the original dataset. Starting from a multi sequence alignment, a new one is constructed, called the bootstrap replicate, by randomly selecting positions from the original alignment. This new set of sequences has the same length as the original, so given that each position is picked at random, certain columns may be present more than once, or not at all. A tree is then inferred from each bootstrap replicate and the proportion of topologies supporting each clade is calculated. A support of at least 70% with 500 bootstrap replicates is usually recommended for giving significant confidence to phylogenetic nodes, and otherwise should be interpreted with caution (Zharkikh and Li 1992).

In MP, the principle is to find the tree that minimizes the amount of evolutionary change needed to explain the data, so the previously mentioned nucleotide substitution models are not taken into account. Although this is a fast and useful method in specific contexts, its simplistic approach carries some limitations. Thus, when computationally feasible, ML and Bayesian methods may be preferred for a more robust analysis.

2.1.2.1. Maximum likelihood and Bayesian methods

For ML predictions, the likelihood (L) is calculated from the probability estimated for different phylogenetic hypotheses. Each hypothesis represents the possible tree structures and the branch lengths τ, as well as the parameters of the substitution model selected θ. Hence, the likelihood can be represented as the probability of the data (sequences observed), given the hypothesis (tree structure and substitution model), as follows:

$$\mathrm{L}\,(\tau, \theta) \;=\; \Pr(\mathrm{Data}|\tau, \theta)$$

The final tree topology selected will be the one that maximizes the probability of observing the data, given the parameters, the so-called maximum likelihood estimate (MLE). However, given that the number of tree topologies increases enormously with the number of taxa involved, it is not feasible to compute the likelihood for each possible tree. Therefore, a heuristic approach is used to obtain reasonable starting trees based on distance methods. In this regard, the neighbour-joining (NJ) method is one of the most widely used. Using this approach, the aim is to find the topology that minimizes the total sum of the branch lengths of the tree, also referred to as the minimum evolution criterion. Then, to reach the MLE, the starting tree is rearranged and a new likelihood value is calculated. Once no better tree is found, this rearrangement stops and a final tree is selected (Lemey et al 2009).

Bayesian inference rose from the application of a theorem proposed by Thomas Bayes. In it, he applies the concept of conditional probability, stating that the probability of a certain event depends on our prior knowledge of factors that influence said event. In other words, the probability of a hypothesis is updated as new information becomes available. In Bayesian phylogenetics, the probability function of the parameters of the phylogeny θ, given the DNA sequences X is represented as follows:

$$f(\theta|X) = \frac{f(X|\theta)\, f(\theta)}{f(X)}$$

$f(\theta|X)$ is termed the posterior probability distribution, denoting the probability of all possible values of θ conditioned on the available sequence data. Given the tremendous amount of parameters that would need to be considered, the posterior probability distribution cannot be calculated analytically. Therefore, an approximation is estimated by Markov Chain Monte Carlo (MCMC) sampling using the Metropolis-Hasting method (Hastings 1970). Starting from an arbitrary point, this method carries out a random "walk" along the probability distribution, by making small changes to the parameters and rejecting or accepting those changes based on how they influence the posterior probability. The chain eventually converges towards an equilibrium state — a parameter space with a high posterior probability — and the final tree is obtained by choosing the one with the highest estimated value. Although this might resemble the

maximum likelihood approach, there are a few important differences. On one hand, the Bayesian approach may not find the best tree, as this depends on the number of samples recovered from the MCMC search. Moreover, for a ML approach, the confidence of the topology needs to be further assessed using a bootstrap analysis, whereas in the Bayesian case each tree and clade has an associated posterior probability. Ultimately, ML and Bayesian approaches have their own advantages. ML methods present a more consistent approach to parameter estimation, whereas Bayesian analyses are more appropriate when taking into account prior information and using more complex models that are computationally demanding.

2.1.2.2. Estimating the evolutionary rate

Early observations of the genetic distance between similar proteins suggested that variation increases linearly in relation to the divergence time (Zuckerkandl and Pauling 1962, Zuckerkandl and Pauling 1965). This gave rise to the notion of a molecular clock, a constant evolutionary rate at which each gene accumulates new mutations over time (Lemey et al 2009). Therefore, by knowing the evolutionary rate for a given gene or species, inferences on their divergence time can be made, and vice-versa. With the divergence rate of a bacterial population, we can then date past transmission events and estimate how long specific lineages have been colonizing their environment. However, empirical studies have shown that many genes evolve at a variable rate, which presented new challenges for molecular dating (Jenkins et al 2002).

Current phylogenetic methods, such as those in the "Bayesian Evolutionary Analysis Sampling Trees" (BEAST) program accommodate a number of clock models to either estimate one global and constant rate for the population (strict clock model) or take a more flexible approach (Drummond et al 2012) (Figure 9). The local molecular clock model, originally proposed by Yoder and Yang (Yoder and Yang 2000), allows the rate of evolution to be constant within certain clusters of the phylogenetic tree, while assuming a different rate for more distant branches. Along similar lines, the autocorrelated relaxed clock model (Thorne et al 1998) is centred on the assumption that the evolutionary rate is more similar among recently diverged taxa, so greater variation lies between the ancestral and terminal branches. Lastly, the uncorrelated relaxed clock

models assume a branch-specific rate with no correlation with the evolutionary history of the population (Drummond et al 2006).

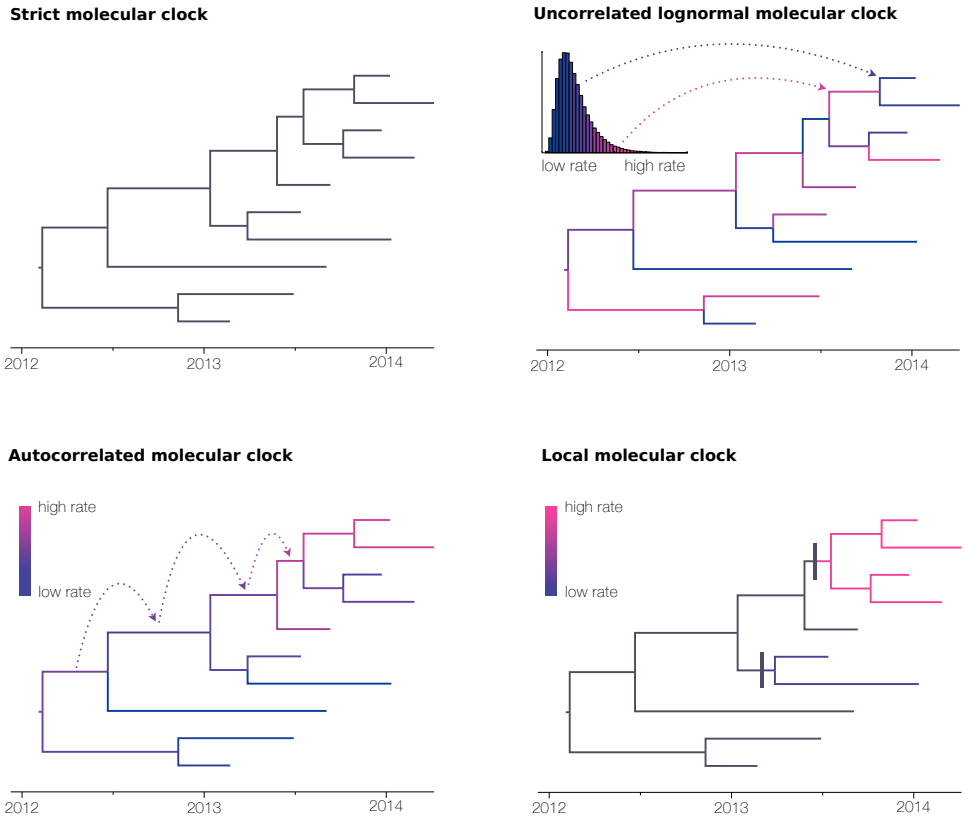


Figure 9. Main clock models used in Bayesian phylogenetic approaches (Drummond et al 2012).

2.1.3. Detecting natural selection

The presence of natural selection can be inferred in a number of ways. Since mutations are expected to arise randomly across the genome, observing recurrent substitutions at specific positions or genes throughout the population may suggest the presence of a similar evolutionary pressure acting on independent lineages. However, a higher than expected substitution rate can also be due to variations in mutation rate between genes, or undetected recombination. Thus, this global approach should be further complemented by analysis of the direction of natural selection. This is inferred by calculating the rate of accumulation of nonsynonymous changes per nonsynonymous site (dN) in relation to the rate of synonymous substitutions per synonymous site (dS). In principle, nonsynonymous mutations have a more significant impact on protein structure and function. Therefore, if the rate of substitutions with a functional effect on

the protein is higher (dN/dS > 1) than what would be expected under neutral evolution, this is indicative of positive selection. By contrast, a dN/dS < 1 suggests the presence of purifying selection and the conservation of protein function. The direction of natural selection, inferred with dN/dS, can be measured at individual sites, across particular set of genes or over the entire genome. In practice, there are several methods to obtain the dN/dS ratio. One of the simplest models, the Nei-Gojobori (Nei and Gojobori 1986), involves assessing the expected number of nonsynonymous and synonymous codons within individual sites or genes and normalize the observed number by this expectation. In microbial genomics, given that sequence data is usually in the form of an alignment of variable positions, calculation of dN/dS can be inferred from modelling the expected nonsynonymous rate for each possible nucleotide substitution (Lieberman et al 2011, Lieberman et al 2014). Although the dN/dS ratio can be a useful metric to infer the presence of natural selection, it should be interpreted with caution. For instance, the small number of substitutions occurring per site may limit the statistical significance of dN/dS. Therefore, to increase its statistical power, dN/dS can be calculated per gene or groups of genes. However, by averaging out the dN/dS across multiple sites, the presence of few positively selected mutations within a gene may be masked and underestimated by an overabundance of negatively selected substitutions (Hedge and Wilson 2016).

As previously mentioned, substitutions within non-coding or intergenic regions may also have an important biological significance. To be able to infer the impact of genomic variants within non-coding segments, there are several tools currently available, such as snpEff (Cingolani et al 2012), that predict the most likely effect of mutations within a given reference genome. However, this requires well-curated and annotated databases containing detailed information on promoter regions, operators, transcription start sites and terminators.

2.2. Within-host diversity and evolution

The perspective on how fast bacterial pathogens evolve in response to external environmental pressures has changed throughout history. Before, evolutionary studies of bacterial populations diverging for millions of years estimated that the rate of

evolutionary change was between 10^{-10} to 10^{-9} substitutions per site per year (Ochman and Wilson 1987, Wilson et al 1987), which would suggest that a population recently colonizing a single host would likely be completely homogeneous. But, until recently, the level of within-host variation of bacterial pathogens was unknown. Indeed, as traditional genotyping tools focused on the diversity of a limited number of traits within a population, epidemiological methods such as MLST (Maiden et al 1998) and PFGE (Schwartz and Cantor 1984) were only able to perform a broad differentiation of bacterial lineages with limited resolution. It is also increasingly clear that to make reliable inferences on pathogen evolution, diversity and transmission it is important to sample a sufficient portion of the within-host population instead of focusing on single isolates.

The development and increasingly widespread implementation of whole-genome sequencing approaches has spawned many reports further detailing the diversity of a population colonizing a single environment. Combined with the application of molecular clock models, it has also brought significantly new understandings of the evolution of bacterial pathogens, showing that within short time periods, bacteria are able to evolve much faster than previously thought (Ho et al 2007). Indeed, by analysing the number of mutations across samples collected either simultaneously or during a short amount of time, the rate of accumulation of new substitutions can be estimated much more accurately. For instance, a retrospective analysis of longitudinal samples of *Burkholderia dolosa* collected throughout 16 years estimated the evolutionary rate to be around 2 SNPs per year per genome (Lieberman et al 2011). Investigation of the substitution rate for other human pathogens showed that it also varies significantly between species. In one extreme is the case of *Helicobacter pylori*, a well-known stomach pathogen found to acquire approximately 30 mutations per year per genome (Falush et al 2001). For other species, such as *E. coli, Clostridium difficile*, GBS, *K. pneumonia* and *S. aureus,* the average rates of evolutionary change range from 1 to 10 mutations per genome per year (Da Cunha et al 2014, Didelot et al 2012a, Mathers et al 2015, Reeves et al 2011, Young et al 2012). On the other end are the cases of *Mycobacterium tuberculosis* and *Legionella pneumophila*, with an average of only 0.6-0.8 mutations per genome year (David et al 2016, Ford et al 2011).

Differences in evolutionary rates are determined not only by differences in selective pressures and the number of generations, but also on the efficiency of DNA repair mechanisms (Bromham and Penny 2003). Indeed, when a mismatch repair system becomes disrupted, it can lead to a much higher rate of point mutations, a phenomenon known as hypermutation (Taddei et al 1997). Along the same lines, specific loci mutate at an above-average rate, producing a pool of genetic variability and phenotypic heterogeneity among a mostly homogeneous population. This rapid adaptability, known as phase variation (Henderson et al 1999), has been implicated as a potential virulence strategy (Alamro et al 2014, Moxon et al 2006). For instance, genomic variation due to replication slippage may swiftly increase or decrease the number of tandem repeating domains within certain immunogenic surface proteins (Figure 10). Then, resulting changes to gene expression may facilitate the selection of variants that can more easily evade the host immune response (Alamro et al 2014) (Figure 10).

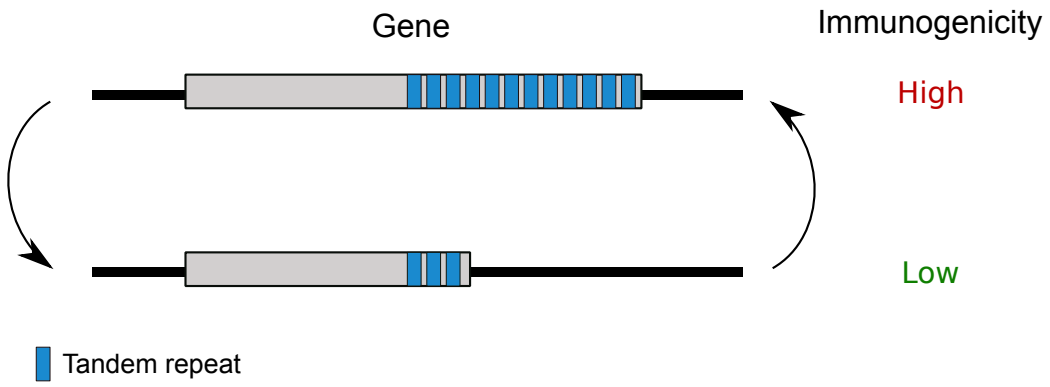


Figure 10. Example of phase variation. The number of repeated domains in particular surface proteins can affect their immunogenicity. Therefore, depending on the immune selective pressures, an increase or decrease in the number of tandem repeats due to replication slippage is rapidly selected.

2.2.1. Transition from carriage to infection in opportunistic pathogens

Most bacterial pathogens are also commensals, so an important consideration when studying bacterial adaptation is to investigate whether the progression of disease is related with the acquisition of pathoadaptive genes and mutations during colonization, in order to find potential triggers of infection. In reality, causes of disease are multifactorial and are dependent on both host- and pathogen-related properties and of their interaction. Nonetheless, studies have looked at the impact of specific mutations in

many clinically important species. In *S. aureus*, a particular study tracked the genomic changes accumulated throughout 13 months in an individual that developed a severe case of septicaemia (Young et al 2012). Genes affected by the mutations accompanying this shift from commensal to pathogen included a major AraC-type transcriptional regulator — a family of genes shown to regulate carbon metabolism, stress responses and virulence. Analysis of 474 clinical isolates of *Pseudomonas aeruginosa* colonizing individuals with cystic fibrosis for an average of five years, identified convergent adaptation in 52 genes involved in regulation, metabolism and virulence (Marvig et al 2014).

Genomic adaptation during colonization does not exclusively relate to progression of disease. Indeed, a recent study showed that a strain of *Burkholderia pseudomallei* underwent several deletions and loss of function mutations — including those affecting an essential virulence factor — allowing it to persistently colonize a single individual (Price et al 2013). This exemplifies that the evolutionary success of a pathogen within its host might depend on the balance between the potential for virulence and for a stable long-term colonization.

The ongoing development of whole-genome techniques will continue to provide a greater insight into how pathogens adapt to the human host, ultimately contributing to more informed therapeutic decisions. One major difficulty will be linking the genomic variation highlighted from evolutionary studies to actual pathogenic phenotypes. Even though assessing the impact of genomic mutations on protein function and structure might be feasible to test, predicting the specific environmental conditions where these mutations could have been positively selected is a more challenging endeavour.

2.2.2. Emergence of antibiotic resistance

One of the most obvious examples of the adaptive potential of bacterial pathogens is the emergence of antibiotic resistance as a result of the overuse and misuse of antibiotics. Under these strong selective pressures, a new clone arising with a point mutation or gene that provides resistance to the antibiotic used sweeps through the population. Therefore, being able to study the evolution of antibiotic resistance using whole-genome

approaches has provided crucial insights into new resistance mechanisms. For clinical laboratories, this has also helped predict treatment failure and develop new drugs and therapeutic strategies (Koeser et al 2014). Certain genes have been found to be frequent targets of adaptive evolution as a result of antibiotic treatment. Such examples are *rpoB*, a target of selection in response to treatment with rifampicin (Gao et al 2010); *gyrA* and *parC*, in which mutations are selected after administering fluoroquinolones (Kawamura et al 2003, Lieberman et al 2014); and *vraRS*, whose variants are associated with decreased vancomycin susceptibility (Mwangi et al 2007). Mutations in these genes were shown to be selected independently in different bacteria that are exposed to the same drug, representing a process of convergent evolution towards resistance. As an example of a more controlled experiment, a microbial cultivation system maintaining a constant antibiotic pressure was used to monitor the evolution of resistance in *E. coli* over 20 days (Toprak et al 2012).

Nonetheless, evolutionary changes conferring antibiotic resistance also possess some inherent costs, explaining why certain resistance mechanisms have not been able to efficiently spread. Mutations involved in antibiotic resistance acquired within key enzymes may affect the overall efficiency of replication and transcription, while specific proteins, whose function is mostly limited to providing antibiotic resistance, may be especially costly to produce (Comas et al 2011, Melnyk et al 2015). Therefore, without the antibiotic pressure, resistant strains might eventually disappear from the population. On the other hand, additional mutations termed “compensatory” might be selected to counteract the adverse effects of carrying the resistance traits, allowing the resistant bacteria to persist (Levin et al 2000).

2.2.3. Transmission and recurrent infections

One of the main applications of whole-genome sequencing has been in the study of pathogen transmission networks during an epidemic. Using high-resolution techniques, it may be possible to trace the direction of transmission from the mutation directionality inferred with conventional phylogenetic approaches. However, the existing bacterial diversity present within each host must be taken into account. Within short timeframes, transmission of bacterial pathogens with a low evolutionary rate may be difficult to

track. At the same time, in cases of long-term carriage and high rates of evolution, substantial variation within each individual may impede a reliable reconstruction of transmission events between multiple hosts sharing a closely related population (Worby et al 2014). Therefore, instead of sequencing a single isolate within the population, recent studies are now either focusing on independently collecting several isolates from each individual or sample (Tong et al 2015), or on performing direct sequencing of a bacterial pool consisting of hundreds of colonies from a single plate (Lieberman et al 2014).

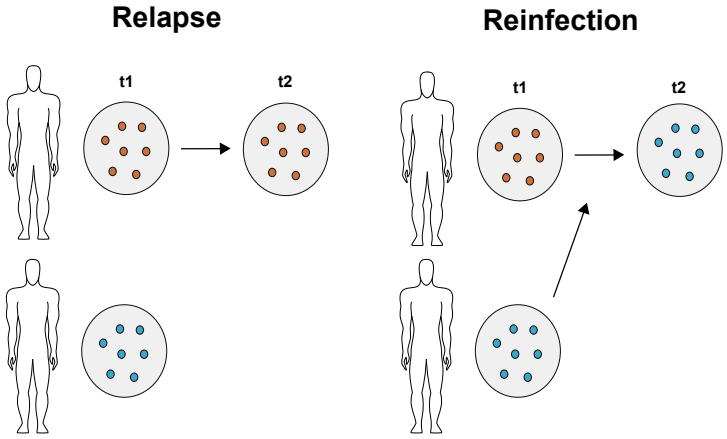


Figure 11. Distinction between relapse and reinfection. In the case of relapse, the same strain detected in the first time point subsequently re-emerges, whereas when reinfection occurs the most recent strain is newly transmitted from another source.

One important application of the reconstruction of pathogen transmission networks is to distinguish reinfection from relapse within individuals who experience multiple episodes of disease (Figure 11). This distinction is usually based on the number of genomic differences observed from consecutive isolates collected between each diagnosis, and by sampling other potential transmission sources. When recurrent infections result from identical or almost identical strains with no closely related isolates detected in additional host or environmental sources, then relapse is the most likely explanation (Okoro et al 2012). This is of extreme importance in selecting the right approach to treat recurrent cases of infections (Eyre et al 2014). Although specific guidelines have been set to help distinguish isolate relationships (van Belkum et al 2007), the highly discriminative resolution of whole-genome data poses additional challenges in defining the right threshold for epidemiologic relatedness to decide therapeutic and control measures during an outbreak.

2.3. Host adaptation

When colonizing a new environment, bacteria encounter multiple challenges forcing adaptive strategies to arise for it to survive and be able to continue to spread. Among these many obstacles are physical barriers within the host, competition with the existing microbial communities, the host immune response and the nutritional conditions in the environment (Didelot et al 2016). The relative contribution of each of these factors in shaping the adaptive response of a bacterial pathogen naturally varies between each individual reservoir.

Sequencing data has now started to unravel the main forces of natural selection present within several host ecosystems. When switching to a different environment, bacteria undergo a period of rapid adaptation that slows down as the population improves in fitness. This has been shown both *in vitro* (Barrick et al 2009) and in evolutionary studies investigating parallel bacterial adaptation in clinical patients (Lieberman et al 2011).

Studies have predicted that mutations among major transcriptional regulators recurrently contribute to host tropism (Yang et al 2011). Nevertheless, experimentally validating these hypotheses is particularly difficult, as recreating the host environment for functional studies, if possible, is especially challenging.

Adaptation is not exclusively dependent on environmental pressures exerted by the host. Neighbouring bacteria, belonging to the same ecosystem, play a dramatic role in driving adaptation (Xavier 2011). HGT is ubiquitous among cohabiting species, helping shape their genomic structure through the exchange of large segments of genetic material. Homologous recombination, for instance, has been shown to be an important process for genomic evolution and adaptation in species such as *S. agalactiae* (Brochet et al 2008a), *E. coli* (Didelot et al 2012b) and *Streptococcus pneumoniae* (Croucher et al 2011), as well as in the *Campylobacter* genus (Sheppard et al 2013a). Beyond representing a genetic pool of mobile, exchangeable genes, microbial communities may also compete for nutritional resources, encouraging bacteria to frequently exploit complementary phenotypes within the population. An example of this complex social

interaction are so-called “selfish cheaters”, microorganisms that reap the benefits of public goods, without the added cost (Andersen et al 2015).

Studies of bacterial host adaptation have provided important insights into the evolutionary history of many host generalist species. For instance, *S. aureus* is a clinically important pathogen known to infect a wide-range of hosts. Genomic studies have inferred that human-to-animal host jumps led to the emergence of livestock-specific strains. However, recent host switches from animals back to humans have shown that bovine-associated clones also have the potential to cause disease in humans (Udo et al 2011). Remarkably, a single SNP within the *dltB* gene was discovered to cause a human-adapted *S. aureus* strain to be able to infect rabbits (Viana et al 2015). Likewise, the flux of MGEs encoding virulence-associated proteins was revealed to underlie *S. aureus* ruminant host tropism (Guinane et al 2010).

Salmonella enterica is an example of a species with both generalist and host-specific lineages. The strain *Salmonella enterica* serovar Enteriditis is able to colonize various hosts, such as chickens, cattle, mice and humans, whereas *Salmonella enterica* serovar Typhi and *Salmonella enterica* serovar Gallinarum are restricted to humans and avian species, respectively. Comparative genomics showed that a host-restricted lifestyle of *S*. Typhi and *S*. Gallinarum led to extensive genomic decay and pseudogene formation, affecting functional traits such as cell motility and metabolism (Thomson et al 2008).

Campylobacter jejuni, a common cause of gastroenteritis, presents frequent host switching among bird and mammal species. In this regard, genome-wide association studies (GWAS) showed that vitamin B5 biosynthesis is an important colonization factor in cattle (Sheppard et al 2013b). Moreover, phase variation of poly C/G tracts among hundreds of genes has also been associated with the rapid adaptation of this species (Bayliss et al 2012).

Genomic analysis also investigated how the host-restricted *Streptococcus equi* subspecies *equi* (*S. equi*) evolved from the zoonotic strain *Streptococcus equi* subspecies *zooepidemicus* (*S. zooepidemicus*) (Holden et al 2009). It was evidenced that host specialization of *S. equi* was driven by functional loss due to nonsense mutations and

deletions, combined with the acquisition of MGEs carrying genes encoding a phospholipase A_2 toxin, four superantigens and an iron acquisition system (Holden et al 2009).

Further characterization of inter- and intra-host evolutionary dynamics and transmission will facilitate the control of both human and animal pathogens. In this respect, *S. agalactiae* is a prime example of an opportunistic and host generalist species that is yet to be fully understood.

3. Genomic adaptation of group B *Streptococcus*

3.1. Genome structure

The first genomes of GBS to be sequenced were of strains NEM316 and 2603V/R (Glaser et al 2002, Tettelin et al 2002), and were followed by the genomic analysis of six additional human-derived isolates (Tettelin et al 2005). There are now 36 completely assembled genomes of GBS publicly available (Table 2), with a genome size ranging from 1.8 to 2.3 Mb, a GC content between 35-36% and an average of 1945 CDS. In addition, more than 1000 draft genomes of GBS have been analysed from various studies published in the last 15 years (Table 3).

In 2011, the first bovine-isolated strain was sequenced and compared to the human-specific genomes previously published (Richards et al 2011). Then, following an in-depth evolutionary analysis of six strains isolated from fish (Rosinski-Chupin et al 2013), the genomic data obtained from 229 isolates was used to investigate how human-specific clones emerged in the 1960s (Da Cunha et al 2014). Later, the specific evolution of the CC1 lineage was studied in more detail from a panel of 185 strains (Flores et al 2015), while the molecular basis for capsule loss was investigated in 128 GBS isolates of various origins (Rosini et al 2015). More recently, the distinction between disease-associated lineages of ST17 among neonates and adults was assessed in a GBS population from Canada (Teatero et al 2016). Lastly, a large panel of 915 GBS strains isolated in Kenya was sequenced and analysed to relate GBS diversity with maternal colonization and neonatal disease (Seale et al 2016). Each of these studies has brought important insights into the evolutionary dynamics of GBS, highlighting the importance of mobile genetic elements in shaping the evolution of this species, and providing clues of the host-specific properties contributing to adaptation and pathogenesis.

Table 2. Properties of the 36 complete GBS genomes publicly available.

Strain	Host	Size (bp)	Proteins	MLST	Year published
09mas018883	Bovine	2138690	2030	ST1	2013
GBS ST-1	Dog	2165970	2055	ST1	2015
SS1	Human	2092070	1977	ST1	2015
A909	Human	2127840	2026	ST7	2005
GD201008-001	Fish	2063110	1940	ST7	2012
GX064	Fish	2064350	1941	ST7	2015
HN016	Fish	2064720	1944	ST7	2015
WC1535	Fish	2212570	2066	ST7	2016
YM001	Fish	2047960	1930	ST7	2015
BM110	Human	2170276	2167	ST17	2016
COH1	Human	2065070	1929	ST17	2014
NGBS128	Human	2079120	1914	ST17	2016
H002	Human	2147420	2008	ST19	2015
GBS1-NY	Human	2243710	2083	ST22	2014
GBS2-NM	Human	2214300	2057	ST22	2014
GBS6	Human	2231480	2079	ST22	2014
NEM316	Unknown	2211490	2103	ST23	2002
CNCTC 10/84	Human	2013840	1903	ST26	2014
SA111	Bovine	2275140	2095	ST61	2016
GBS85147	Human	1996150	1876	ST103	2015
2603V/R	Human	2160270	2127	ST110	2002
138P	Fish	1838700	1593	ST261	2014
138spar	Fish	1838130	1590	ST261	2014
2-22	Fish	1838870	1548	ST261	2013
GX026	Fish	1840650	1613	ST261	2015
CU_GBS_08	Human	2084510	1999	ST283	2016
CU_GBS_98	Human	2029670	1935	ST283	2016
SG-M1	Human	2116810	2037	ST283	2015
NGBS357	Human	2172870	2045	ST297	2016
NGBS572	Human	2061430	1937	ST452	2014
NGBS061	Human	2221210	2109	ST459	2014
S25	Fish	1838990	1640	ST552	2016
SA20-06	Fish	1820890	1646	ST552	2012
ILRI005	Camel	2109760	2004	ST609	2013
ILRI112	Camel	2029200	2073	ST617	2013
FWL1402	Frog	2090290	1994	ST739	2016

Table 3. Main genomic studies of GBS published so far.

Authors	Year	Genomes[1]	Host[2]	MLST[3]
Glaser et al.	2002	1	Unknown	ST23
Tettelin, et al.	2002	1	Human	ST110
Tettelin, et al.	2005	6	Human	Various
Richards, et al.	2011	1	Bovine	ST67
Rosinski-Chupin, et al.	2013	6	Fish	ST7/260/261
Da Cunha, et al.	2014	229	Human	Various
Flores, et al.	2015	185	Human	ST1
Rosini, et al.	2015	128	Human	Various
Teatero, et al.	2016	93	Human	ST17
Seale, et al.	2016	915	Human	Various

[1]Number of genomes sequenced and analysed.
[2]Host origin of the majority of isolates studied.
[3]MLST sequence type (ST) of the isolates studied.

3.2. The role of mobile genetic elements

Genomic analysis of the first eight sequenced genomes of GBS identified a set of 69 genomic islands absent in at least one of the genomes (Tettelin et al 2005). This study showed that GBS possesses a set of core genes conserved across the species equivalent to approximately 80% of each individual genome (Tettelin et al 2005). In the case of NEM316, genomic analysis showed that 55% of the genes predicted in this strain had an orthologous variant in *S. pyogenes*. Among the remaining regions, half were described as being clustered within 14 genomic islands containing known and putative virulence genes (Glaser et al 2002). GBS is described as having an "open" pan-genome, seeing that the number of accessory genes increases with each new sequenced strain (Tettelin et al 2005). Functional classification of both the core and accessory genes showed that the dispensable or accessory part was related to hypothetical proteins with unknown function, while the set of conserved genes was mostly associated with housekeeping functions, the cell envelope, gene regulation and protein transport (Tettelin et al 2005). Closer inspection of the mobilome revealed that two thirds of the genomic islands corresponded to ICEs or related elements with a high plasticity and wide distribution across the species (Brochet et al 2008a). Of the ICE detected in GBS, a novel family was identified, termed *TnGBSs*, whose mobility is mediated by a transposase with a DDE

motif, instead of a traditional site-specific phage-like integrase (Brochet et al 2009, Guerillot et al 2013). Importantly, *TnGBS* promotes the mobilization of chromosomal DNA through an Hfr-type mechanism.

Globally, HGT has been shown to play a significant role in the diversification and adaptive potential of GBS. In particular, conjugative transfer of large DNA segments was shown to have driven the emergence of major GBS clones (Brochet et al 2008b), affecting important regions such as the capsule operon. Recently, it was shown that a subset of serotype IV strains— a CPS type that has been increasing in frequency among carriers and infected individuals — originated within the CC17 lineage through recombination of the entire *cps* operon (Bellais et al 2012). Recombination and HGT was also suggested as a means to generate diversity among the immunogenic alpha-like family of proteins (Broker and Spellerberg 2004, Da Cunha et al 2014, Lachenauer et al 2000). Moreover, the *scpB* gene, encoding the surface protein C5a peptidase, is located alongside *lmb*, coding for the laminin-binding protein, within a composite transposon that was possibly exchanged with *S. pyogenes* (Chmouryguina et al 1996, Franken et al 2001). HGT also contributed to the current distribution of the pilus islands PI-1 and PI-2 among the GBS population (Springman et al 2014). In addition, comparative genomics revealed that HGT between GBS and other streptococci sharing the same environment involved the frequent exchange of resistance and metabolism traits, such as the lactose operon (Richards et al 2011). It was more recently shown that the acquisition of the ICEs *Tn916* and *Tn5801*, harbouring the tetracycline resistance gene *tetM*, was the main reason for the selection of disease-causing GBS clones in the mid-20th century (Da Cunha et al 2014). In the same study, a *Tn916* variant containing the *ermB* gene for macrolide resistance was also associated with the expansion of a specific lineage (Da Cunha et al 2014). The *Tn916* family of transposons has long been studied in *Enterococcus faecalis* (Franke and Clewell 1981) and is found among a wide range of bacterial species (Roberts and Mullany 2009).

Regulating the acquisition of mobile DNA elements is the CRISPR system. Its locus encodes a set of unique "spacer" sequences, together with CRISPR-associated (CAS) proteins that altogether provide immunity against MGEs (Deveau et al 2010, Horvath and Barrangou 2010) (Figure 12). Once bacteria are first invaded by an external

element, such as a bacteriophage, part of its DNA is cleaved and integrated into the CRISPR locus as a spacer of 26-72 nucleotides (Kupczok and Bollback 2013). The CRISPR RNAs (crRNAs) will then be able to target newly invading foreign DNAs for degradation, if similar enough to the sequences incorporated into the spacer array (Bhaya et al 2011). Two CRISPR systems have been identified in GBS (CRISPR1 and CRISPR2), but only CRISPR1 was shown to be active and suggested to contribute to the diversity of the mobilome in the GBS population (Lopez-Sanchez et al 2012).

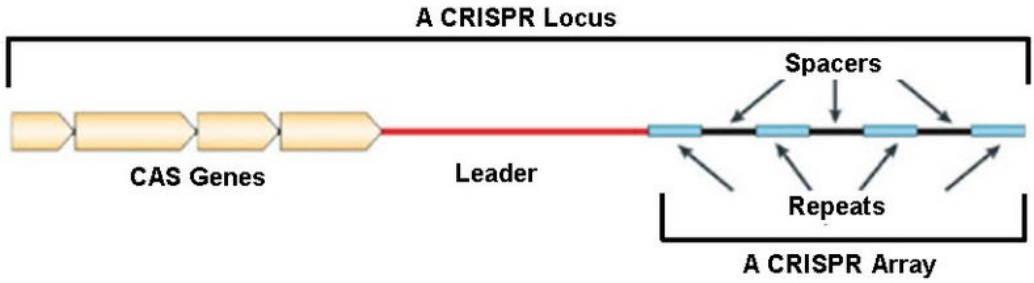


Figure 12. Composition and structure of the CRISPR locus (Sorek et al 2008).

3.3. Current insights into host adaptation and pathogenesis

Knowing what contributes to the virulence and colonization potential of GBS, the question that arises is whether there are particular functional traits of host-specific populations that might help explain their affinity towards a more restricted environment. Although large chromosomal exchanges underlied the main evolutionary events of this species, phylogenetic analysis of the human-specific CC17 showed that it represents a highly conserved group, reflecting a recent divergence unaffected by recombination (Da Cunha et al 2014, Sorensen et al 2010). Various studies have tried to decipher the hypervirulent properties of CC17. In particular, a distinctive set of surface proteins and adhesion factors appear to distinguish this lineage from other GBS clones. For instance, the hypervirulent GBS adhesin HvgA is unique to CC17 strains, since other GBS isolates possess different variants of an equivalent gene known as *bibA* (Tazi et al 2010). Of the serine-rich repeat proteins, Srr2, together with a homolog of its secretory locus *secA2*, are exclusively detected in CC17-specific strains (Brochet et al 2006, Seifert et al 2006). Moreover, the most recently identified gene belonging to the *fbs* family, *fbsC*, was shown to be important for the adhesion of most clinical GBS isolates, but not of those belonging to the CC17 complex, as the protein is not expressed (Buscetta et al

2014). Similarly, evolutionary studies of serotype V strains from CC1 showed that positive selection of small genetic changes, rather than extensive recombination, contributed the most to their successful dissemination (Flores et al 2015). Of the positively selected mutations, most were located in the capsule operon and in regions coding for the pilus structures and two-component systems (Flores et al 2015).

PCR-based methods were able to identify genomic regions unique to human- or bovine-specific populations of GBS, suggesting the presence of distinct selective pressures driving acquisition and loss of function through horizontal exchange (Richards et al 2011). In particular, the ability to metabolize lactose through the action of a functional lactose operon (Lac.2) was shown to be a common trait exclusively among GBS colonizing the bovine mammary gland (Richards et al 2013). Additionally, transcriptome analysis of a bovine-adapted strain grown in bovine milk revealed an up-regulation of functions related to copper homeostasis, metabolism of pyrimidine, purine, glycerol and glucose, and of putative aminoglycoside antibiotic resistant genes. Another distinguishing trait between the human and bovine population is that the genomic island harbouring *scpB* and *lmb* is frequently absent from bovine GBS isolates (Sorensen et al 2010). Moreover, it was shown that expression of the C5a peptidase gene *scpB* is induced in human, but not bovine, serum (Gleich-Theurer et al 2009), further suggesting that this gene is most likely differentially regulated in different host environments. Another genomic island distributed differently between human- and bovine-specific populations harbours an ABC transporter coupled with a two-component regulatory system. This region, termed *vexp*-*vncRS*, is present exclusively in CCs of human origin (Brochet et al 2008b). Other similar systems were identified in *E. faecalis* (Paulsen et al 2003) and in *C. difficile* (Sebaihia et al 2006). These two species are also known to colonize the human intestinal tract, so the acquisition of the *vexp*-*vncRS* region could be a contributing factor for adaptation to this host reservoir.

Antibiotic resistance determinants also differ between the GBS populations in humans and cattle. One major breakthrough in the study of the evolutionary history of GBS was the observation that most human-associated lineages are tetracycline resistant due to the acquisition of the *tetM* gene (Da Cunha et al 2014, Poyart et al 2003). In bovines, there is a variable but generally low prevalence of resistance to tetracycline, and is most

frequently conferred by the *tetO* gene, as opposed to *tetM*. The *lsa* genes, conferring cross-resistance to lincosamides, streptogramin A and pleuromutilins were also recently shown to present a host association. Although only 19 strains were identified as carrying *lsa* genes, *lsaC*, *lsaC* variant 1 and *lsaE* were detected only among strains of human origin, while *lsaC* variant 2 was restricted to GBS strains responsible for bovine mastitis (Douarre et al 2015).

The relation between the pilus composition in GBS and its host-associated populations has also been investigated. In particular, it was shown that human STs were more likely to carry a combination of PI-1 and PI-2, whereas bovine-specific isolates normally lack the PI-1 locus altogether, suggesting that the pilus contributes to host-specific properties (Springman et al 2014). For instance, the human-associated PI-1 was shown to primarily serve to protect against intracellular killing by human-derived macrophages (Maisey et al 2008). Moreover, PI-1 was seen in combination with PI-2a most frequently among CC19, and with PI-2b in CC17 strains, suggesting that PI-2's contribution to human colonization might be more significant in this particular genomic background (Springman et al 2014). However, comparative genomics of a CC17 population from Canada detected the replacement of PI-1 by an MGE carrying resistance genes against tetracycline, macrolides and other antibiotics (Teatero et al 2016). In the end, these reports suggest that the role and contribution of GBS pili to host adaptation is multifaceted, as it depends on pilus type, genomic background and the environmental conditions.

The association between the alpha-like family of proteins with host and CC is more tenuous, owing to frequent recombination events (Broker and Spellerberg 2004, Lachenauer et al 2000). Nonetheless, in the case of the hypervirulent CC17 strains, Rib is the only alpha-like variant that has been detected (Brochet et al 2006). Moreover, one notable observation is that even though almost all GBS strains carry at least one gene encoding an alpha protein, none are present in the fish-adapted ST260/261. In fact, the ST260/261 fish-restricted group underwent a substantial reductive evolution, resulting in a chromosome 10-25% smaller than that of other GBS (Rosinski-Chupin et al 2013). Besides not having an Alp gene, ST260/261 is also devoid of the genomic island harbouring *scpB* and *lmb*, and of the *vexp-vncRS* region. Given its distinctive properties,

it was hypothesised that this lineage diverged earlier from other GBS clones, shedding some light into the evolutionary history of the species. The other CC from which fish-derived strains can be found (CC7) is able to colonize humans, cattle and fish and has not undergone extensive genetic decay, reflecting a more recent divergence, and adaptation to a host generalist lifestyle.

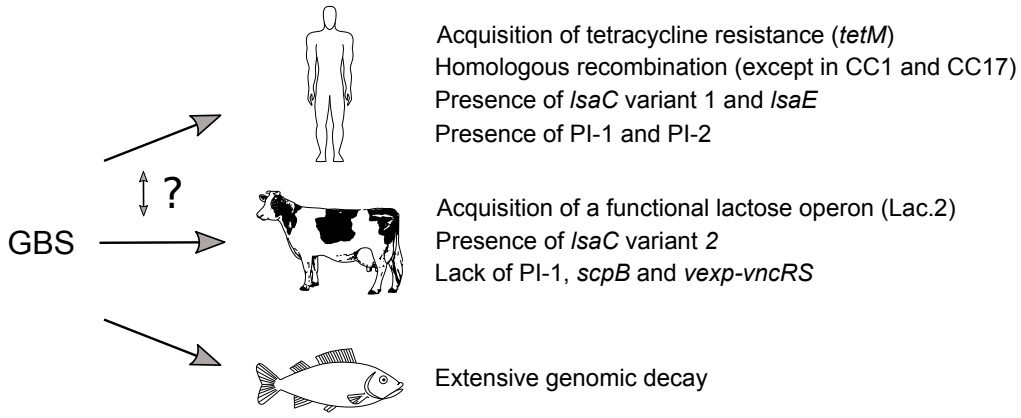


Figure 13. Main features of GBS host adaptation currently known.

Genomic studies presented thus far have contributed to a better understanding of the host specificities of GBS (Figure 13). However, these works have mostly focused on the analysis of human-specific isolates. Furthermore, the selective advantage of these host-specific features remains unclear, as they may reflect spurious associations confounded by population structuring and linkage disequilibrium. The fact that several CCs are found in both humans and bovines also raises important questions on the host evolutionary history of GBS. Likewise, its zoonotic potential and the risk of inter-host transmission are still a matter of debate (Figure 13). Contradictory reports have speculated that some human GBS clones have originated from a bovine reservoir (Bisharat et al 2004) or that bovine-specific strains have a human origin (Manning et al 2010). Furthermore, fish consumption has been considered a potential risk factor for colonization of specific serotypes of GBS in humans (Foxman et al 2007). The initial hypothesis that the CC17 hypervirulent clone could have derived from a bovine ancestor was later discarded and accredited to the composite genome structure of GBS, the result of frequent recombination (Brochet et al 2006, Sorensen et al 2010). However, additional evidence inferred from whole-genome data is crucial to assess the potential of GBS host-switching and cross-species infection. Additionally, as new methods and

more genomes become available, larger and more thorough analyses will further elucidate the complex history and specific properties of this host generalist species.

Main objectives

The main goal of my PhD project was to decipher the evolutionary history of GBS and provide new valuable insights into its host-specific lifestyle.

The first part of the work consisted in the detailed analysis of a bovine-specific population of GBS that has long persisted as a cause of bovine mastitis in Portugal. We were able to infer the main adaptive changes and functional traits determining the evolutionary success of this population in the bovine environment. At the same time, it allowed me to setup a methodological approach to analyse the evolution of large bacterial populations using whole-genome data.

In the second part I have investigated the human-adapted population of GBS at two evolutionary scales. One of the goals was to better understand the short-term evolution of GBS in the human host by identifying the genomic mutations positively selected in GBS strains transmitted from mother to child during the perinatal period. We hypothesised that given a certain level of within-host diversity of GBS in each mother, particular GBS variants with a selective advantage could take over the newborn population after being vertically acquired during or just after birth. In parallel, a comprehensive genomic analysis was performed on 612 isolates belonging to the hypervirulent and human-specific CC17 lineage. The objective was to further develop and apply the method established at the start of the project to analyse the main evolutionary events that have shaped the CC17 population, understand how it disseminated worldwide and find the most important genetic factors behind its hypervirulent nature.

Results

A. Parallel evolution and genomic signatures of adaptation in the bovine host

At the end of the 20th century, several control strategies were implemented to reduce the incidence of GBS mastitis throughout Europe. However, in Portuguese dairy farms GBS continues to be frequently associated with bovine intramammary infections, causing a significant financial burden. For the first part of the project I collaborated with research and veterinary laboratories in Portugal, where I previously developed my Master's thesis, to closely study this particular GBS population. An initial genotyping analysis revealed that a single CC61 clone was the sole cause of GBS mastitis in the north of Portugal. Having uncovered this unexpected scenario in the northern region, we decided to extend our analysis to the rest of the country where we observed an ongoing trend towards the replacement of the local GBS population by this same clone. The lack of population genomics studies aiming to understand the adaptation of GBS to the bovine environment prompted us to take advantage of this unique situation and closely follow the *in vivo* evolution of GBS in the bovine host. In parallel, we performed the same evolutionary analysis on a bovine-isolated CC2 population as a means to compare the adaptive changes occurring within two distinct genomic backgrounds.

Our SNP-based genomic analysis showed that the most evident signal of parallel evolution was within an iron/manganese transporter operon, in which all CC61 isolates acquired at least one mutation. From the genetic specialization and decay of GBS in the bovine ecosystem we were able to gain new insights into the accessory functions of GBS isolated from cattle, as well as into the risk of inter-host transmission between the bovine- and human-derived populations. This study illustrates the value of genomics in pinpointing important mechanisms of bacterial adaptation and how they might be functionally addressed. It also allowed me to develop a bioinformatics workflow that can be further applied to accurately study the evolution of other microbial populations from large volumes of sequencing data. In addition to the work here presented, preliminary assessment of the functional heterogeneity of this population, namely regarding growth

in bovine milk and biofilm formation, revealed promising data (not shown) that may be explored in the future to link the genotypic diversity detected to specific phenotypic outcomes.

Publication nº1

environmental microbiology

Environmental Microbiology (2016) **18**(11), 4216–4229 doi:10.1111/1462-2920.13550

Persistence of a dominant bovine lineage of group B *Streptococcus* reveals genomic signatures of host adaptation

Alexandre Almeida,[1,2,3] Cinthia Alves-Barroco,[4] Elisabeth Sauvage,[1,2] Ricardo Bexiga,[5] Pedro Albuquerque,[6] Fernando Tavares,[6,7] Ilda Santos-Sanches[4] and Philippe Glaser[1,2]*

[1]*Institut Pasteur, Unité Evolution et Ecologie de la Résistance aux Antibiotiques, Paris, France.*

[2]*CNRS UMR 3525, Paris, France.*

[3]*Université Pierre et Marie Curie, Paris, France.*

[4]*Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, UCIBIO – Unidade de Ciências Biomoleculares Aplicadas, Universidade NOVA de Lisboa, Lisboa, Portugal.*

[5]*Faculdade de Medicina Veterinária, Centro de Investigação Interdisciplinar em Sanidade Animal, Universidade de Lisboa, Lisboa, Portugal.*

[6]*CIBIO – Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório Associado, Universidade do Porto, Campus Agrário de Vairão, Vairão, Portugal.*

[7]*Faculdade de Ciências, Departamento de Biologia, Universidade do Porto, Porto, Portugal.*

Summary

Group B *Streptococcus* (GBS) is a host-generalist species, most notably causing disease in humans and cattle. However, the differential adaptation of GBS to its two main hosts, and the risk of animal to human infection remain poorly understood. Despite improvements in control measures across Europe, GBS is still one of the main causative agents of bovine mastitis in Portugal. Here, by whole-genome analysis of 150 bovine GBS isolates we discovered that a single CC61 clone is spreading throughout Portuguese herds since at least the early 1990s, having virtually replaced the previous GBS population. Mutations within an iron/manganese transporter were independently acquired by all of the CC61 isolates, underlining a key adaptive strategy to persist in the bovine host. Lateral transfer of bacteriocin production and antibiotic resistance genes also underscored the contribution of the microbial ecology and genetic pool within the bovine udder environment to the success of this clone. Compared to strains of human origin, GBS evolves twice as fast in bovines and undergoes recurrent pseudogenizations of human-adapted traits. Our work provides new insights into the potentially irreversible adaptation of GBS to the bovine environment.

Received 20 July, 2016; revised 20 September, 2016; accepted 26 September, 2016. *For correspondence. E-mail pglaser@pasteur.fr; Tel. (+33) 1 45 68 89 96

Introduction

Bovine mastitis is an inflammatory disease of the mammary gland, causing a significant burden for animal welfare and for the dairy industry worldwide (Heikkila *et al.*, 2012; Deb *et al.*, 2013). One of the major pathogens responsible for bovine intramammary infections is *Streptococcus agalactiae* (group B *Streptococcus*, GBS) (Keefe, 1997; Wyder *et al.*, 2011). Control strategies implemented since 1960 were able to reduce the incidence of GBS mastitis in several European countries. However, many dairy farms worldwide continue to be predominantly infected by GBS, or have observed a recent re-emergence of GBS mastitis (Kalmus *et al.*, 2011; Klimiene *et al.*, 2011; Mweu *et al.*, 2012; Bi *et al.*, 2016; Jorgensen *et al.*, 2016). Particularly in Portuguese herds, GBS is among the most frequently detected species in animals diagnosed with mastitis (Almeida *et al.*, 2013; Rato *et al.*, 2013). Besides infecting cattle, GBS asymptomatically colonizes the gastrointestinal tract of 10–30% of the human population (Schuchat, 1998), and is a leading cause of infections in neonates and in immunocompromised individuals (Dermer *et al.*, 2004).

Multilocus sequence typing (MLST) studies of GBS have described various clonal complexes (CCs), and corresponding sequence types (STs) with different host specificities. Phylogenetic studies showed that human GBS throughout the world encompass a generally conserved population of a few clones that were selected following the use of tetracycline in the 1950s onwards (Da Cunha *et al.*, 2014). They correspond most frequently to

clonal complexes 1, 6-8-10, 17, 19 and 23 (Jones *et al*., 2003). Although fewer population analyses of bovine isolates have been performed, the distribution of CCs found in cattle is more variable across different studies. For instance, although CC61/67 is known as a bovine-specific lineage (Sorensen *et al*., 2010) and found to have been widely distributed across both the UK (Bisharat *et al*., 2004) and the USA (Springman *et al*., 2014), a recent study of Norwegian dairy farms did not identify any CC61/67 strains in the 19 herds that were infected by GBS (Jorgensen *et al*., 2016). Specifically in the south-western region of Portugal, genotyping of GBS strains isolated between 2002 and 2003 revealed that they clustered mainly into ST2, ST23, ST61, plus a novel ST61-related lineage, ST554 (Rato *et al*., 2013). In France, a high proportion of bovine GBS belonging to CC23 and CC61/67 were isolated from various geographical areas, while a smaller number of strains were found belonging to CC1, CC2, CC6, CC17 and CC19 (Haenni *et al*., 2010). Indeed, several clonal complexes, such as CC1, CC7, CC23 and CC26 comprise GBS strains adapted to both humans and bovines, suggesting the possibility of interspecies transmission (Oliveira *et al*., 2006; Sorensen *et al*., 2010; Zadoks *et al*., 2011). A PCR-based screening (Richards *et al*., 2011) and a transcriptome analysis of a bovine-specific strain (Richards *et al*., 2013) have looked at some of the genetic factors unique to GBS strains of bovine origin. However, population genomics studies are still lacking to understand the specific adaptation of GBS to the bovine host.

In this work, we performed an in-depth whole-genome analysis of 150 isolates, revealing that cattle throughout Portugal is infected almost exclusively by a single GBS clone. We investigated the adaptive strategies contributing to its evolutionary success, while unveiling new insights into the distinct dynamic between human- and bovine-adapted GBS.

Results

Genotyping of the bovine GBS population

We first analyzed a set of 197 GBS isolates collected between 2011 and 2014 from mastitic milk samples from 15 dairy farms throughout Portugal (Supporting Information Table S1). For a preliminary assessment of their genetic diversity, we sequenced the CRISPR1 locus due to its high discriminatory resolution (Lopez-Sanchez *et al*., 2012). Spacers identified in the leader end of the CRISPR1 locus presented a high degree of heterogeneity with a farm-specific distribution, suggesting that isolates from a single farm are genetically closer (Supporting Information Table S2). Surprisingly, spacers 7 and 147, characteristically found in CC61 strains at the trailer end of the locus (Lopez-Sanchez *et al*., 2012), were completely conserved across all but one isolate that showed an ST2-specific profile.

Based on the CC predicted for each isolate using this approach, a epidemiological analysis of these 197 isolates was performed in combination with earlier collections of GBS strains from the south-west of Portugal (Rato *et al*., 2013) and France (Haenni *et al*., 2010), and a set of 47 isolates from the north of Spain previously genotyped by our group (Lopez-Sanchez *et al*., 2012) (Fig. 1). GBS populations analyzed in these earlier collections were genetically more diverse, with a significant number of isolates belonging to CC2, CC23 or CC61 (Fig. 1). In particular, there was a considerable shift in the proportion of CC61 isolates found in the south-west of Portugal, rising from 53% in 2002–2003 to 98% in 2011–2013, and leading to a replacement of the GBS population (Fig. 1). Indeed, in 2002–2003, only two (33%) out of the six herds sampled were purely infected by CC61 (Rato *et al*., 2013). However, in the 2011–2013 collection, four (80%) out of the five farms analyzed were infected by this clone (Supporting Information Table S1).

Dairy herds in Portugal are almost exclusively infected by a single CC61 clone

To reconstruct the evolutionary history of this population and gain a broader overview of their phylogenetic structure, we selected 128 CC61 isolates from Portugal (118 collected between 2011 and 2014, and 10 between 2002 and 2003), together with the three CC61 isolates from France, to perform whole-genome analysis. The genome of SA111, used as a reference sequence, was completely assembled into a single contig of 2,275,139 base pairs (bp), using long-read sequencing (PacBio). A total of 2212 protein coding genes, 7 rRNA loci and 80 tRNAs genes were predicted and annotated with Prokka (Seemann, 2014). Whole-genome sequencing of 19 ST2 isolates from Portugal and Spain (Supporting Information Tables S1 and S2) was also performed to infer their genetic relatedness and search for analogous mechanisms of adaptation between the CC61 and ST2 bovine GBS clones.

The 131 CC61 isolates selected from Portugal and France were analyzed and compared with 25 CC61 genomes available on the NCBI database (Supporting Information Fig. S1). Phylogenetic analysis showed that all CC61 isolates from Portugal collected since 2002 correspond to one monophyletic clade (Supporting Information Fig. S1). Strikingly, this reveals that the entire CC61 population of GBS infecting dairy farms in Portugal resulted from the recent dissemination of a single clone. Likewise, whole-genome comparison of the 19 ST2 representative isolates, with 19 publicly available genomes, demonstrated a tight clustering of isolates from Portugal and the north of

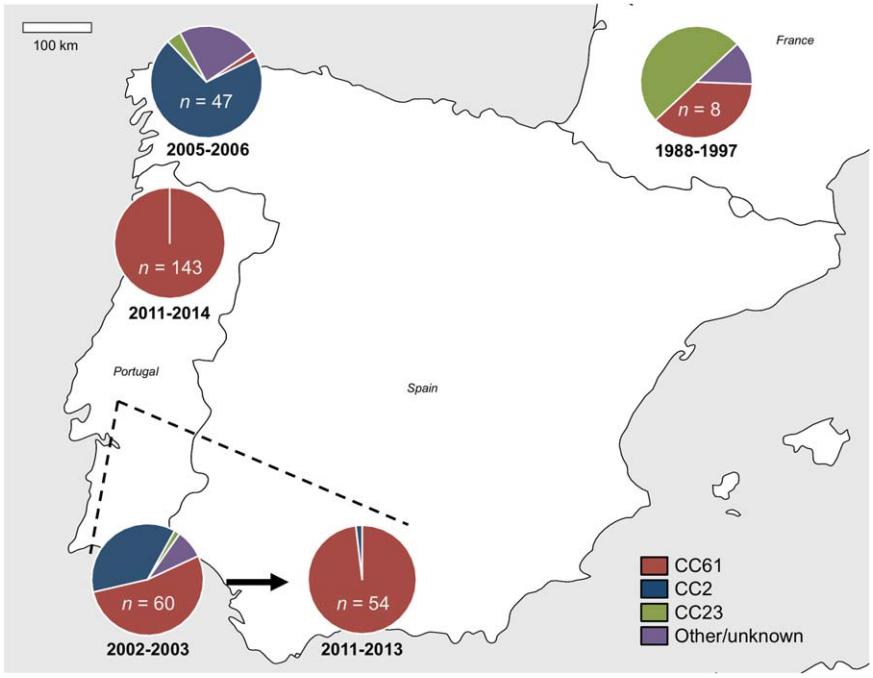


Fig. 1. Genotyping of GBS isolates identified in the south-west of Europe. Map depicting the total number of isolates identified in Portugal, the north of Spain and the south-west of France belonging to different clonal complexes (CCs) according to the figure key. Isolates included in the analysis correspond to those sampled in this work, in addition to other GBS collections isolated from the south-west of Portugal between 2002 and 2003 (Rato *et al.*, 2013), the south-west of France (Haenni *et al.*, 2010) and a collection from the north of Spain.

Spain (Supporting Information Fig. S2), with an average of only 20 SNPs relative to their most recent common ancestor (MRCA).

Recurrent infections are due to contagious transmission and resilience to treatment

For a more detailed analysis of the genetic variation among the CC61 GBS isolates from Portugal, their genomes were mapped and compared with the complete assembly of strain SA111, used as an internal reference. All CC61 GBS isolates from Portugal clustered into two major groups showing an average of 92 SNPs since their MRCA (Fig. 2). The well-defined and herd-specific structure of this population underscores the swift dissemination of GBS within individual farms. One exception was seen in farms C (Setúbal) and W (Centre-West), as isolates collected from these two herds are phylogenetically intermixed – possibly the result of a recent exchange of GBS-infected cattle (Fig. 2). Furthermore, aside from all Portuguese strains being monophyletic, there is no clear geographical distribution of the farm-specific clusters (Fig. 2).

To gain additional insight into the persistence of GBS mastitis in the country, we analyzed all isolates collected within farm PV2 during a 14-month timespan (Fig. 3). For one particular animal, seven isolates were collected at two different time points (t4 and t5) from the four quarters of the udder. Interestingly, isolates that were collected from the same quarter were genetically more similar than to the others obtained from this farm. This suggests that the same strain might have persisted in the udder between the two distinct time points. However, our limited sampling and the presence of within-host GBS heterogeneity in the udder does not exclude the possibility that the later isolates could have also been newly transmitted from other animals within this herd.

The CC61 epidemic clone has persisted in Portugal since at least the early 1990s

The low level of divergence between the CC61 isolates suggests that this epidemic clone disseminated recently. Taking advantage of the sampling of this population across 12 years (Supporting Information Table S1), together with older isolates from France collected between 1996 and 1997, we applied Bayesian phylogenetics to deduce the age of their MRCA (Supporting Information Fig. S3). BEAST (Drummond *et al.*, 2012) analyses estimated that this lineage started to expand in Portugal between 1960 and 1990 (Supporting Information Fig. S3), while diverging at a mean evolutionary rate of 1.69×10^{-6} substitutions/site/year (95% highest posterior densities [HPDs], $1.18–2.23 \times 10^{-6}$ substitutions/site/year). These results infer that the CC61 GBS strains colonizing and infecting cattle in Portugal have persisted since at least 1990. Furthermore, the divergence rate determined for this bovine-specific population is twice as high as that previously inferred for GBS of human origin ($0.56–0.93 \times 10^{-6}$ substitutions/site/year) (Da Cunha *et al.*, 2014).

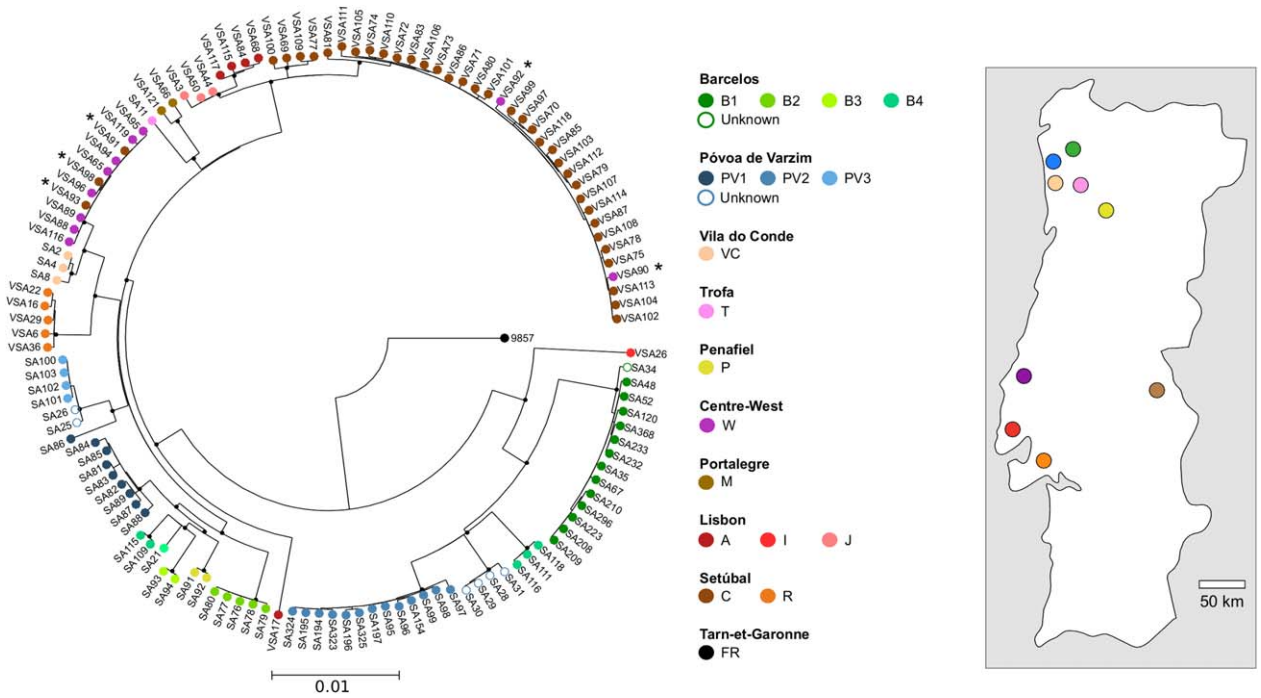


Fig. 2. Geographic distribution and phylogeny of the CC61 epidemic clone from Portugal. The ML phylogenetic tree was built using RAxML (Stamatakis, 2014), based on a total of 1340 core-genome SNPs over 1.51 Mb. Tree includes the 128 CC61 representative isolates from Portugal rooted with one CC61 genome from France (9857). Isolates are colour-coded according to the farm they were isolated from, as depicted in the figure key. Black dots in nodes denote a bootstrap support > 95%. The five isolates that did not cluster exclusively with those within the same farm are indicated with an asterisk. The map of Portugal is depicted on the right with each location coloured as follows: Barcelos (green), Póvoa de Varzim (blue), Vila do Conde (beige), Trofa (pink), Penafiel (yellow), Centre-West (purple), Portalegre (brown), Lisbon (red) and Setúbal (orange).

Convergent adaptations were detected in major transcriptional regulators and within an iron/manganese transporter operon

The clonal expansion of the CC61 population in Portugal allowed us to compare the parallel *in vivo* evolution of multiple lineages during adaptation to the bovine environment. Using an approach adapted from Lieberman and colleagues (2011), we searched for coding regions under natural selection and with recurrent patterns of mutations. We detected a signal of purifying selection at the genome level of all CC61 isolates from Portugal (dN/dS average of 0.56), which suggests that selection is predominantly removing genetic variation from this population. However, we hypothesized that loci independently mutated in different isolates would be under local positive selection. For this analysis, SNPs that were detected in multiple isolates as a result of a single mutational event in their common ancestor were counted as one independent SNP. A total of 1012 independent mutations were detected among coding genes, and 164 independent SNPs within intergenic sequences. After excluding regions with a low density of SNPs, we identified 417 genes and 36 intergenic regions with at least one mutation (Supporting Information Fig. S4). Under a neutral evolutionary model, the mutations we detected would be randomly distributed across the core genome of GBS (Supporting Information Fig. S4). Yet, the number of genes and intergenic regions containing more

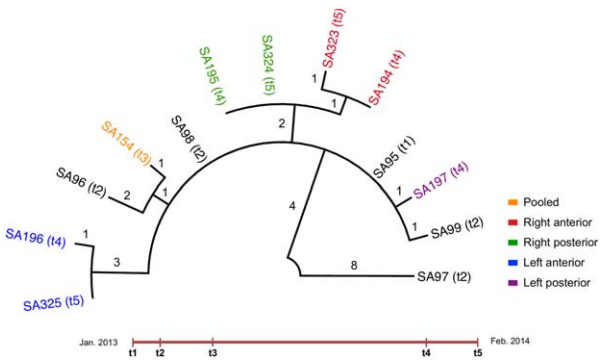


Fig. 3. ML phylogenetic tree built using RAxML (Stamatakis, 2014), highlighting longitudinal samples isolated from farm PV2. Eight isolates collected from one animal are colour-coded according to the quarter of the mammary gland they were isolated from, as indicated in the figure key. Black-coloured isolates correspond to those collected from other animals within the same farm. Digits match the number of SNPs underlying each branch of the tree. Time points t1 to t5 represent the following isolation dates: t1 – January 2013, t2 – February 2013, t3 – April 2013, t4 – December 2013, t5 – February 2014.

Table 1. Mutations in regulatory systems displaying a signal of positive selection.

Locus	Product	Length (bp)	NS[a]	S[b]	Isolates[c]
SA111_00141	HrcA family transcriptional regulator	1083	3	0	37
SA111_00245	Transcriptional antiterminator, BglG family	2037	4	0	56
SA111_01487	Two component system histidine kinase[d]	1230	3	0	45
SA111_01488	Two component system response regulator[d]	693	3	1	12
SA111_01781	Transmembrane histidine kinase CovS	1506	2	1	All
SA111_01923	Transcriptional regulator, GlnR	372	4	1	10
SA111_02131	Phosphate regulon sensor protein PhoR	1656	3	0	All
SA111_02142	Transcriptional regulator, MerR family	717	2	1	10
SA111_02249	TetR family transcriptional regulator	540	5	0	61

a. Number of nonsynonymous substitutions.
b. Number of synonymous substitutions.
c. Number of isolates, out of the 128 CC61 representatives from Portugal that were sequenced, affected by at least one of the mutations.
d. These two genes encode the sensor and regulator of the same two-component system.

than three independent mutations was significantly higher ($P < 0.001$) than expected by neutral drift (Supporting Information Fig. S4). The 85 genes that acquired at least three mutations are involved in a wide range of functions (Supporting Information Table S3). Among those, nine genes with 29 nonsynonymous SNPs and 4 synonymous substitutions correspond to regulatory systems involved in transcription control. dN/dS estimates further reinforced the hypothesis that this group of mutations is under positive selection (dN/dS = 2.25, CI = 1.84–2.47). One of the regulators affected is the CovS sensor histidine kinase (*SA111_01781*; Table 1), part of a two-component system known as CovRS, and a major regulator of virulence in GBS (Lamy *et al*., 2004; Santi *et al*., 2009). Two nonsynonymous mutations were detected in the C-terminal cytoplasmic kinase region of CovS, which is involved in the phosphorylation of its cognate response regulator CovR. An equivalent analysis of the ST2 bovine strains from Spain and Portugal uncovered three independent nonsynonymous substitutions also affecting the *covS* locus (data not shown). This further underscores the contribution of modifications within CovRS to virulence and colonization of the bovine host.

As for the intergenic regions, we saw a strong signal of convergent adaptation in a non-coding region located within a three-gene operon involved in iron and manganese transport (Bray *et al*., 2009) (Table 2 and Fig. 4). Indeed, 14 independent SNPs in 11 different positions were found downstream the first gene of the operon that encodes a metal-binding lipoprotein (Fig. 4A). Altogether, these mutations affected all of the CC61 isolates from Portugal (Table 2 and Fig. 4B). Furthermore, nine of these eleven positions were specifically located in a Rho-independent transcriptional terminator (Fig. 4A) (Rosinski-Chupin *et al*., 2015).

To assess the impact of these mutations, we quantified by RT-qPCR the expression of the genes flanking this internal terminator (*mtsA* and *mtsB*; Fig. 4C). We analyzed six bovine isolates (SA4, SA80, SA109, SA111, VSA66,

Table 2. Intergenic regions that acquired more than three independent mutations.

Region	Length (bp)	Locus[a]	Product	Strand[b]	SNPs	Isolates[c]
IR 1	178	*SA111_00084*	Alcohol/Acetaldehyde dehydrogenase	+	3	8
		SA111_00085	Alcohol dehydrogenase	+		
IR 2	36	*SA111_00149*	Hypothetical protein ywlG	+	3	2
		SA111_00150	Small-conductance channel	−		
IR 3	180	*SA111_00166*	Ribose operon repressor	−	3	39
		SA111_00167	Heme efflux system permease HrtB	+		
IR 4	109	*SA111_01084*	ABC transporter permease protein	+	4	3
		SA111_01085	DNA-binding response regulator	+		
IR 5	143	*SA111_01511*	tRNA-dependent ligase	−	4	3
		SA111_01512	Surface antigen-related protein	−		
IR 6	170	*SA111_01697*	Iron/manganese ABC transporter	−	14	All
		SA111_01698	Iron/manganese ABC transporter	−		
IR 7	138	*SA111_02248*	Phage infection protein	−	3	46
		SA111_02249	TetR family transcriptional regulator	+		

a. For each intergenic region, the first and second rows correspond to the genes identified upstream and downstream, respectively.
b. Gene direction. "+" denotes forward strand and "−" indicates reverse strand.
c. Number of isolates, out of the 128 CC61 representatives from Portugal that were sequenced, affected by at least one of the mutations.

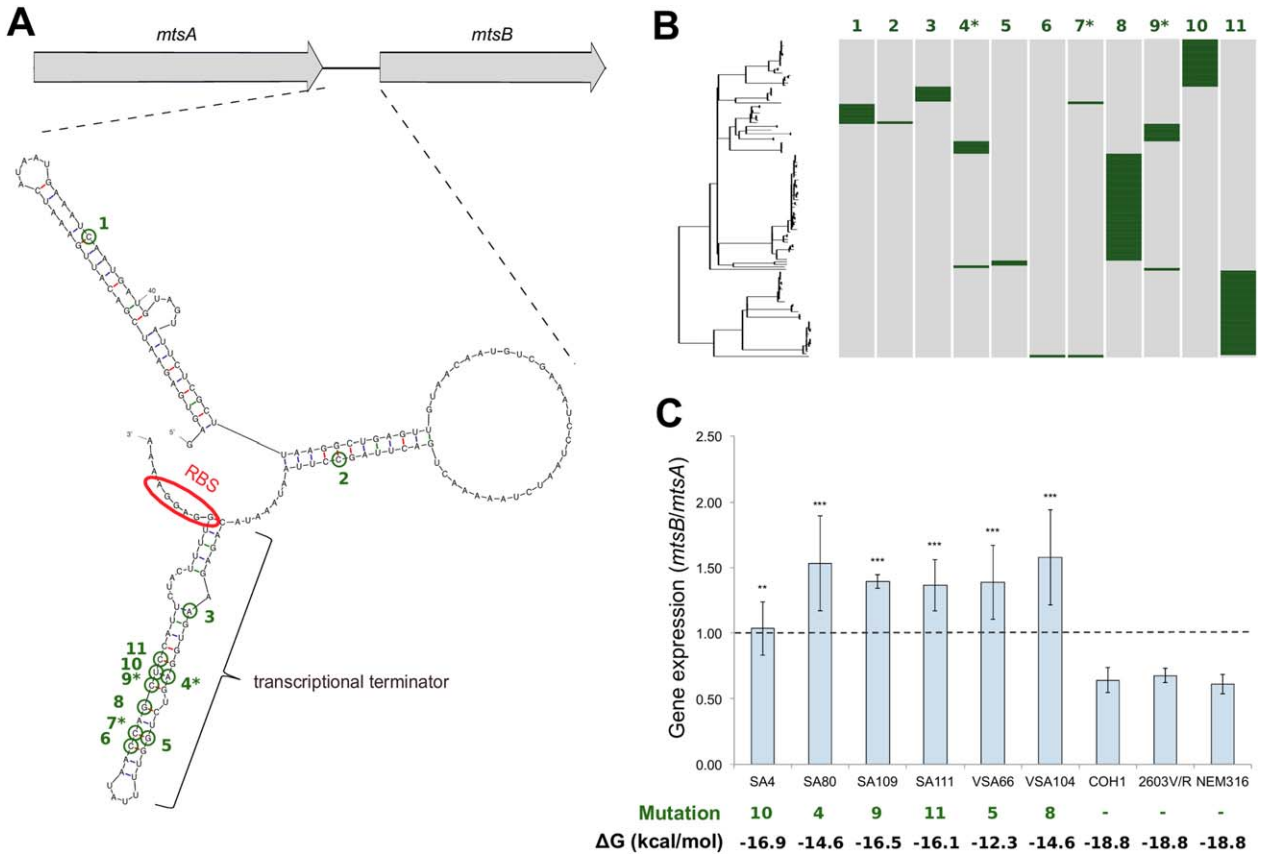


Fig. 4. Intergenic mutations within an operon for iron/manganese transport.
(**a**) Location of the independent mutations found between *mtsA* (SA111_01698) and *mtsB* (SA111_01697), relative to the mRNA secondary structure predicted with mfold (http://unafold.rna.albany.edu/?q = mfold) using the sequence of the MRCA of the CC61 isolates from Portugal. The ribosomal binding site (RBS) is indicated in red and all the positions where independent SNPs were found are circled in dark green and numbered from one to 11. Asterisks indicate that there were two independent mutational events in that position.
(**b**) Mutations acquired in each of the 11 sites (green-coloured boxes), in relation to the core-genome phylogeny of the 128 CC61 isolates from Portugal.
(**c**) RT-qPCR results obtained with six CC61 isolates from Portugal (SA4, SA80 SA109, SA111, VSA66 and VSA104) and three strains used as control (COH1, 2603V/R and NEM316). Gene expression is represented as a ratio between the transcription level of *mtsB* and *mtsA*. Isolate names are indicated below each graph, together with the mutation acquired, as shown in panel (a), and the Gibbs free energy (ΔG) of the terminator structure predicted with mfold. Experiments were performed in triplicate with at least three independent cultures. Error bars represent standard deviation (SD) +/−. A two-tailed *t* test was performed by comparing each isolate against NEM316 (***$P < 0.001$; **$P < 0.01$).

VSA104) that acquired six different mutations within the stem-loop structure of the terminator (Fig. 4C), together with three strains carrying the ancestral sequence (COH1, 2603V/R and NEM316). Although in these control strains tested there was a consistent 30–40% decrease in the expression of *mtsB* after the terminator, none of the bovine CC61 isolates showed any decline in transcription levels ($P < 0.01$; Fig. 4C). By analyzing the impact of each variant on the RNA secondary structure, all the mutations were predicted to reduce the thermodynamic stability of the terminator structure in relation to the parental sequence (Fig. 4C). These results suggest that the mutations present in the CC61 bovine isolates we tested affect the structure of the terminator and lead to an increased expression of the downstream gene (*mtsB*), which encodes the ATP-binding protein of the iron/manganese transporter.

Cohabiting streptococci have contributed to the adaptation of this clone

We used the software Roary (Page *et al.*, 2015) to characterize both the core and accessory genome (i.e., the pan-genome) of the CC61 population. Overall, among the 128 CC61 isolates, a total of 1744 genes were identified as core (present in $\geq$ 99% of the isolates) and 1551 as accessory (missing from at least two isolates). All of the core genes were also present in closely related genomes (9857, 10058, 10059 and LDS610; Supporting Information

Fig. S1) and, therefore, before the divergence of this epidemic lineage in Portugal.

The majority of accessory genes were related to bacteriophages, integrative and conjugative elements or of unknown origin, so we focused on those with functional relevance, specifically for drug resistance and sugar metabolism. We detected independent acquisitions of genes involved in resistance to tetracycline (*tetO*), macrolides (*ermB*), streptomycin (*aadE*) and lincosamides (*lnuC*), as well as of two gene clusters involved in the production and resistance to the lantibiotics macedocin and nisin (Fig. 5). The gene cluster responsible for the biosynthesis and immunity to macedocin was acquired by one of the major sublineages of the CC61 Portuguese population (Fig. 5). By reconstructing the mobile genetic elements (MGEs) carrying the drug-related genes, we found that they were not exclusively present with a significant similarity in GBS, but were also detected among *Streptococcus dysgalactiae*, *Streptococcus suis* and *Streptococcus uberis* (Supporting Information Table S4). Interestingly, *S. dysgalactiae* and *S. uberis* are frequently associated with bovine mastitis. In the ST2-specific isolates, a similar analysis of the accessory and drug-related genes also showed the acquisition of *tetO*, *ermB*, *aadE* and the nisin operon by their MRCA (Supporting Information Fig. S5). Focusing on sugar metabolism, the Lac.2-2 variant of the lactose operon comprising *lacABCDFEGX* was conserved among all of the isolates studied. Additionally, a second copy of the Lac.2 operon (Lac.2-1), also found in *S. dysgalactiae* and *S. uberis* (Supporting Information Table S4), was acquired by two clades of the CC61 population (Fig. 5), but not by the ST2 isolates (Supporting Information Fig. S5).

Recurrent pseudogenizations reveal that adaptation to the bovine host is likely irreversible

We combined the data obtained from the phylogenetic and pan-genome analyses to detect specific functions that may have been lost during the adaptation of GBS to the bovine host. Variant detection and subsequent prediction of their functional effect identified 114 frameshift and 33 nonsense mutations generating a total of 119 pseudogenes (with an average of 12 per isolate).

Intriguingly, several independent mutations were found reoccurring at different loci. A total of 43 isolates acquired altogether seven nonsense or frameshift mutations within the *secA2-Y2* locus (Table 3). Within the *cps* operon, involved in the biosynthesis of the type II capsular polysaccharides, seven inactivating events affected a total of 56 isolates (Table 3 and Fig. 5). In the ST2 population we analyzed, loss or truncation of three genes of the *cps* operon was also detected (Supporting Information Fig. S5). The synthesis of this polysaccharide poses a significant nutritional cost and has also been shown to inhibit biofilm formation (Qin *et al*., 2013; Smitran *et al*., 2013). Additionally, four pseudogenes were detected within a region involved in glycogen biosynthesis (Table 3). Presumably, while colonizing the bovine milk – a sugar-rich environment – GBS might not require an active storage and synthesis of glycogen to survive and grow. Similarly, we observed four independent nonsense or frameshift mutations in the *opuC* operon, involved in the uptake of osmoprotectants which are recruited in high osmolarity conditions (Sleator *et al*., 2001) (Table 3). The milk has a moderate and consistent osmotic concentration (Jackson and Rothera, 1914), so this operon might cause an unnecessary fitness burden during colonization of the udder. Supporting this hypothesis, we also detected an independent frameshift mutation of the *opuCA* gene in the ST2 bovine isolates additionally studied.

Looking at the pan-genome distribution (Fig. 5), two independent events of gene loss were detected within a segment known as the lactose operon Lac.1, which has already been shown to be dispensable for the metabolism of lactose (Loughman and Caparon, 2007). Moreover, a 5-kb region encompassing an ABC transporter coupled with a two-component regulatory system (*vexp*) showed a progressing decay among the CC61 bovine strains studied (Fig. 5). Also in the ST2 isolates, independent loss and inactivation of this genomic island was detected (Supporting Information Fig. S5). Lastly, we found that among the GBS strains that acquired a genomic island for macedocin production and immunity, several isolates have lost part of this genomic cluster (Fig. 5). This suggests that the evolutionary pressure to maintain the functional integrity of this element might have diminished after contributing to the selection of their ancestral clone.

Discussion

Bovine mastitis caused by GBS continues to be a major veterinary and economic issue worldwide, representing a particularly prevalent problem in Portugal. We have performed a whole-genome analysis of GBS strains colonizing Portuguese dairy herds and discovered that they are infected almost exclusively by a single clone belonging to CC61. A Bayesian phylogenetic inference allowed us to estimate that this particular CC61 clone has been expanding in Portugal since at least the early 1990s (Supporting Information Fig. S3). The development of the agriculture sector, concurrent with Portugal's entry into the European Economic Community (EEC) in 1986, led to the expansion of dairy herds and the implementation of milking parlours in the country. This might have been conducive to the spread of the CC61 GBS clone we uncovered, and to its replacement of the existing GBS population.

Our whole-genome epidemiological analysis tracked bovine infections across multiple farms and within a single

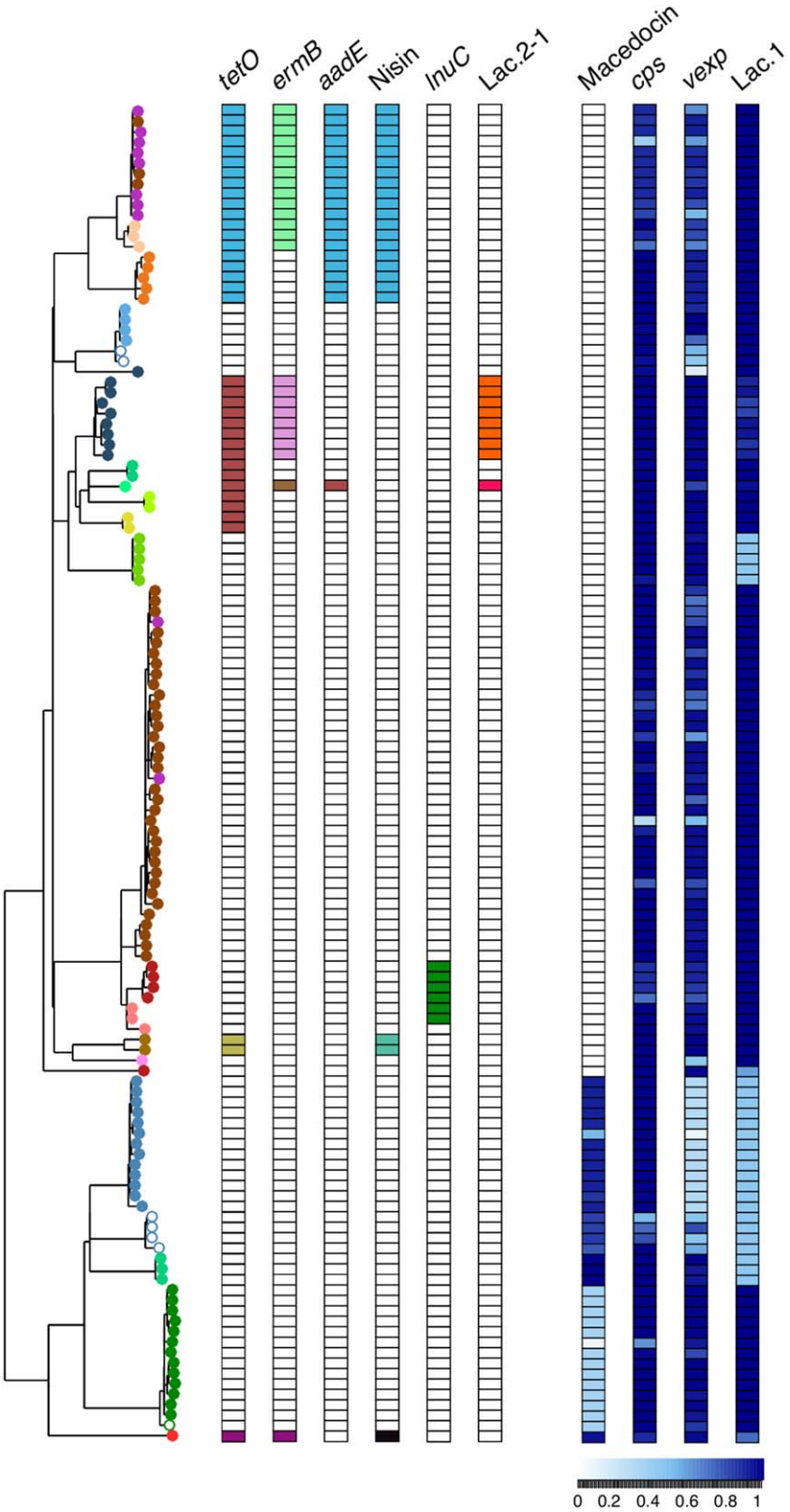


Fig. 5. Distribution of the most functionally relevant mobile genetic elements, involved in drug resistance (*tetO*, *ermB*, *aadE*, Nisin, *lnuC* and Macedocin) and sugar transport/metabolism (Lac.2-1 and Lac.1) in relation to the core-genome phylogeny of the 128 CC61 isolates from Portugal colour-coded as in Fig. 2. The ongoing degradation of the capsule operon (*cps*) and the *vexp* region are depicted on the right. White denotes absence, and for *tetO*, *ermB*, *aadE*, *nisin*, *lnuC* and Lac.2-1, equally coloured boxes mean they were detected within the same contiguous sequence. The genetic decay of the *cps*, *vexp* and the Lac.1 loci is depicted by a gradient from white to dark blue corresponding to a blast score ratio ranging from 0 to 1, as indicated in the figure key.

Table 3. Genomic regions most frequently inactivated.

Locus	Product	NS[a]	FS[b]	Isolates[c]
SA111_00292	L-carnitine/choline ABC transporter, OpuCD	0	1	11
SA111_00293	L-carnitine/choline ABC transporter, OpuCC	0	1	
SA111_00295	L-carnitine/choline ABC transporter, OpuCA	0	2	
SA111_00962	Glycogen debranching protein	0	2	17
SA111_00963	Glycogen branching enzyme, GlgB	0	1	
SA111_00964	Glucose-1-phosphate adenylyltransferase, GlgC	1	0	
SA111_00965	Glycogen biosynthesis protein GlgD	Missing[d]		
SA111_01329	Capsular polysaccharide repeat unit transporter CpsL	Missing[d]		56
SA111_01330	Capsular polysaccharide biosynthesis protein CpsK	0	1	
SA111_01333	Polysaccharide biosynthesis glycosyl transferase CpsJ	Missing[d]		
SA111_01334	Capsular biosynthesis protein CpsI	Missing[d]		
SA111_01335	Polysaccharide biosynthesis protein CpsH	Missing[d]		
SA111_01337	Polysaccharide biosynthesis protein CpsF	0	1	
SA111_01338	Galactosyl transferase CpsE	Missing[d],[e]		
SA111_01616	Glycosyltransferase GftA	1	0	43
SA111_01617	Protein export cytoplasm protein SecA2	1	1	
SA111_01619	Accessory secretory protein Asp2	0	1	
SA111_01621	Preprotein translocase SecY2 subunit	0	2	
SA111_01622	Glycosyl transferase, putative	1	0	

a. Number of nonsense mutations.
b. Number of frameshift mutations.
c. Number of isolates, out of the 128 CC61 representatives from Portugal that were sequenced, affected by at least one of the mutations.
d. Missing = loss/truncation of the gene.
e. Insertion sequence (IS) inserted within the gene in 34 isolates.

bovine host (Figs 2 and 3). Based on these observations, we suggest that the prevalence of this CC61 GBS lineage in Portugal is due both to persistent colonization of the same animal, as well as to its dissemination throughout the herd by cross-contamination. While GBS may be most frequently transferred during the milking process – through contamination of the milking apparatus – a recent study also found alternative environmental routes of transmission (Jorgensen *et al.*, 2016).

Distinguishing neutral from adaptive mutations is essential to understand how a pathogen adapts to its environment. dN/dS analyses of the CC61 isolates that expanded in Portugal underlined a general level of strong purifying selection. This suggests that the bovine environment imposes significant selective constraints and is quickly purging stochastic mutations that diminish the fitness of the CC61 population. Therefore, we reason that the ubiquitous prevalence of this epidemic clone in Portugal is most likely a result of its competitive advantage in the bovine host, rather than a product of chance. Conversely, the frequent occurrence of nonsynonymous mutations among major transcriptional regulators suggests these specific changes might be providing a competitive advantage during colonization of the bovine udder and milk (Table 1). Indeed, regulatory systems like CovRS are known to affect pathogenesis by controlling the expression of virulence-associated genes in response to different environmental stimuli (Lamy *et al.*, 2004; Santi *et al.*, 2009; Almeida *et al.*, 2015). Of special interest were the functionally significant mutations detected within an iron/manganese transporter operon that independently affected all of the isolates studied (Fig. 4). Trace metals are naturally found in low concentrations in the bovine milk (Lonnerdal *et al.*, 1981). Additionally, during the dry period of the bovine lactation cycle, the increased concentration of lactoferrin in the mammary gland further depletes the milk of metal nutrients essential for bacterial growth and survival (Hurley and Rejman, 1993). Thus, the mutations we observed might contribute to a higher uptake of iron and manganese by GBS, potentially underlying an adaptive strategy for this dominant clone to persist in more challenging conditions. Supporting this hypothesis, the acquisition of manganese during growth in milk was shown to be essential for infection of the bovine mammary gland by *S. uberis* (Smith *et al.*, 2003).

Mobile genetic elements (MGEs) have been shown to play a major role in the evolution of GBS (Brochet *et al.*, 2008a; Richards *et al.*, 2011), while representing a genomic record of its interaction with other neighbouring bacteria. The accessory genes shared with other mastitis-causing streptococci reveal the contribution of the genetic repertoire of bacterial species circulating in the bovine environment to the evolutionary success of the CC61 clone (Fig. 5 and Supporting Information Table S4). In particular, the acquisition of genes involved in the synthesis and resistance to macedocin correlates with the expansion of one of the two major clades of the CC61 population in Portugal (Fig. 5). Macedocin is known to inhibit the growth of a wide range of lactic acid bacteria and several pathogenic

species (Georgalaki *et al*., 2002). The genomic island specifically acquired by the CC61 isolates is similar to that first described in *Streptococcus macedonicus* (Papadelli *et al*., 2007), a species normally isolated from fermented milk products. Therefore, macedocin production and resistance might have contributed to the selective advantage of this clone and to its replacement of the local GBS population.

How the adaptation of GBS differs within humans and bovines has been largely unknown, albeit crucial to understand the risk of interspecies transmission. The long-term persistence of this dominant CC61 clone allowed us to uncover the genomic footprint of GBS's adaptation to the bovine host. The faster evolutionary rate that was estimated for the bovine-adapted population probably stems from the different conditions encountered by GBS in humans and cattle. In the milk, a more nutrient-rich and bacteria-poor environment, GBS might display a faster growth rate, while the milking process promotes the continuous growth of the bacteria by regularly replenishing their natural medium. Evolutionary studies have shown that host specialization may be reflected by reductive evolutionary processes associated with gene loss and inactivation (Toft and Andersson, 2010). Bovine-specific strains analyzed in this work showed a directed and recurrent pseudogenization of multiple genes, such as the *secA2-Y2*, *cps* and *vexp* loci. The *secA2-Y2* system was shown in GBS to be involved in the synthesis and export of the serine-rich glycoprotein Srr1, which promotes bacterial adhesion to human lung epithelial cells (Mistou *et al*., 2009). Moreover, the *vexp* region is highly conserved exclusively among human GBS strains and was suggested to play a role in their adaptation to the human gut (Brochet *et al*., 2008b). Therefore, these data reflect distinct requirements for the adaptation of GBS to humans and bovines. The loss of human-adapted traits might not allow GBS to as efficiently invade and recolonize the human host, suggesting that interspecies transmission is more permissive from humans to cattle. In addition, our results were further supported by the occurrence of a similar evolutionary trend in the ST2 clone dominant in Galicia, which represents a recent adaptation to the bovine environment (Supporting Information Fig. S2).

Our study illustrates an intriguing example of the differential host adaptation of a generalist species, and how whole-genome analysis may be leveraged to understand its underlying potential for causing disease in different environments. These results will also be instrumental to eradicate this dominant CC61 clone and prevent its ongoing dissemination.

Experimental procedures

Sequencing, genome assembly and annotation

A total of 234 GBS isolates collected from Portugal, Spain and France were analyzed in this study (Supporting Information Table S1). The CRISPR1 locus of all GBS isolates was sequenced and analyzed as previously described (Lopez-Sanchez *et al*., 2012). One hundred and fifty GBS isolates (131 from CC61 and 19 from ST2) were then chosen for whole-genome sequencing, based on the diversity of CRISPR spacer profiles identified, as well as on their temporal and geographical distribution (Supporting Information Tables S1 and S2). Genomes were sequenced using the Illumina HiSeq 2000 platform with single- or paired-end read runs of 101 bp. Reads were filtered for quality and genome sequences were assembled using the Velvet software (Zerbino and Birney, 2008) with an optimized k-mer length, a minimum coverage of 10 and a contig length of at least 200 bp. Strain SA111 was additionally selected for single molecule real-time sequencing (PacBio RS II system). PacBio subreads were assembled with both PBcR (Berlin *et al*., 2015) and the RS_HGAP_Assembly.3 protocol from the SMRT analysis toolkit v2.3. A finished assembly was achieved by complementing the predictions obtained by both tools. Consensus accuracy was further polished using Quiver (Chin *et al*., 2013) and any remaining indels within homopolymer regions were manually corrected with the corresponding Illumina reads. Assembled genomes were annotated with Prokka (Seemann, 2014) and global pan-genome analyses were carried out using Roary (Page *et al*., 2015).

Sequencing reads and corresponding genome assemblies have been deposited in the EMBL nucleotide sequence database (http://www.ebi.ac.uk/ena) under study accession number PRJEB12926. For the accession numbers of the individual samples, see Supporting Information Table S1.

Genome mapping, variant calling and phylogenetic inference

Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) was used to map reads of each CC61 isolate against the complete genome of SA111, and of each ST2 isolate against the draft assembly of VSA10. Variant calling was performed with Genome Analysis ToolKit (GATK) (McKenna *et al*., 2010) according to the published recommendations (Van der Auwera *et al*., 2013; DePristo *et al*., 2011).

Phylogenetic trees of the CC61 and ST2 isolates studied in combination with publicly available GBS genomes were built from the core-genome alignment of their assemblies with Parsnp (Treangen *et al*., 2014) (Supporting Information Figs S1 and S2). Genomes with no more than one allele difference in relation to the ST61 profile were considered CC61. Phylogeny of the CC61 isolates from Portugal (Fig. 2) was inferred from the polymorphic positions detected in the variant calling workflows. Recombinant sites were removed following identification with Gubbins (Croucher *et al*., 2015). Variants present in the accessory genome of these strains, determined using the filter_BSR_variome.py script from the Large Scale BLAST score ratio (LS-BSR) pipeline (Sahl *et al*., 2014), were also not considered. Maximum-likelihood (ML) phylogenies were inferred with RAxML (Stamatakis, 2014), using a General Time-Reversible (GTR) substitution model with a gamma-distributed rate across sites combined with an ascertainment bias to take into account the sole use of variable positions in the alignment. ML trees were bootstrapped with 1000 replicates.

To investigate the temporal evolution of the GBS isolates included in this study, the Bayesian phylogenetic software BEAST v1.8.2 (Drummond *et al.*, 2012) was used to calibrate the evolutionary rate with the corresponding sampling date of each isolate (Supporting Information Table S1). We used a GTR substitution model considering a gamma-distributed rate across sites with a proportion of invariant positions. To identify the most suitable tree and clock models, we compared the strict, uncorrelated lognormal relaxed and uncorrelated exponential relaxed clock models together with coalescent constant, exponential growth, expansion growth and Bayesian skyline tree models. BEAST runs for model testing were conducted in duplicate for 50 million Markov Chain Monte Carlo (MCMC) generations with samples taken every 1000 generations. The best model was deduced by marginal likelihood estimation with stepping-stone/path sampling. The constant size tree model and the uncorrelated exponential relaxed clock model were preferred. Final phylogenetic analysis was run for 100 million MCMC generations with 10% burn-in using the above parameters. Final run presented good mixing and convergence with effective sample size (ESS) values above 200 for all parameters. The presence of temporal structure was further validated by comparison with 10 simulated datasets with randomized sampling dates, as previously described (Firth *et al.*, 2010).

Variant annotation, parallel evolution and dN/dS

The functional effect of each SNP detected within coding sequences was classified with snpEff (Cingolani *et al.*, 2012) as either nonsynonymous (N) or synonymous (S). In the case of indels, only frameshift mutations were considered. To detect parallel evolution in a given number of mutations observed (m_t), 1000 simulations of m_t in the reference genome were performed to establish an expected distribution of the number of mutations per gene. One-tailed *P* values were calculated by assessing the frequency in which the number of mutations per gene observed was higher than the simulated expectation. Subsequently, the same analysis was performed while only taking into account intergenic regions. Genes and intergenes with a low SNP density, i.e. fewer than one SNP per their average length, were excluded. For dN/dS calculations, the observed spectrum of *N* and *S* mutations per nucleotide change was normalized by an expected frequency simulated for the whole genome, as previously described (Lieberman *et al.*, 2014). Values above 1 are indicative of positive selection. Clopper–Pearson confidence intervals (CI) were calculated by the binomial test.

RT-qPCRs

Bacterial cultures were grown to the exponential growth phase (OD_{600} = 0.4–0.5) in 15 ml Todd Hewitt (TH) broth at 37°C. RNA extraction, reverse transcription and RT-qPCRs were performed as previously described (Lamy *et al.*, 2004), using the primers indicated in Supporting Information Table S5. Relative gene expression was quantified with a standard curve-based method, in which a regression analysis was performed using serial dilutions of genomic DNA from strain VSA104. Each assay was performed with three experimental replicates starting from at least three independent cultures. A two-tailed *t*-test was carried out to determine whether the expression differences were statistically significant. RNA secondary structures and thermodynamic stabilities were predicted with mfold (http://unafold.rna.albany.edu/?q = mfold).

Acknowledgements

This work was supported by ANR-LabEx project IBEID. Partial support was also obtained through UCIBIO (UID/Multi/04378/2013 and POCI-01-0145-FEDER-007728) and Project PTDC/CVT-EPI/6685/2014. Sequencing was performed at the Pasteur Genopole, a member of France Génomique (ANR10-IBNS-09-08). A.A. is a scholar in the Pasteur – Paris University (PPU) International PhD program and received a stipend from the ANR-LabEx IBEID. Authors would like to thank Laurence Ma for her help in performing the Illumina sequencing. They also thank SEGALAB, the Laboratorio de Sanidade e Produción Animal de Galicia, Xunta de Galicia and Sophie Payot for providing isolates used in this study. Authors also thank Isabelle Rosinski-Chupin, Pierre-Emmanuel Douarre, Niza Ribeiro, Bruno Gonzalez-Zorn and Helena Madeira for fruitful discussions.

References

Almeida, A., Albuquerque, P., Araujo, R., Ribeiro, N., and Tavares, F. (2013) Detection and discrimination of common bovine mastitis-causing streptococci. *Vet Microbiol* **164:** 370–377.

Almeida, A., Villain, A., Joubrel, C., Touak, G., Sauvage, E., Rosinski-Chupin, I., *et al.* (2015) Whole-genome comparison uncovers genomic mutations between group B streptococci sampled from infected newborns and their mothers. *J Bacteriol* **197:** 3354–3366.

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., and Phillippy, A.M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33:** 623.

Bi, Y.L., Wang, Y.J., Qin, Y., Vallverdu, R.G., Garcia, J.M., Sun, W., *et al.* (2016) Prevalence of bovine mastitis pathogens in bulk tank milk in China. *PLos One* **11:** 13.

Bisharat, N., Crook, D.W., Leigh, J., Harding, R.M., Ward, P.N., Coffey, T.J., *et al.* (2004) Hyperinvasive neonatal group B *Streptococcus* has arisen from a bovine ancestor. *J Clin Microbiol* **42:** 2161–2167.

Bray, B.A., Sutcliffe, I.C., and Harrington, D.J. (2009) Expression of the MtsA lipoprotein of *Streptococcus agalactiae* A909 is regulated by manganese and iron. *Antonie Van Leeuwenhoek* **95:** 101–109.

Brochet, M., Couve, E., Glaser, P., Guedon, G., and Payot, S. (2008a) Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *J Bacteriol* **190:** 6913–6917.

Brochet, M., Rusniok, C., Couve, E., Dramsi, S., Poyart, C., Trieu-Cuot, P., *et al.* (2008b) Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. *Proc Natl Acad Sci USA* **105:** 15961–15966.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., *et al.* (2013) Nonhybrid, finished microbial

genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10:** 563.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Tung, N., Wang, L., *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly* **6:** 80–92.

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., *et al*. (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43.**

Da Cunha, V., Davies, M.R., Douarre, P.E., Rosinski-Chupin, I., Margarit, I., Spinali, S., *et al*. (2014) *Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nat Commun* **5:** 11.

Deb, R., Kumar, A., Chakraborty, S., Verma, A.K., Tiwari, R., Dhama, K., *et al*. (2013) Trends in diagnosis and control of bovine mastitis: A review. *Pakistan J Biol Sci* **16:** 1653–1661.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., *et al*. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491.

Dermer, P., Lee, C., Eggert, J., and Few, B. (2004) A history of neonatal group B *Streptococcus* with its related morbidity and mortality rates in the United States. *J Pediatric Nurs* **19:** 357–363.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29:** 1969–1973.

Firth, C., Kitchen, A., Shapiro, B., Suchard, M.A., Holmes, E.C., and Rambaut, A. (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol* **27:** 2038–2051.

Georgalaki, M.D., Van den Berghe, E., Kritikos, D., Devreese, B., Van Beettmen, J., Kalantzopoulos, G., *et al*. (2002) Macedocin, a food-grade lantibiotic produced by *Streptococcus macedonicus* ACA-DC 198. *Appl Environ Microbiol* **68:** 5891–5903.

Haenni, M., Saras, E., Bertin, S., Leblond, P., Madec, J.Y., and Payot, S. (2010) Diversity and mobility of integrative and conjugative elements in bovine isolates of *Streptococcus agalactiae*, *S. dysgalactiae subsp. dysgalactiae*, and *S. uberis*. *Appl Environ Microbiol* **76:** 7957–7965.

Heikkila, A.M., Nousiainen, J.I., and Pyorala, S. (2012) Costs of clinical mastitis with special reference to premature culling. *J Dairy Sci* **95:** 139–150.

Hurley, W.L., and Rejman, J.J. (1993) Bovine lactoferrin in involuting mammary tissue. *Cell Biol Int* **17:** 283–289.

Jackson, L.C., and Rothera, A.C. (1914) Milk – Its milk sugar, conductivity and depression of freezing point. *Biochem J* **8:** 1–27.

Jones, N., Bohnsack, J.F., Takahashi, S., Oliver, K.A., Chan, M.S., Kunst, F., *et al*. (2003) Multilocus sequence typing system for group B *Streptococcus*. *J Clin Microbiol* **41:** 2530–2536.

Jorgensen, H.J., Nordstoga, A.B., Sviland, S., Zadoks, R.N., Solverod, L., Kvitle, B., and Mork, T. (2016) *Streptococcus agalactiae* in the environment of bovine dairy herds – Rewriting the textbooks? *Vet Microbiol* **184:** 64–72.

Kalmus, P., Aasmaee, B., Kaerssin, A., Orro, T., and Kask, K. (2011) Udder pathogens and their resistance to antimicrobial agents in dairy cows in Estonia. *Acta Vet Scand* **53.**

Keefe, G.P. (1997) *Streptococcus agalactiae* mastitis: A review. *Can Vet J* **38:** 429–437.

Klimiene, I., Ruzauskas, M., Spakauskas, V., Matusevicius, A., Mockeliunas, R., Pereckiene, A., *et al*. (2011) Antimicrobial resistance patterns to beta-lactams of gram-positive cocci isolated from bovine mastitis in Lithuania. *Polish J Vet Sci* **14:** 467–472.

Lamy, M.C., Zouine, M., Fert, J., Vergassola, M., Couve, E., Pellegrini, E., *et al*. (2004) CovS/CovR of group B *Streptococcus*: A two-component global regulatory system involved in virulence. *Mol Microbiol* **54:** 1250–1268.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Lieberman, T.D., Michel, J.B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R., Jr, *et al*. (2011) Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* **43:** 1275.

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

Lonnerdal, B., Keen, C.L., and Hurley, L.S. (1981) Iron, copper, zinc, and manganese in milk. *Annu Rev Nutr* **1:** 149–174.

Lopez-Sanchez, M.J., Sauvage, E., Da Cunha, V., Clermont, D., Hariniaina, E.R., Gonzalez-Zorn, B., *et al*. (2012) The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol Microbiol* **85:** 1057–1071.

Loughman, J.A., and Caparon, M.G. (2007) Comparative functional analysis of the *lac* operons in *Streptococcus pyogenes*. *Mol Microbiol* **64:** 269–280.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., *et al*. (2010) The Genome Analysis ToolKit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.

Mistou, M.Y., Dramsi, S., Brega, S., Poyart, C., and Trieu-Cuot, P. (2009) Molecular dissection of the *secA2* locus of group B *Streptococcus* reveals that glycosylation of the Srr1 LPXTG protein is required for full virulence. *J Bacteriol* **191:** 4195–4206.

Mweu, M.M., Nielsen, S.S., Halasa, T., and Toft, N. (2012) Annual incidence, prevalence and transmission characteristics of *Streptococcus agalactiae* in Danish dairy herds. *Prev Vet Med* **106:** 244–250.

Oliveira, I.C.M., de Mattos, M.C., Pinto, T.A., Ferreira-Carvalho, B.T., Benchetrit, L.C., Whiting, A.A., *et al*. (2006) Genetic relatedness between group B streptococci originating from bovine mastitis and a human group B *Streptococcus* type V cluster displaying an identical pulsed-field gel electrophoresis pattern. *Clin Microbiol Infect* **12:** 887–893.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., *et al*. (2015) Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31:** 3691–3693.

Papadelli, M., Karsioti, A., Anastasiou, R., Georgalaki, M., and Tsakalidou, E. (2007) Characterization of the gene

cluster involved in the biosynthesis of macedocin, the lantibiotic produced by *Streptococcus macedonicus*. *FEMS Microbiol Lett* **272:** 75–82.
Qin, L., Kida, Y., Imamura, Y., Kuwano, K., and Watanabe, H. (2013) Impaired capsular polysaccharide is relevant to enhanced biofilm formation and lower virulence in *Streptococcus pneumoniae*. *J Infect Chemother* **19:** 261–271.
Rato, M.G., Bexiga, R., Florindo, C., Cavaco, L.M., Vilela, C.L., and Santos-Sanches, I. (2013) Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. *Vet Microbiol* **161:** 286–294.
Richards, V.P., Lang, P., Bitar, P.D.P., Lefebure, T., Schukken, Y.H., Zadoks, R.N., and Stanhope, M.J. (2011) Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. *Infect Genet Evol* **11:** 1263–1275.
Richards, V.P., Choi, S.C., Bitar, P.D.P., Gurjar, A.A., and Stanhope, M.J. (2013) Transcriptomic and genomic evidence for *Streptococcus agalactiae* adaptation to the bovine environment. *BMC Genom* **14:** 15.
Rosinski-Chupin, I., Sauvage, E., Sismeiro, O., Villain, A., Da Cunha, V., Caliot, M.E., *et al*. (2015) Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen *Streptococcus agalactiae*. *BMC Genom* **16.**
Sahl, J.W., Caporaso, J.G., Rasko, D.A., and Keim, P. (2014) The large-scale blast score ratio (LS-BSR) pipeline: A method to rapidly compare genetic content between bacterial genomes. *Peerj* **2.**
Santi, I., Grifantini, R., Jiang, S.M., Brettoni, C., Grandi, G., Wessels, M.R., and Soriani, M. (2009) CsrRS regulates group B *Streptococcus* virulence gene expression in response to environmental pH: A new perspective on vaccine development. *J Bacteriol* **191:** 5387–5397.
Schuchat, A. (1998) Epidemiology of group B streptococcal disease in the United States: Shifting paradigms. *Clin Microbiol Rev* **11:** 497.
Seemann, T. (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30:** 2068–2069.
Sleator, R.D., Wouters, J., Gahan, C.G.M., Abee, T., and Hill, C. (2001) Analysis of the role of OpuC, an osmolyte transport system, in salt tolerance and virulence potential of *Listeria monocytogenes*. *Appl Environ Microbiol* **67:** 2692–2698.
Smith, A.J., Ward, P.N., Field, T.R., Jones, C.L., Lincoln, R.A., and Leigh, J.A. (2003) MtuA, a lipoprotein receptor antigen from *Streptococcus uberis*, is responsible for acquisition of manganese during growth in milk and is essential for infection of the lactating bovine mammary gland. *Infect Immunity* **71:** 4842–4849.
Smitran, A., Opavski, N.V., Eric-Marinkovic, J., Gajic, I., and Ranin, L. (2013) Adherence and biofilm production of invasive and non-invasive isolates of *Streptococcus pyogenes* after hyaluronidase treatment. *Arch Biol Sci* **65:** 1353–1361.
Sorensen, U.B., Poulsen, K., Ghezzo, C., Margarit, I., and Kilian, M. (2010) Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. *mBio* **1.**
Springman, A.C., Lacher, D.W., Waymire, E.A., Wengert, S.L., Singh, P., Zadoks, R.N., *et al*. (2014) Pilus distribution among lineages of group B *Sreptococcus*: An evolutionary and clinical perspective. *BMC Microbiol* **14.**
Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.
Toft, C., and Andersson, S.G.E. (2010) Evolutionary microbial genomics: Insights into bacterial host adaptation. *Nat Rev Genet* **11:** 465–475.
Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. (2014) The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* **15.**
Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., *et al*. (2013) From fastQ data to high-confidence variant calls: The Genome Analysis ToolKit best practices pipeline. In *Current Protocols in Bioinformatics*. Alex Bateman, Sorin Draghici, Ekta Khurana, Sandra Orchard and William R. Pearson (eds.). Wiley: Hoboken, NJ, USA. pp. 11.10.1-11.10.33.
Wyder, A.B., Boss, R., Naskova, J., Kaufmann, T., Steiner, A., and Graber, H.U. (2011) *Streptococcus* spp. and related bacteria: Their identification and their pathogenic potential for chronic mastitis – A molecular approach. *Res Vet Sci* **91:** 349–357.
Zadoks, R.N., Middleton, J.R., McDougall, S., Katholm, J., and Schukken, Y.H. (2011) Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. *J Mamm Gland Biol Neoplasia* **16:** 357–372.
Zerbino, D.R., and Birney, E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Core-genome phylogeny of CC61. The ML phylogeny was built using RAxML (Stamatakis, 2014), based on a total of 4524 polymorphic positions. Tree depicts all the publicly available CC61 genomes, as well as the 128 CC61 representative isolates from Portugal analyzed in this work and the three additional isolates sampled from the southwest of France (9857, 10058, 10059). Sequence type (ST) and location of each isolate (if available) are plotted in the tree as depicted in the figure key. All genomes depicted are of bovine origin, except for ATCC13813, which was isolated from the oral cavity of a human host. All tree nodes presented a bootstrap support $>75\%$. Tree was rooted with the CC1 genome of GBS 09mas01883 (not shown).
Fig. S2. Core-genome phylogeny of ST2 genomes. ML phylogenetic tree built using RAxML (Stamatakis, 2014) based on a total of 1403 SNPs. Tree depicts all the publicly available ST2 genomes, together with the 19 ST2 isolates from Portugal and Spain analyzed in this work. Host origin and location of each isolate (if available) are plotted in the tree as depicted in the figure key. All tree nodes presented a bootstrap support $>75\%$. Tree was rooted with the CC19 genome of GBS 2603V/R (not shown).
Fig. S3. Timed phylogeny of the CC61 isolates. Bayesian phylogenetic inference of the 128 CC61 isolates sequenced from Portugal and the three additional CC61 isolates from France, performed with BEAST (Drummond *et al*., 2012). The CC61 isolates are colour-coded as in Fig. 2. Interval

dates in brackets represent 95% HPDs of the estimated date of each node.

Fig. S4. Parallel evolution in genes and in intergenic regions. Number of genes (**a**) or intergenic regions (**b**) with at least *m* mutations, as a function of *m*. Expected values correspond to the average of 1000 random distributions of the number of mutations observed. Error bars represent standard deviation (SD) +/−. $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$.

Fig. S5. Analysis of most functionally relevant accessory genes identified in the CC61 population, against the core-genome phylogeny of the 19 ST2 isolates from Portugal and Spain. Isolates are colour-coded as in Supporting Information Fig. S2. White denotes absence, and for *tetO*, *ermB* and *aadE*, different coloured boxes mean they were detected within distinct contiguous sequences. Similarities with the nisin operon, *cps*, *vexp* and Lac.1 loci are depicted by a gradient from white to dark blue corresponding to a blast score ratio ranging from 0 to 1, as indicated in the figure key.

Table S1. *S. agalactiae* strains used in this study.

Table S2. CRISPR spacer profiles of the *S. agalactiae* strains used in this study.

Table S3. Genes that acquired more than three independent mutations.

Table S4. BLASTn results of the drug resistance and metabolism genes acquired by the CC61 clone.

Table S5. Primer pairs used in the RT-qPCRs.

Table S1. *S. agalactiae* strains used in this study.

Portugal (2011-2014)

Strain	Farm	Location[a]	Year	MLST	Accession ID
SA2	VC	Vila do Conde, Portugal	2011	ST61	ERS1075201
SA3	VC	Vila do Conde, Portugal	2011	ST61	n.s.[b]
SA4	VC	Vila do Conde, Portugal	2011	ST61	ERS1075203
SA6	VC	Vila do Conde, Portugal	2011	ST61	n.s.
SA7	VC	Vila do Conde, Portugal	2011	ST61	n.s.
SA8	VC	Vila do Conde, Portugal	2011	ST61	ERS1075206
SA9	VC	Vila do Conde, Portugal	2011	ST61	n.s.
SA10	VC	Vila do Conde, Portugal	2011	ST61	n.s.
SA11	T	Trofa, Portugal	2011	ST61	ERS1075209
SA21	Unknown	Barcelos, Portugal	2011	ST61	ERS1075210
SA25	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075211
SA26	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075212
SA28	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075213
SA29	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075214
SA30	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075215
SA31	Unknown	Póvoa de Varzim, Portugal	2011	ST61	ERS1075216
SA32	Unknown	Barcelos, Portugal	2011	ST61	n.s.
SA33	Unknown	Barcelos, Portugal	2011	ST61	n.s.
SA34	Unknown	Barcelos, Portugal	2011	ST61	ERS1075219
SA35	B1	Barcelos, Portugal	2012	ST61	ERS1075220
SA36	B1	Barcelos, Portugal	2012	ST61	n.s.
SA37	B1	Barcelos, Portugal	2012	ST61	n.s.
SA38	B1	Barcelos, Portugal	2012	ST61	n.s.
SA39	B1	Barcelos, Portugal	2012	ST61	n.s.
SA40	B1	Barcelos, Portugal	2012	ST61	n.s.
SA41	B1	Barcelos, Portugal	2012	ST61	n.s.
SA42	B1	Barcelos, Portugal	2012	ST61	n.s.
SA43	B1	Barcelos, Portugal	2012	ST61	n.s.
SA44	B1	Barcelos, Portugal	2012	ST61	n.s.
SA45	B1	Barcelos, Portugal	2012	ST61	n.s.
SA46	B1	Barcelos, Portugal	2012	ST61	n.s.
SA47	B1	Barcelos, Portugal	2012	ST61	n.s.
SA48	B1	Barcelos, Portugal	2012	ST61	ERS1075233
SA49	B1	Barcelos, Portugal	2012	ST61	n.s.
SA50	B1	Barcelos, Portugal	2012	ST61	n.s.
SA51	B1	Barcelos, Portugal	2012	ST61	n.s.
SA52	B1	Barcelos, Portugal	2012	ST61	ERS1075237
SA55	B1	Barcelos, Portugal	2012	ST61	n.s.
SA56	B1	Barcelos, Portugal	2012	ST61	n.s.
SA57	B1	Barcelos, Portugal	2012	ST61	n.s.
SA58	B1	Barcelos, Portugal	2012	ST61	n.s.
SA59	B1	Barcelos, Portugal	2012	ST61	n.s.
SA60	B1	Barcelos, Portugal	2012	ST61	n.s.

SA62	B1	Barcelos, Portugal	2012	ST61	n.s.
SA65	B1	Barcelos, Portugal	2012	ST61	n.s.
SA66	B1	Barcelos, Portugal	2012	ST61	n.s.
SA67	B1	Barcelos, Portugal	2012	ST61	ERS1075247
SA69	B1	Barcelos, Portugal	2012	ST61	n.s.
SA74	B1	Barcelos, Portugal	2012	ST61	n.s.
SA76	B2	Barcelos, Portugal	2012	ST61	ERS1075250
SA77	B2	Barcelos, Portugal	2012	ST61	ERS1075251
SA78	B2	Barcelos, Portugal	2012	ST61	ERS1075252
SA79	B2	Barcelos, Portugal	2012	ST61	ERS1075253
SA80	B2	Barcelos, Portugal	2012	ST61	ERS1075254
SA81	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075255
SA82	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075256
SA83	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075257
SA84	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075258
SA85	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075259
SA86	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075260
SA87	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075261
SA88	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075262
SA89	PV1	Póvoa de Varzim, Portugal	2012	ST61	ERS1075263
SA90	B1	Barcelos, Portugal	2013	ST61	n.s.
SA91	P	Penafiel, Portugal	2013	ST61	ERS1075265
SA92	P	Penafiel, Portugal	2013	ST61	ERS1075266
SA93	B3	Barcelos, Portugal	2013	ST61	ERS1075267
SA94	B3	Barcelos, Portugal	2013	ST61	ERS1075268
SA95	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075269
SA96	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075270
SA97	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075271
SA98	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075272
SA99	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075273
SA100	PV3	Póvoa de Varzim, Portugal	2013	ST61	ERS1075274
SA101	PV3	Póvoa de Varzim, Portugal	2013	ST61	ERS1075275
SA102	PV3	Póvoa de Varzim, Portugal	2013	ST61	ERS1075276
SA103	PV3	Póvoa de Varzim, Portugal	2013	ST61	ERS1075277
SA104	B1	Barcelos, Portugal	2013	ST61	n.s.
SA105	B4	Barcelos, Portugal	2013	ST61	n.s.
SA106	B4	Barcelos, Portugal	2013	ST61	n.s.
SA107	B4	Barcelos, Portugal	2013	ST61	n.s.
SA108	B4	Barcelos, Portugal	2013	ST61	n.s.
SA109	B4	Barcelos, Portugal	2013	ST61	ERS1075283
SA110	B4	Barcelos, Portugal	2013	ST61	n.s.
SA111	B4	Barcelos, Portugal	2013	ST61	ERS1075285
SA112	B4	Barcelos, Portugal	2013	ST61	n.s.
SA113	B4	Barcelos, Portugal	2013	ST61	n.s.
SA114	B4	Barcelos, Portugal	2013	ST61	n.s.
SA115	B4	Barcelos, Portugal	2013	ST61	ERS1075289
SA116	B4	Barcelos, Portugal	2013	ST61	ERS1075290

SA117	B4	Barcelos, Portugal	2013	ST61	n.s.
SA118	B4	Barcelos, Portugal	2013	ST61	ERS1075292
SA119	B1	Barcelos, Portugal	2013	ST61	n.s.
SA120	B1	Barcelos, Portugal	2013	ST61	ERS1075294
SA121	B1	Barcelos, Portugal	2013	ST61	n.s.
SA122	B1	Barcelos, Portugal	2013	ST61	n.s.
SA154	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075297
SA185	B1	Barcelos, Portugal	2013	ST61	n.s.
SA186	B1	Barcelos, Portugal	2013	ST61	n.s.
SA194	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075300
SA195	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075301
SA196	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075302
SA197	PV2	Póvoa de Varzim, Portugal	2013	ST61	ERS1075303
SA201	PV2	Póvoa de Varzim, Portugal	2013	ST61	n.s.
SA206	B1	Barcelos, Portugal	2013	ST61	n.s.
SA208	B1	Barcelos, Portugal	2013	ST61	ERS1075306
SA209	B1	Barcelos, Portugal	2013	ST61	ERS1075307
SA210	B1	Barcelos, Portugal	2013	ST61	ERS1075308
SA213	B1	Barcelos, Portugal	2013	ST61	n.s.
SA219	B1	Barcelos, Portugal	2013	ST61	n.s.
SA222	B1	Barcelos, Portugal	2013	ST61	n.s.
SA223	B1	Barcelos, Portugal	2013	ST61	ERS1075312
SA224	B1	Barcelos, Portugal	2013	ST61	n.s.
SA225	B1	Barcelos, Portugal	2013	ST61	n.s.
SA226	B1	Barcelos, Portugal	2013	ST61	n.s.
SA228	B1	Barcelos, Portugal	2013	ST61	n.s.
SA230	B1	Barcelos, Portugal	2013	ST61	n.s.
SA232	B1	Barcelos, Portugal	2013	ST61	ERS1075318
SA233	B1	Barcelos, Portugal	2013	ST61	ERS1075319
SA235	B1	Barcelos, Portugal	2013	ST61	n.s.
SA237	B1	Barcelos, Portugal	2013	ST61	n.s.
SA240	B1	Barcelos, Portugal	2013	ST61	n.s.
SA242	B1	Barcelos, Portugal	2013	ST61	n.s.
SA243	B1	Barcelos, Portugal	2013	ST61	n.s.
SA244	B1	Barcelos, Portugal	2013	ST61	n.s.
SA245	B1	Barcelos, Portugal	2013	ST61	n.s.
SA247	B1	Barcelos, Portugal	2013	ST61	n.s.
SA248	B1	Barcelos, Portugal	2013	ST61	n.s.
SA249	B1	Barcelos, Portugal	2013	ST61	n.s.
SA250	B1	Barcelos, Portugal	2013	ST61	n.s.
SA251	B1	Barcelos, Portugal	2013	ST61	n.s.
SA253	PV2	Póvoa de Varzim, Portugal	2014	ST61	n.s.
SA254	PV2	Póvoa de Varzim, Portugal	2014	ST61	n.s.
SA255	PV2	Póvoa de Varzim, Portugal	2014	ST61	n.s.
SA257	PV2	Póvoa de Varzim, Portugal	2014	ST61	n.s.
SA258	PV2	Póvoa de Varzim, Portugal	2014	ST61	n.s.
SA262	B1	Barcelos, Portugal	2014	ST61	n.s.

SA296	B1	Barcelos, Portugal	2014	ST61	ERS1075338
SA323	PV2	Póvoa de Varzim, Portugal	2014	ST61	ERS1075339
SA324	PV2	Póvoa de Varzim, Portugal	2014	ST61	ERS1075340
SA325	PV2	Póvoa de Varzim, Portugal	2014	ST61	ERS1075341
SA368	B1	Barcelos, Portugal	2014	ST61	ERS1075342
SA408	B1	Barcelos, Portugal	2014	ST61	n.s.
VSA65	W	Centre-West, Portugal	2012	ST61	ERS1075354
VSA66	M	Portalegre, Portugal	2012	ST61	ERS1075355
VSA68	A	Lisbon, Portugal	2012	ST554	ERS1075356
VSA69	R	Setúbal, Portugal	2012	ST554	ERS1075357
VSA70	R	Setúbal, Portugal	2013	ST554	ERS1075358
VSA71	R	Setúbal, Portugal	2012	ST554	ERS1075359
VSA72	R	Setúbal, Portugal	2012	ST554	ERS1075360
VSA73	R	Setúbal, Portugal	2013	ST554	ERS1075361
VSA74	R	Setúbal, Portugal	2012	ST554	ERS1075362
VSA75	R	Setúbal, Portugal	2012	ST554	ERS1075363
VSA77	R	Setúbal, Portugal	2012	ST554	ERS1075364
VSA78	R	Setúbal, Portugal	2012	ST554	ERS1075365
VSA79	R	Setúbal, Portugal	2012	ST554	ERS1075366
VSA80	R	Setúbal, Portugal	2012	ST554	ERS1075367
VSA81	R	Setúbal, Portugal	2012	ST554	ERS1075368
VSA83	R	Setúbal, Portugal	2012	ST554	ERS1075369
VSA84	A	Lisbon, Portugal	2012	ST554	ERS1075370
VSA85	R	Setúbal, Portugal	2012	ST554	ERS1075371
VSA86	R	Setúbal, Portugal	2011	ST554	ERS1075372
VSA87	R	Setúbal, Portugal	2012	ST554	ERS1075373
VSA88	W	Centre-West, Portugal	2012	ST61	ERS1075374
VSA89	W	Centre-West, Portugal	2012	ST61	ERS1075375
VSA90	W	Centre-West, Portugal	2012	ST554	ERS1075376
VSA91	R	Setúbal, Portugal	2013	ST61	ERS1075377
VSA92	W	Centre-West, Portugal	2012	ST554	ERS1075378
VSA93	R	Setúbal, Portugal	2012	ST61	ERS1075379
VSA94	W	Centre-West, Portugal	2012	ST61	ERS1075380
VSA95	W	Centre-West, Portugal	2012	ST61	ERS1075381
VSA96	W	Centre-West, Portugal	2012	ST61	ERS1075382
VSA97	R	Setúbal, Portugal	2013	ST554	ERS1075383
VSA98	R	Setúbal, Portugal	2013	ST61	ERS1075384
VSA99	R	Setúbal, Portugal	2013	ST554	ERS1075385
VSA100	R	Setúbal, Portugal	2012	ST554	ERS1075386
VSA101	R	Setúbal, Portugal	2012	ST554	ERS1075387
VSA102	R	Setúbal, Portugal	2013	ST554	ERS1075388
VSA103	R	Setúbal, Portugal	2012	ST554	ERS1075389
VSA104	R	Setúbal, Portugal	2013	ST554	ERS1075390

VSA105	R	Setúbal, Portugal	2012	ST554	ERS1075391
VSA106	R	Setúbal, Portugal	2012	ST554	ERS1075392
VSA107	R	Setúbal, Portugal	2012	ST554	ERS1075393
VSA108	R	Setúbal, Portugal	2012	ST554	ERS1075394
VSA109	R	Setúbal, Portugal	2012	ST554	ERS1075395
VSA110	R	Setúbal, Portugal	2012	ST554	ERS1075396
VSA111	R	Setúbal, Portugal	2013	ST554	ERS1075397
VSA112	R	Setúbal, Portugal	2012	ST554	ERS1075398
VSA113	R	Setúbal, Portugal	2013	ST554	ERS1075399
VSA114	R	Setúbal, Portugal	2013	ST554	ERS1075400
VSA115	A	Lisbon, Portugal	2012	ST554	ERS1075401
VSA116	W	Centre-West, Portugal	2012	ST61	ERS1075402
VSA117	A	Lisbon, Portugal	2012	ST554	ERS1075403
VSA118	R	Setúbal, Portugal	2012	ST554	ERS1075404
VSA119	W	Centre-West, Portugal	2013	ST61	ERS1075405
VSA121	M	Portalegre, Portugal	2011	ST61	ERS1075406
VSA76	Z	Setúbal, Portugal	2012	ST2	ERS1075408

[a]Farm location is depicted in Figure 2.
[b]n.s. = not sequenced.

Portugal (2002-2003)

Strain	Farm	Location[a]	Year	MLST	Accession ID
VSA3	J	Lisbon, Portugal	2002	ST554	ERS1075344
VSA6	C	Setúbal, Portugal	2002	ST61	ERS1075345
VSA10	A	Centre-West, Portugal	2002	ST2	ERS1075407
VSA15	G	Lisbon, Portugal	2003	ST2	n.s.[b]
VSA16	C	Setúbal, Portugal	2003	ST61	ERS1075346
VSA17	A	Lisbon, Portugal	2002	ST61	ERS1075347
VSA22	C	Setúbal, Portugal	2003	ST61	ERS1075348
VSA26	I	Lisbon, Portugal	2003	ST61	ERS1075349
VSA27	C	Setúbal, Portugal	2003	ST61	n.s.
VSA29	C	Setúbal, Portugal	2002	ST61	ERS1075350
VSA32	J	Lisbon, Portugal	2003	ST61	n.s.
VSA36	C	Setúbal, Portugal	2002	ST61	ERS1075351
VSA43	J	Lisbon, Portugal	2002	ST61	n.s.
VSA44	J	Lisbon, Portugal	2003	ST554	ERS1075352
VSA50	J	Lisbon, Portugal	2003	ST554	ERS1075353
VSA53	J	Lisbon, Portugal	2002	ST554	n.s.
VSA54	J	Lisbon, Portugal	2002	ST554	n.s.

[a]Farm location is depicted in Figure 2.
[b]n.s. = not sequenced.

France

Strain	Farm	Location	Year	MLST	Accession ID
9857	FR1	Tarn-et-Garonne, France	1996	ST61	ERS1075198
10058	FR1	Tarn-et-Garonne, France	1997	ST416	ERS1075199
10059	FR1	Tarn-et-Garonne, France	1997	ST416	ERS1075200

Spain

Strain	Farm	Location	Year	MLST	Accession ID
mad3	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075409
mad11	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075410
mad12	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075411
mad14	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075412
mad15	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075413
mad24	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075414
mad25	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075415
mad29	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075416
mad33	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075417
mad34	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075418
mad35	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075419
mad40	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075420
mad42	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075421
mad43	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075422
mad44	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075423
mad53	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075424
mad60	Unknown	Galicia, Spain	2005-2006	ST2	ERS1075425

Table S2. CRISPR spacer profiles of the *S. agalactiae* strains used in this study.

Strain	Farm	Location	MLST	WGS[a]	CRISPR spacers
SA2	VC	Vila do Conde, Portugal	ST61	Y	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA3	VC	Vila do Conde, Portugal	ST61	N	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA4	VC	Vila do Conde, Portugal	ST61	Y	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA6	VC	Vila do Conde, Portugal	ST61	N	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA7	VC	Vila do Conde, Portugal	ST61	N	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA8	VC	Vila do Conde, Portugal	ST61	Y	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA9	VC	Vila do Conde, Portugal	ST61	N	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA10	VC	Vila do Conde, Portugal	ST61	N	6 1317 1312 1325 948 1319 1318 137 138 1313 6 7 147
SA11	T	Trofa, Portugal	ST61	Y	1315 1316 1319 1317 1312 948 190 936 137 138 1313 6 7 147
SA21	Unknown	Barcelos, Portugal	ST61	Y	1315 1316 1317 1318 137 138 1313 6 7 147
SA25	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1316 1319 1318 137 137 147
SA26	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 138 1313 7 147
SA28	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1321 1312 190 1313 6 6 7 147
SA29	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1321 1312 190 1313 6 6 7 147
SA30	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1321 1312 190 1313 6 6 7 147
SA31	Unknown	Póvoa de Varzim, Portugal	ST61	Y	1324 1323 1322 1321 1312 190 1313 6 6 7 147
SA32	Unknown	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA33	Unknown	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA34	Unknown	Barcelos, Portugal	ST61	Y	1312 190 1313 6 6 7 147
SA35	B1	Barcelos, Portugal	ST61	Y	1312 190 1313 6 6 7 147
SA36	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA37	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA38	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA39	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA40	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA41	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA42	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA43	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA44	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA45	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA46	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA47	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA48	B1	Barcelos, Portugal	ST61	Y	1312 190 1313 6 7 147
SA49	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA50	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA51	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA52	B1	Barcelos, Portugal	ST61	Y	1312 190 1313 6 6 7 147
SA55	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA56	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA57	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA58	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA59	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA60	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA62	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA65	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA66	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA67	B1	Barcelos, Portugal	ST61	Y	1312 190 1313 6 6 7 147
SA69	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA74	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA76	B2	Barcelos, Portugal	ST61	Y	1315 1318 137 138 1313 6 7 147
SA77	B2	Barcelos, Portugal	ST61	Y	1315 137 138 1313 6 7 147
SA78	B2	Barcelos, Portugal	ST61	Y	1315 1318 137 138 1313 6 7 147
SA79	B2	Barcelos, Portugal	ST61	Y	1315 1318 137 138 1313 6 7 147
SA80	B2	Barcelos, Portugal	ST61	Y	1315 1318 137 138 1313 6 7 147
SA81	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 6 7 147
SA82	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 138 1313 6 7 147
SA83	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 1313 6 7 147
SA84	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 6 7 147
SA85	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 138 147
SA86	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 137 138 1312 1325 1318 137 138 1313 7 147
SA87	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 138 1313 6 7 147
SA88	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 138 1313 6 7 147
SA89	PV1	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1318 1312 1325 948 190 936 138 1313 6 7 147
SA90	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA91	P	Penafiel, Portugal	ST61	Y	1317 1318 1312 1325 936 1316 1319 138 1313 1325 7 147
SA92	P	Penafiel, Portugal	ST61	Y	1317 1318 1312 1325 936 1316 1319 138 1313 1325 7 147
SA93	B3	Barcelos, Portugal	ST61	Y	1315 1316 1317 1318 137 138 1313 6 7 147
SA94	B3	Barcelos, Portugal	ST61	Y	1315 1316 1317 1318 137 138 1313 6 7 147
SA95	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA96	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA97	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 147
SA98	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA99	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA100	PV3	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 138 1313 7 147
SA101	PV3	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 1316 1319 1318 137 137 138 1313 7 147
SA102	PV3	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 137 138 1313 7 147
SA103	PV3	Póvoa de Varzim, Portugal	ST61	Y	1315 1316 1319 1317 137 138 1313 7 147
SA104	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA105	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA106	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA107	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA108	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA109	B4	Barcelos, Portugal	ST61	Y	1314 1315 1316 1317 1318 137 138 1313 6 7 147
SA110	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA111	B4	Barcelos, Portugal	ST61	Y	1312 190 1313 6 7 147
SA112	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA113	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA114	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA115	B4	Barcelos, Portugal	ST61	Y	1314 1315 1316 1317 1318 137 138 1313 6 7 147
SA116	B4	Barcelos, Portugal	ST61	Y	1312 190 1313 6 7 147
SA117	B4	Barcelos, Portugal	ST61	N	1312 190 1313 6 7 147
SA118	B4	Barcelos, Portugal	ST61	Y	1312 190 1313 6 7 147
SA119	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA120	B1	Barcelos, Portugal	ST61	Y	1312 190 1313 6 6 7 147
SA121	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA122	B1	Barcelos, Portugal	ST61	N	1312 190 1313 6 6 7 147
SA154	PV2	Póvoa de Varzim, Portugal	ST61	Y	1312 190 6 6 7 147
SA185	B1	Barcelos, Portugal	ST61	N	1320 1312 190 6 6 7 147
SA186	B1	Barcelos, Portugal	ST61	N	1312 190 6 6 7 147
SA194	PV2	Póvoa de Varzim, Portugal	ST61	Y	1312 190 6 6 7 147
SA195	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA196	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA197	PV2	Póvoa de Varzim, Portugal	ST61	Y	1320 1312 190 6 6 7 147
SA201	PV2	Póvoa de Varzim, Portugal	ST61	N	1320 1312 190 6 6 7 147

SA206	B1	Barcelos, Portugal	ST61	N						1312				190												6	6	7	147
SA208	B1	Barcelos, Portugal	ST61	Y						1312				190												6	6	7	147
SA209	B1	Barcelos, Portugal	ST61	Y						1312				190												6	6	7	147
SA210	B1	Barcelos, Portugal	ST61	Y						1312				190										1313		6	6	7	147
SA213	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA219	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA222	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA223	B1	Barcelos, Portugal	ST61	Y						1312				190										1313		6	6	7	147
SA224	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA225	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA226	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA228	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA230	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA232	B1	Barcelos, Portugal	ST61	Y						1312				190										1313		6	6	7	147
SA233	B1	Barcelos, Portugal	ST61	Y						1312				190										1313		6	6	7	147
SA235	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA237	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA240	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA242	B1	Barcelos, Portugal	ST61	N						1312				190														7	147
SA243	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA244	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA245	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA247	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA248	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA249	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA250	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA251	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
SA253	PV2	Póvoa de Varzim, Portugal	ST61	N						1312				190												6	6	7	147
SA254	PV2	Póvoa de Varzim, Portugal	ST61	N					1320	1312				190												6	6	7	147
SA255	PV2	Póvoa de Varzim, Portugal	ST61	N					1320	1312				190												6	6	7	147
SA257	PV2	Póvoa de Varzim, Portugal	ST61	N					1320	1312				190												6	6	7	147
SA258	PV2	Póvoa de Varzim, Portugal	ST61	N					1320	1312				190												6	6	7	147
SA262	B1	Barcelos, Portugal	ST61	N					1320	1312				190										1313		6	6	7	147
SA296	B1	Barcelos, Portugal	ST61	Y						1312				190										1313		6	6	7	147
SA323	PV2	Póvoa de Varzim, Portugal	ST61	Y						1312				190												6	6	7	147
SA324	PV2	Póvoa de Varzim, Portugal	ST61	Y					1320	1312				190												6	6	7	147
SA325	PV2	Póvoa de Varzim, Portugal	ST61	Y					1320	1312				190												6	6	7	147
SA368	B1	Barcelos, Portugal	ST61	Y					1320	1312				190										1313		6	6	7	147
SA408	B1	Barcelos, Portugal	ST61	N						1312				190										1313		6	6	7	147
VSA3	J	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA6	C	Setúbal, Portugal	ST61	Y	1315				1318	1312	1325		948		936	1316	1319	1317	1316			137	138	1313			6	7	147
VSA16	C	Setúbal, Portugal	ST61	Y	1315				1318	1312	1325		948		936	1316	1319	1317	1318		6	137	138	1313			6	7	147
VSA17	A	Lisbon, Portugal	ST61	Y	1315					1312	1325		948		936	1316	1319	1317	1318					1313			6		147
VSA22	C	Setúbal, Portugal	ST61	Y	1315				1318	1312	1325		948		936	1316	1319	1317				137	138	1313			6	7	147
VSA26	I	Lisbon, Portugal	ST61	Y									948	190	936						137	138	138	1313	1313			7	147
VSA27	C	Setúbal, Portugal	ST61	N	1315																6	137	138	1313			6	7	147
VSA29	C	Setúbal, Portugal	ST61	Y	1315				1318	1312	1325		948		936	1316	1319	1317	1316		6	137	138	1313			6	7	147
VSA32	J	Lisbon, Portugal	ST61	N	1315	1316										1319	1318	1317					138				6	7	147
VSA36	C	Setúbal, Portugal	ST61	Y	1315				1318	1312	1325		948		936	1316	1319	1317				137	138	1313			6	7	147
VSA43	J	Lisbon, Portugal	ST61	N	1315		1316	1319	1318	1312			948	190	936	1316	1319	1317	1318			137	138	1313			6	7	147
VSA44	J	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA50	J	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA53	J	Lisbon, Portugal	ST554	N	1315		1316	1319	1318	1346	1347	1348												1313			6	7	147
VSA54	J	Lisbon, Portugal	ST554	N	1315	1316		1319	1318	1312			948	190					1318			137	138	1313			6	7	147
VSA65	W	Centre-West, Portugal	ST61	Y						1312	1325		948	190			1319					137	138	1313				7	147
VSA66	M	Portalegre, Portugal	ST61	Y											936	1316	1319							1313			6	7	147
VSA68	A	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA69	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA70	R	Setúbal, Portugal	ST554	Y	1315	1316				1312			948	190	936							137	138	1313			6	7	147
VSA71	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA72	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA73	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA74	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA75	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA77	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA78	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA79	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA80	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA81	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA83	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA84	A	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA85	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA86	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA87	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA88	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA89	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA90	W	Centre-West, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA91	R	Setúbal, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA92	W	Centre-West, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA93	R	Setúbal, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA94	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA95	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA96	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA97	R	Setúbal, Portugal	ST554	Y	1315	1316				1312			948	190	936							137	138	1313			6	7	147
VSA98	R	Setúbal, Portugal	ST61	Y	1315					1312	1325		948				1319		1318			137	138	1313				7	147
VSA99	R	Setúbal, Portugal	ST554	Y	1315	1316								190	936							137	138	1313			6	7	147
VSA100	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA101	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA102	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA103	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA104	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA105	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA106	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA107	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA108	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA109	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA110	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA111	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA112	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA113	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA114	R	Setúbal, Portugal	ST554	Y	1315								948	190	936							137	138	1313			6	7	147
VSA115	A	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA116	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA117	A	Lisbon, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA118	R	Setúbal, Portugal	ST554	Y	1315					1312			948	190	936							137	138	1313			6	7	147
VSA119	W	Centre-West, Portugal	ST61	Y						1312	1325		948				1319		1318			137	138	1313				7	147
VSA121	M	Portalegre, Portugal	ST61	Y																				1313			6	7	147
9857	FR1	Tarn-et-Garonne, France	ST61	Y						1312					1327	1328	948	949		936			138	1313	6	7	6	7	147

10058	FR1	Tarn-et-Garonne, France	ST416	Y	1343	1344	1345	126	1021	190							936	887		138	1313			6	7	147
10059	FR1	Tarn-et-Garonne, France	ST416	Y	1343	1344	1345	126	1021	190							936	887		138	1313			6	7	147
VSA10	A	Centre-West, Portugal	ST2	Y													511	512		42	43	44	45	46	6	8
VSA15	G	Lisbon, Portugal	ST2	N													511	512	513	42		44	45	46	6	8
VSA76	Z	Setúbal, Portugal	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad3	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad11	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45		6	8
mad12	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad14	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad15	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad24	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad25	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad29	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44			6	8
mad33	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad34	Unknown	Galicia, Spain	ST2	Y														512	513	42	43	44	45		6	8
mad35	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad40	Unknown	Galicia, Spain	ST2	Y																	43	44	45		6	8
mad42	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8
mad43	Unknown	Galicia, Spain	ST2	Y																	43	44	45		6	8
mad44	Unknown	Galicia, Spain	ST2	Y																	43	44	45		6	8
mad53	Unknown	Galicia, Spain	ST2	Y																	43	44	45		6	8
mad60	Unknown	Galicia, Spain	ST2	Y													511	512	513	42	43	44	45	46	6	8

[a]Whether the strain was sent for whole-genome sequencing (Y - yes, N - no).

Table S3. Genes that acquired more than three independent mutations.

Locus	Product	N[a]	S[b]	Length (bp)
SA111_00073	phosphoribosylaminoimidazole carboxylase II	2 (0)	1	1074
SA111_00115	DNA-directed RNA polymerase alpha subunit	3 (0)	0	939
SA111_00141	HrcA family transcriptional regulator	3 (0)	0	1083
SA111_00150	Potassium efflux system KefA protein / Small-conductance mechanosensitive channel	2 (0)	1	843
SA111_00183	Amino acid ABC transporter, glutamine-binding protein/permease protein	2 (0)	1	1551
SA111_00195	ABC transporter, substrate-binding protein	3 (0)	1	1656
SA111_00245	putative galactitol operon regulator (Transcriptional antiterminator), BglG family	4 (1)	0	2037
SA111_00295	Osmotically activated L-carnitine/choline ABC transporter, ATP-binding protein OpuCA	4 (0)	0	1146
SA111_00335	Transketolase	2 (0)	2	1986
SA111_00344	penicillin-binding protein	3 (0)	1	2154
SA111_00352	Nicotinate phosphoribosyltransferase	3 (0)	0	1461
SA111_00444	Copper-translocating P-type ATPase	4 (0)	1	2235
SA111_00465	UTP--glucose-1-phosphate uridylyltransferase	2 (0)	1	900
SA111_00563	putative phosphomannomutase	1 (0)	2	1695
SA111_00643	Calcium-transporting ATPase	3 (0)	0	2685
SA111_00651	carboxymethylenebutenolidase-related protein	3 (0)	0	711
SA111_00667	Adenosine deaminase	3 (1)	0	1023
SA111_00747	DNA gyrase subunit B	1 (0)	2	1953
SA111_00762	Beta-lactamase class C and other penicillin binding proteins	3 (0)	1	1152
SA111_00763	Daunorubicin resistance ATP-binding protein drrA	2 (0)	1	993
SA111_00775	cylI protein	1 (0)	2	2196
SA111_00804	Uronate isomerase	3 (0)	1	1401
SA111_00809	Beta-hexosaminidase	3 (0)	0	1791
SA111_00815	Threonyl-tRNA synthetase	2 (0)	1	1944
SA111_00830	Chromosome partition protein smc	3 (0)	1	3540
SA111_00859	haloacid dehalogenase	2 (0)	1	903
SA111_00902	6-phospho-beta-glucosidase	3 (0)	0	1428
SA111_00915	Oligoendopeptidase F	4 (0)	0	1806
SA111_00953	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	3 (0)	0	1260
SA111_00962	glycogen debranching protein	3 (0)	1	2301
SA111_00964	Glucose-1-phosphate adenylyltransferase	5 (1)	0	1140
SA111_00971	ATP synthase subunit alpha	3 (0)	1	1506
SA111_00976	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	3 (0)	0	1272
SA111_00980	acetyltransferase, GNAT family	3 (0)	0	522
SA111_00986	GTPase and tRNA-U34 5-formylation enzyme TrmE	2 (0)	1	1377
SA111_00987	ABC transporter ATP-binding protein	3 (0)	1	1911
SA111_00994	putative amino acid ligase found clustered with an amidotransferase	0 (0)	3	1344
SA111_00997	Phosphoglucosamine mutase	2 (0)	1	1353
SA111_01004	hypothetical protein	4 (0)	1	4113
SA111_01015	Cation-transporting ATPase, E1-E2 family	3 (0)	1	2793
SA111_01017	chloramphenicol acetyltransferase	3 (0)	0	639
SA111_01055	Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]	2 (0)	1	1815
SA111_01064	Nucleoside-binding protein	5 (0)	0	1050
SA111_01217	recombinase	2 (0)	1	1209
SA111_01326	Sialic acid biosynthesis protein NeuD, O-acetyltransferase	2 (0)	1	618
SA111_01387	pullulanase	5 (0)	0	3759
SA111_01388	Histidinol-phosphatase	3 (0)	0	735
SA111_01392	Glycerophosphoryl diester phosphodiesterase	2 (0)	1	1782
SA111_01465	putative ATP-dependent Clp proteinase (ATP-binding subunit)	3 (1)	3	2109
SA111_01487	Two component system histidine kinase	3 (0)	0	1230
SA111_01488	Two component response regulator	3 (0)	1	693
SA111_01499	multidrug ABC transporter ATP-binding protein	3 (1)	0	1740
SA111_01508	PTS system, fructose-specific IIABC components	3 (0)	0	1965
SA111_01566	Hemolysins containing CBS domains	2 (0)	1	1335
SA111_01615	GftB: Glycosyl transferase, family 8	2 (0)	1	1326
SA111_01618	Accessory secretory protein Asp3	2 (0)	1	993
SA111_01633	hypothetical protein	3 (0)	0	1473
SA111_01636	Glutamine ABC transporter, glutamine-binding protein/permease protein	2 (0)	1	2184
SA111_01657	3'-to-5' exoribonuclease RNase R	2 (0)	2	2361
SA111_01676	67kDa Myosin-crossreactive streptococcal antigen	2 (0)	1	1773
SA111_01719	Branched-chain amino acid transport system carrier protein	3 (0)	0	1338
SA111_01723	PTS cellobiose-specific IIC component	4 (0)	0	1218

SA111_01740	ABC transporter permease	2 (0)	2	870
SA111_01751	membrane protein	2 (0)	1	252
SA111_01781	Transmembrane histidine kinase CovS	2 (0)	1	1506
SA111_01790	Duplicated ATPase component MtsB of energizing module of methionine-regulated ECF transporter	3 (0)	1	1677
SA111_01826	Pyruvate,phosphate dikinase	4 (0)	2	2646
SA111_01838	ATP-dependent DNA helicase RecG	2 (0)	1	2016
SA111_01846	PTS system, sucrose-specific IIABC components	5 (0)	0	1920
SA111_01923	Transcriptional regulator, repressor of the glutamine synthetase, MerR family	4 (0)	1	372
SA111_02017	Glutathione biosynthesis bifunctional protein gshF	3 (0)	1	2253
SA111_02110	2',3'-cyclic-nucleotide 2'-phosphodiesterase	3 (0)	2	2403
SA111_02130	PTS system, maltose and glucose-specific IIABC components	2 (0)	1	2184
SA111_02131	Phosphate regulon sensor protein PhoR (SphS)	3 (0)	0	1656
SA111_02134	phosphate transporter ATP-binding protein	3 (1)	1	750
SA111_02142	Transcriptional regulator, MerR family	2 (0)	1	717
SA111_02146	acetyltransferase	2 (0)	1	480
SA111_02162	5-methyltetrahydropteroyltriglutamate-- homocysteine methyltransferase	0 (0)	4	2238
SA111_02180	FIG01117180: hypothetical protein	3 (3)	0	1641
SA111_02200	Ribonucleotide reductase of class III (anaerobic), large subunit	1 (0)	2	2199
SA111_02201	Competence-induced protein Ccs4	3 (0)	1	1542
SA111_02223	LSU ribosomal protein L33p @ LSU ribosomal protein L33p, zinc-independent	0 (0)	3	150
SA111_02245	membrane protein	2 (0)	1	2541
SA111_02249	TetR family transcriptional regulator	5 (0)	0	540
SA111_02255	Phosphoesterase, DHH family protein	1 (0)	3	1983

[a]Number of nonsynonymous substitutions. The number of nonsense mutations among those are indicated in brackets.
[b]Number of synonymous substitutions.

Table S4. BLASTn results of the drug resistance and metabolism genes acquired by the CC61 clone.

Genes	Function	BLASTn similarity (Coverage - Identity)[a]									
		S. agalactiae		***S. dysgalactiae***		***S. macedonicus***		***S. suis***		***S. uberis***	
tetO	Tetracycline resistance	100%	99%	-		-		99%	99%	-	
ermB	Macrolide resistance	98%	99%	-		-		98%	99%	-	
aadE	Streptomycin resistance	100%	99%	-		-		100%	99%	-	
Nisin	Nisin production and resistance	100%	100%	-		-		100%	99%	96%	87%
lnuC	Lincosamide resistance	100%	100%	-		-		82%	99%	-	
Lac.2-1	Lactose metabolism	100%	99%	100%	99%	-		-		100%	99%
Macedocin	Macedocin production and resistance	83%	99%	92%	99%	92%	99%	-		-	

[a]BLASTn search was performed with a 1-kb sequence flanking the target region.

Table S5. Primer pairs used in the RT-qPCRs.

Primer	Sequence	Tm (ºC)	GC%	Length (bp)	Gene
mtsA_F	TCAAGTGTTGACAAGCGTCCT	60	48	112	*mtsA*
mtsA_R	AGCTGTCACCCTTTTGACCT	59	50		
mtsB_F	CTGGAGAATCAGGGATTGGGG	60	57	183	*mtsB*
mtsB_R	GGGACACTTTTTCCCAGTCAG	59	52		

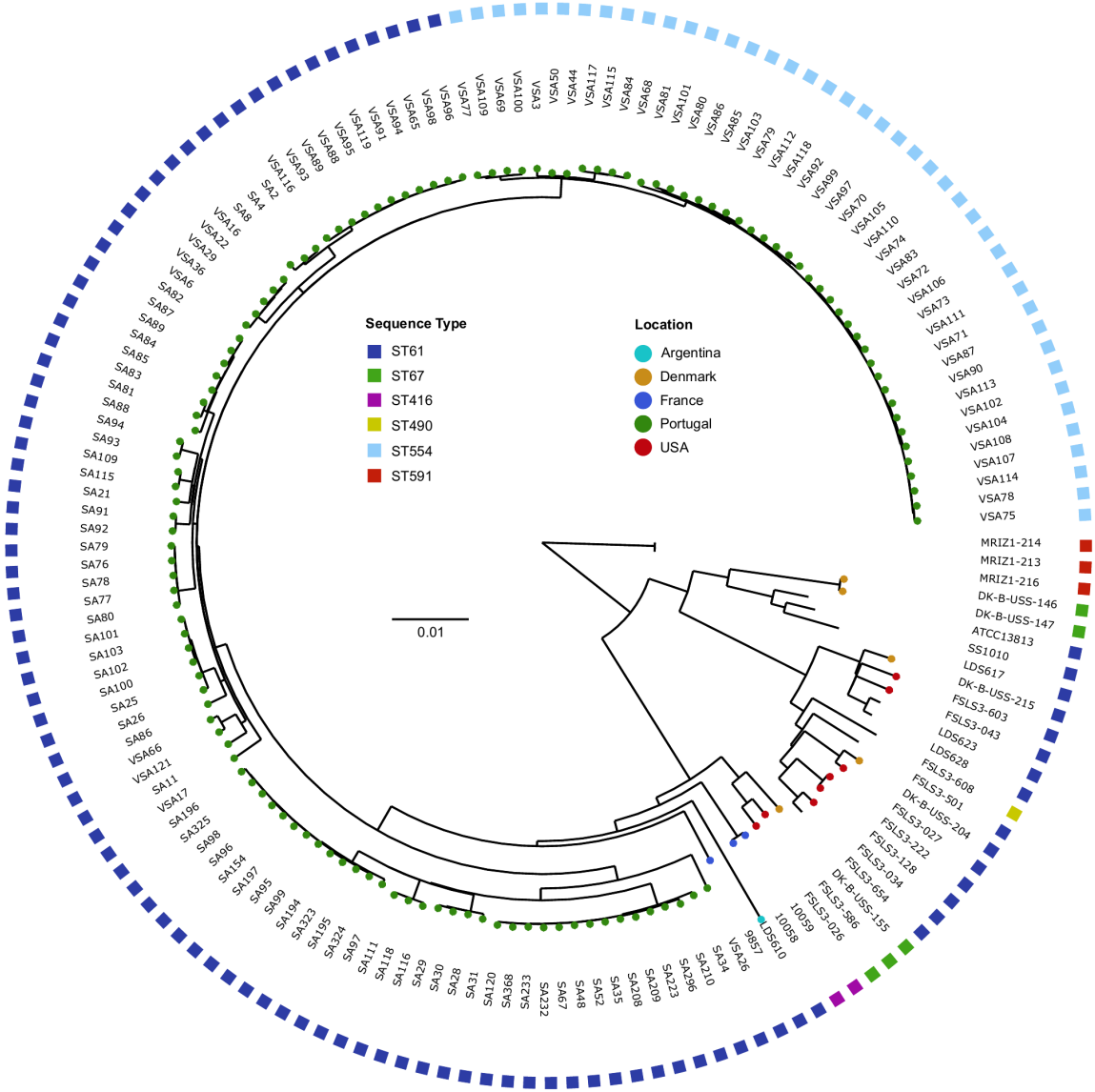


Fig. S1. Core-genome phylogeny of CC61. The ML phylogeny was built using RAxML (Stamatakis, 2014), based on a total of 4524 polymorphic positions. Tree depicts all the publicly available CC61 genomes, as well as the 128 CC61 representative isolates from Portugal analyzed in this work and the three additional isolates sampled from the southwest of France (9857, 10058, 10059). Sequence type (ST) and location of each isolate (if available) are plotted in the tree as depicted in the figure key. All genomes depicted are of bovine origin, except for ATCC13813, which was isolated from the oral cavity of a human host. All tree nodes presented a bootstrap support >75%. Tree was rooted with the CC1 genome of GBS 09mas01883 (not shown).

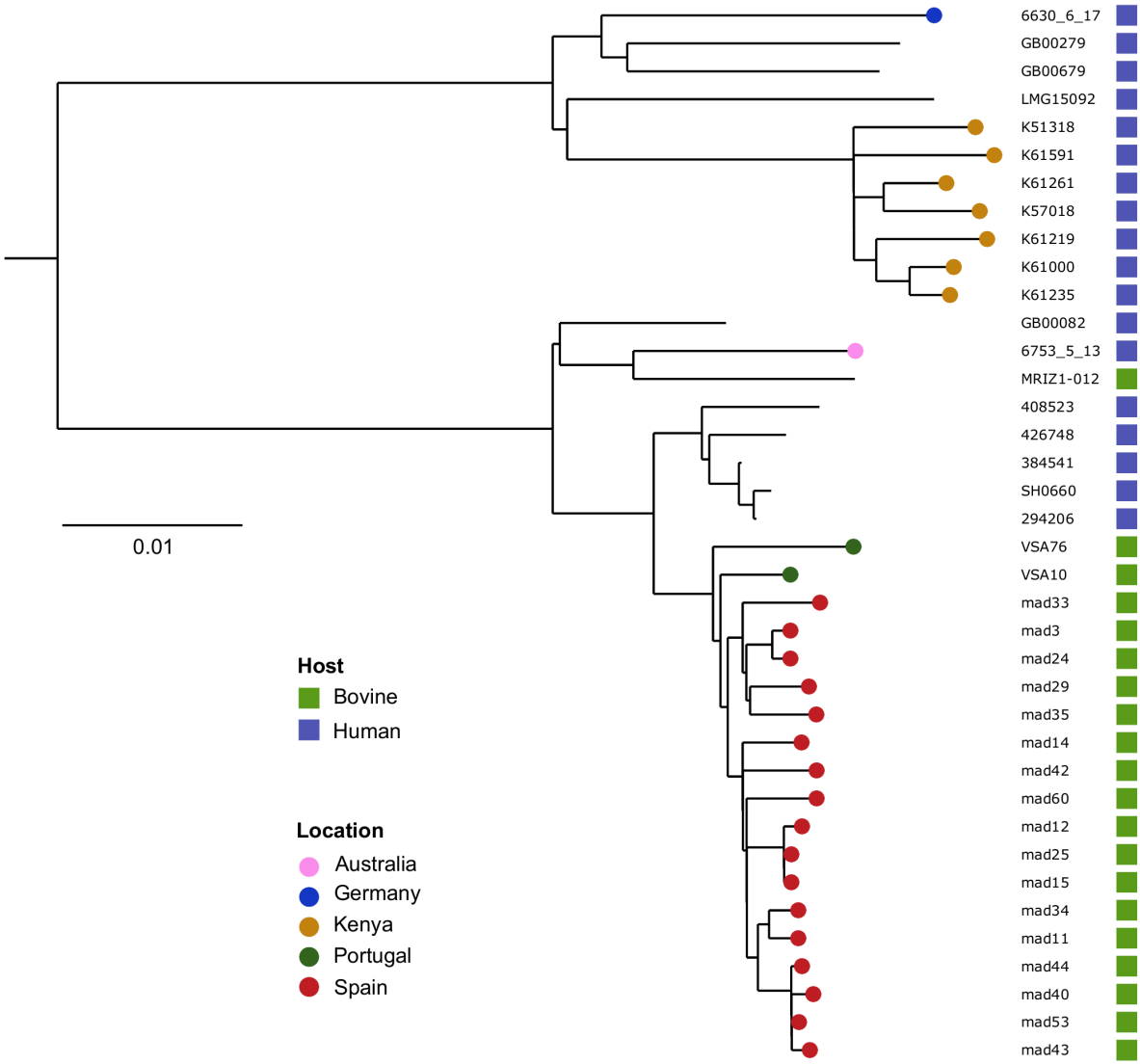


Fig. S2. Core-genome phylogeny of ST2 genomes. ML phylogenetic tree built using RAxML (Stamatakis, 2014) based on a total of 1403 SNPs. Tree depicts all the publicly available ST2 genomes, together with the 19 ST2 isolates from Portugal and Spain analyzed in this work. Host origin and location of each isolate (if available) are plotted in the tree as depicted in the figure key. All tree nodes presented a bootstrap support>75%. Tree was rooted with the CC19 genome of GBS 2603V/R (not shown).

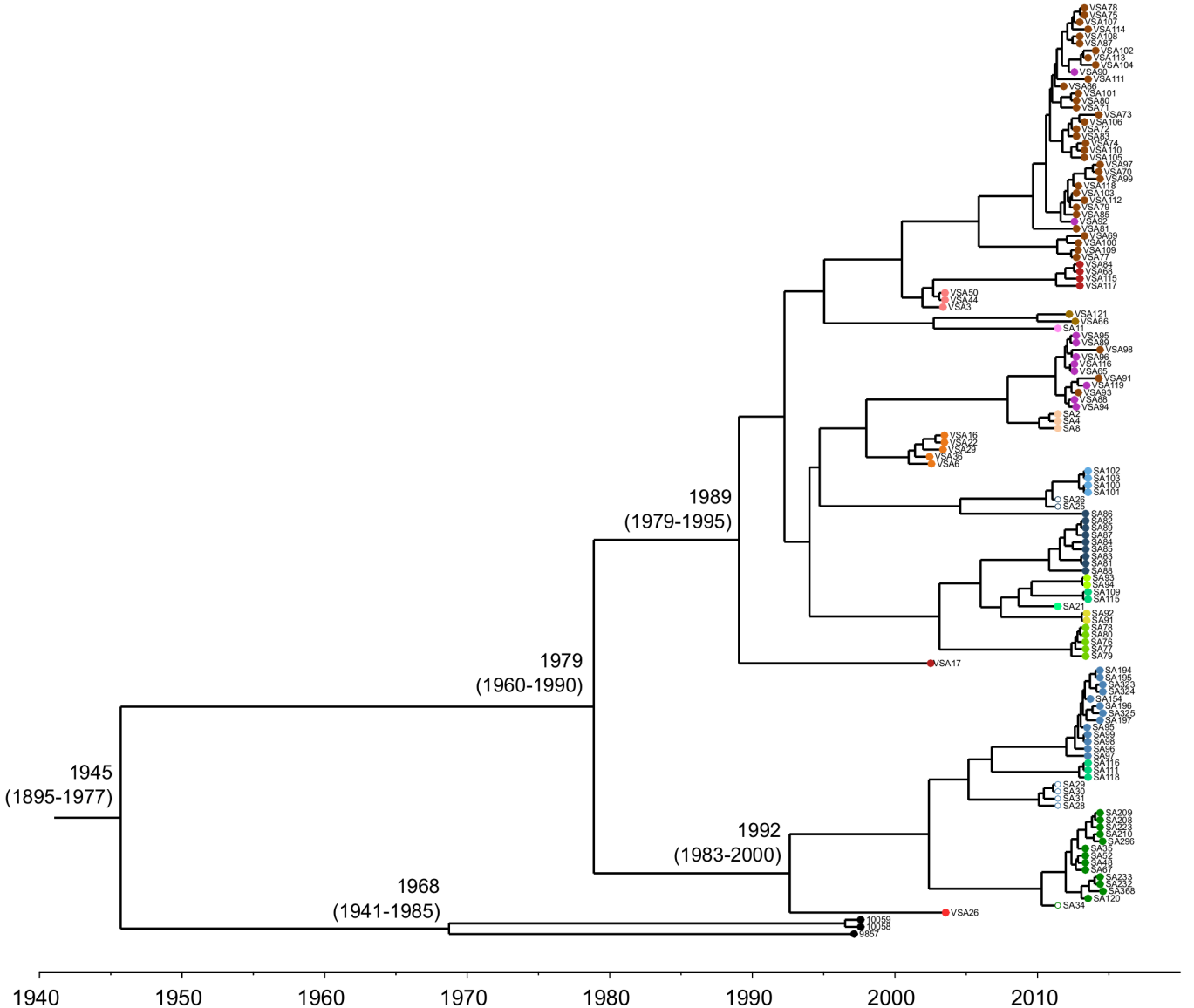


Fig. S3. Timed phylogeny of the CC61 isolates. Bayesian phylogenetic inference of the 128 CC61 isolates sequenced from Portugal and the three additional CC61 isolates from France, performed with BEAST (Drummond et al., 2012). The CC61 isolates are colour-coded as in Fig. 2. Interval dates in brackets represent 95% HPDs of the estimated date of each node.

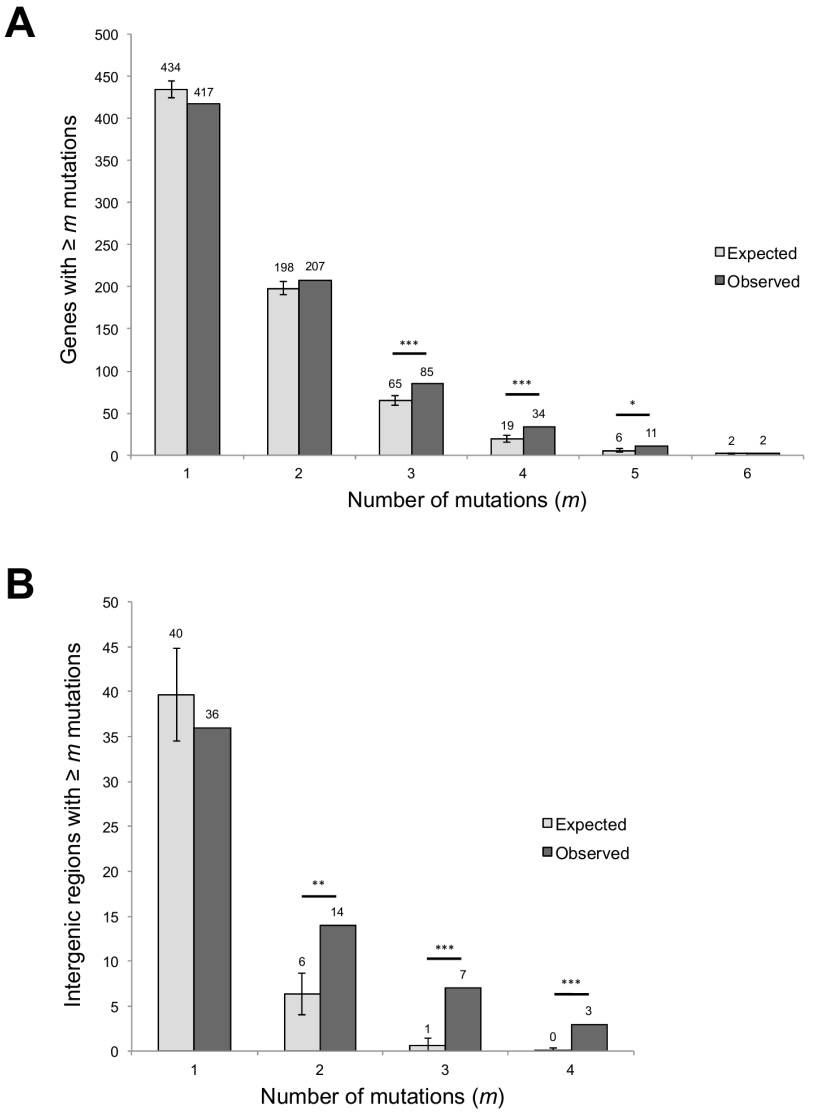


Fig. S4. Parallel evolution in genes and in intergenic regions. Number of genes (a) or intergenic regions (b) with at least *m* mutations, as a function of *m*. Expected values correspond to the average of 1000 random distributions of the number of mutations observed. Error bars represent standard deviation (SD) 1/2. ***P<0.001; **P<0.01; *P<0.05.

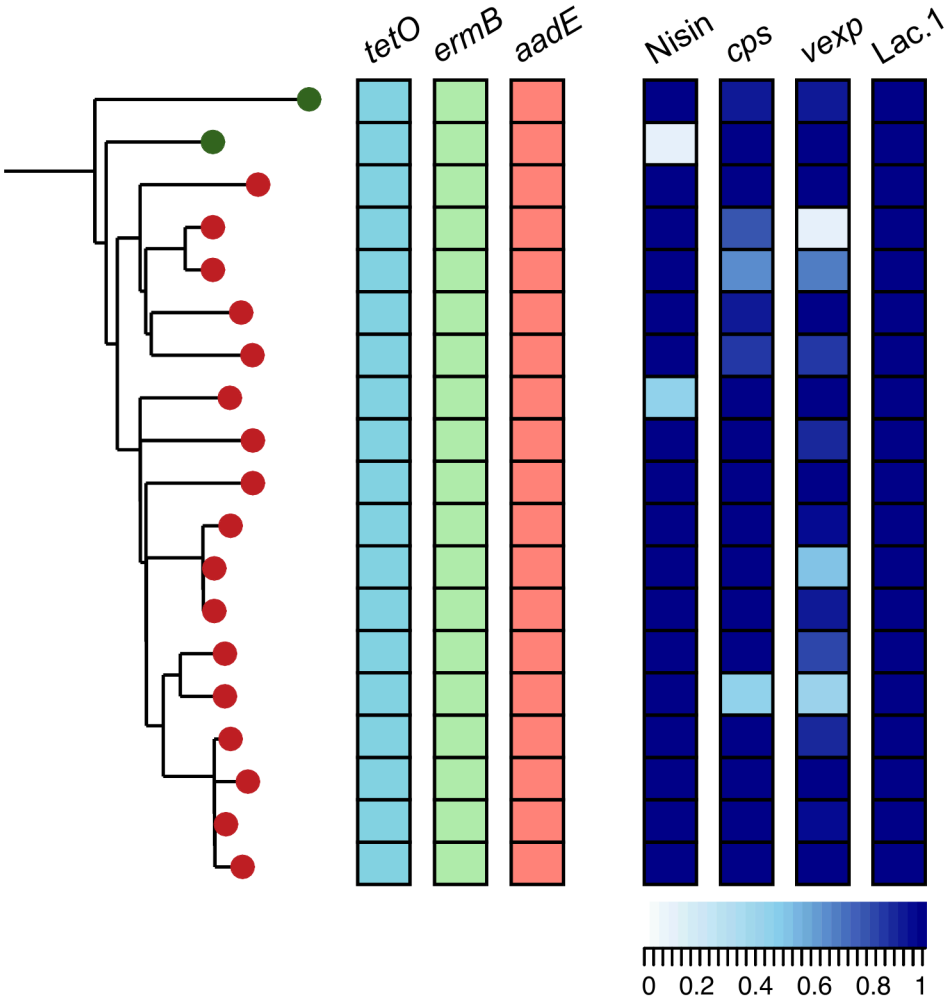


Fig. S5. Analysis of most functionally relevant accessory genes identified in the CC61 population, against the coregenome phylogeny of the 19 ST2 isolates from Portugal and Spain. Isolates are colour-coded as in Supporting Information Fig. S2. White denotes absence, and for *tetO*, *ermB* and *aadE*, different coloured boxes mean they were detected within distinct contiguous sequences. Similarities with the nisin operon, *cps*, *vexp* and Lac.1 loci are depicted by a gradient from white to dark blue corresponding to a blast score ratio ranging from 0 to 1, as indicated in the figure key.

B. Transmission and host-adaptive evolution in humans

Since their emergence in the 1950s, GBS infections represent a prevalent cause of neonatal morbidity and mortality for which preventive therapeutic measures have not been completely effective. In fact, current control strategies have not mitigated the incidence of LOD and its consistent association with CC17. It has also been unclear whether the transmission of GBS from mother to child is accompanied by the emergence of genomic mutations that may increase the propensity to cause disease.

To address this, we analysed the human-derived population of GBS under two evolutionary perspectives. Using a panel of 19 pairs, we followed the transmission of GBS strains from mother to child to look for the selection of clinically associated genomic variants. In a few cases, we found that pathoadaptive mutations are acquired in the newborn after maternal transmission. Notably, mutations within a known virulence regulator and the promoter of the surface protein Rib were shown to alter the expression of virulence-associated genes in human blood. These results suggested that mutations selected in the newborn might contribute to the ability of GBS to infect its host. Extending this work to a larger number of samples will be of critical importance to more confidently understand the transmission dynamics of GBS during the perinatal period.

Complementing this analysis at a broader scale, I investigated the genomic evolution of the hypervirulent and human-specific CC17. Building on the methodology developed for the analysis of the bovine-derived population, I performed an in-depth study of 612 CC17 genomes collected worldwide over 54 years to identify the most important pathways for adaptation and pathogenesis within the human host. Having access to such an extensive dataset allowed us to find evolutionary signals of adaptation among this hypervirulent population. In the end, we concluded that various strategies of adaptation in GBS are employed through modulation of genes involved in metabolism, adhesion, immune evasion and regulation, which might be linked to important phenotypes for survival and virulence in humans.

Publication nº2

JB Journals.ASM.org

Whole-Genome Comparison Uncovers Genomic Mutations between Group B Streptococci Sampled from Infected Newborns and Their Mothers

Alexandre Almeida,[a,b,c] Adrien Villain,[d] Caroline Joubrel,[e,f,g,h,i] Gérald Touak,[e,f] Elisabeth Sauvage,[a,b] Isabelle Rosinski-Chupin,[a,b] Claire Poyart,[e,f,g,h,i] Philippe Glaser[a,b,d]

Institut Pasteur, Unité de Biologie des Bactéries Pathogènes à Gram-Positif, Paris, France[a]; CNRS UMR3525, Paris, France[b]; Université Pierre et Marie Curie, Paris, France[c]; Institut Pasteur, Plateforme de Bio-Analyse Génomique, Paris, France[d]; Service de Bactériologie, Centre National de Référence des Streptocoques, Groupe Hospitalier Paris Centre Cochin-Hôtel Dieu-Broca, Assistance Publique Hôpitaux de Paris, Paris, France[e]; DHU Risques et Grossesse, Assistance Publique Hôpitaux de Paris, Paris, France[f]; INSERM, U1016, Paris, France[g]; CNRS (UMR 8104), Paris, France[h]; Université Paris Descartes, Sorbonne Paris Cité, Paris, France[i]

ABSTRACT

Streptococcus agalactiae **(group B *Streptococcus* or GBS), a commensal of the human gut and genitourinary tract, is a leading cause of neonatal infections, in which vertical transmission from mother to child remains the most frequent route of contamination. Here, we investigated whether the progression of GBS from carriage to disease is associated with genomic adaptation. Whole-genome comparison of 47 GBS samples from 19 mother-child pairs uncovered 21 single nucleotide polymorphisms (SNPs) and seven insertions/deletions. Of the SNPs detected, 16 appear to have been fixed in the population sampled whereas five mutations were found to be polymorphic. In the infant strains, 14 mutations were detected, including two independently fixed variants affecting the *covRS* locus, which is known to encode a major regulatory system of virulence. A one-nucleotide insertion was also identified in the promoter region of the highly immunogenic surface protein Rib gene. Gene expression analysis after incubation in human blood showed that these mutations influenced the expression of virulence-associated genes. Additional identification of three mutated strains in the mothers' milk raised the possibility of the newborns also being a source of contamination for their mothers. Overall, our work showed that GBS strains in carriage and disease scenarios might undergo adaptive changes following colonization. The types and locations of the mutations found, together with the experimental results showing their phenotypic impact, suggest that those in a context of infection were positively selected during the transition of GBS from commensal to pathogen, contributing to an increased capacity to cause disease.**

IMPORTANCE

Group B *Streptococcus* (GBS) is a major pathogen responsible for neonatal infections. Considering that its colonization of healthy adults is mostly asymptomatic, the mechanisms behind its switch from a commensal to an invasive state are largely unknown. In this work, we compared the genomic profile of GBS samples causing infections in newborns with that of the GBS colonizing their mothers. Multiple mutations were detected, namely, within key virulence factors, including the response regulator CovR and surface protein Rib, potentially affecting the pathogenesis of GBS. Their overall impact was supported by differences in the expression of virulence-associated genes in human blood. Our results suggest that during GBS's progression to disease, particular variants are positively selected, contributing to the ability of this bacterium to infect its host.

Streptococcus agalactiae, or group B *Streptococcus* (GBS), is currently regarded as one of the leading causes of neonatal sepsis and meningitis worldwide (1–4). Known as a commensal of the digestive and genitourinary tracts of 10 to 30% of the human population, GBS is also a significant source of disease in immunocompromised adults and in the elderly (4–7). Apart from causing human infections, GBS is an etiological agent of bovine mastitis and invasive disease in fish (8–10). First identified in the 1930s as a cause of neonatal mortality, it became a widespread concern in many developed countries in the second half of the 20th century (11–13). In the United States, throughout the 1990s, 7,600 cases of infection in newborns leading to 310 deaths per year were observed, whereas most recently, the global burden of GBS disease was estimated to be 0.53 per 1,000 live births (4, 14). Furthermore, GBS infections can lead to additional complications. Particularly regarding neonatal meningitis, studies have shown some degree of cognitive impairment in 15 to 20% of affected infants, at least until 5 years after birth (15, 16).

Neonatal infections are classified as either early-onset disease (EOD) when they occur in newborns within the first 6 days of life or late-onset disease (LOD) when they are diagnosed from 7 days

Received 2 June 2015 **Accepted** 5 August 2015

Accepted manuscript posted online 17 August 2015

Citation Almeida A, Villain A, Joubrel C, Touak G, Sauvage E, Rosinski-Chupin I, Poyart C, Glaser P. 2015. Whole-genome comparison uncovers genomic mutations between group B streptococci sampled from infected newborns and their mothers. J Bacteriol 197:3354–3366. doi:10.1128/JB.00429-15.

Editor: V. J. DiRita

Address correspondence to Claire Poyart, claire.poyart@cch.aphp.fr, or Philippe Glaser, pglaser@pasteur.fr.

A.V. and C.J. contributed equally to this work.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.00429-15.

doi:10.1128/JB.00429-15

up to approximately 3 months of age (4). Intrapartum antibioprophylaxis for parturients at risk of GBS infection was able to reduce the incidence of EOD, but it did not affect the rate of LOD (3). In EOD, a GBS strain has colonized the maternal vaginal tract and is subsequently transferred to the baby during or just before birth. The most frequent form of infection is thought to be through aspiration, by the baby, of GBS-contaminated amniotic or vaginal fluid from the mother. This causes an initial dissemination of the bacteria in the respiratory epithelium and in the child's bloodstream, leading to clinical manifestations usually in the form of pneumonia or sepsis (4, 17, 18). Late-onset cases typically result in meningitis, but the mode of transmission remains unclear. LOD may result from community- or hospital-acquired contamination after birth, even though the mother is likely the main source of infection. GBS strains can be grouped into 10 different serotypes according to their polysaccharide capsule (19). Epidemiological data have established a strong link between capsular serotype (CPS) III and a substantial proportion of EOD incidences, as well as the majority of LOD infections (20–23). Within these strains, multilocus sequence typing (MLST) analyses have inferred sequence type 17 (ST17) to be the most frequently associated lineage, which led to its current label as a "hypervirulent lineage" (22, 24).

To investigate the emergence of GBS neonatal infections, population and evolutionary studies have detailed a recent origin of the main human-adapted clonal complexes, with recombination of large chromosomal regions playing a key role in their genomic diversification (25). By Bayesian phylogenetic inference, the widespread use of tetracycline in the 1960s was reported as one of the main driving forces behind the selection of pathogenic clones (26). Nevertheless, not all tetracycline-resistant lineages have disseminated; thus, additional colonization and virulence properties seem to have been responsible for the expansion of certain clones still causing disease today. In this regard, GBS pathogenesis is related to several virulence factors, and their role is to promote the ability of the bacteria to infect and damage the host. Some notable examples include the hypervirulent GBS adhesin (HvgA), the β-hemolysin encoded by the *cyl* operon, the C5a peptidase, and surface proteins of the alpha-like family (17, 18, 27–30). Moreover, the expression of major virulence factors is regulated by the two-component system CovRS (control of virulence regulator and sensor). Inactivation of this system, first described in group A *Streptococcus* (GAS), was proven to enhance the pathogenesis of GAS by derepressing the transcription of virulence-associated genes (31). CovR is also a key regulator of virulence in GBS, depending on the site or mode of infection in the host (32, 33).

With the advent of high-throughput sequencing, novel insights into the *in vivo* evolution of many bacterial pathogens have been obtained. For instance, studies on *Staphylococcus aureus*, *Burkholderia pseudomallei*, and *Pseudomonas aeruginosa* have highlighted the selection of a diverse range of genomic mutations during chronic infection (34–36). Furthermore, whole-genome sequencing was recently used to investigate the within-host diversity and transmission of methicillin-resistant *S. aureus* in a veterinary context (37). Yet, specifically in GBS, the exact extent of genomic identity and virulence potential between strains colonizing women and those acquired by their babies through direct transmission remains to be elucidated. Comparative genomic analyses of GBS strains from carriage and disease would help gain a better understanding of the switch from commensal to pathogen and infer crucial virulence and genetic determinants responsible for neonatal GBS diseases.

In this work, we aimed to determine whether specific mutations are selected throughout the course of disease. We performed whole-genome sequencing of a unique collection of GBS samples from infected neonates and from their mothers in either a carriage or a disease context. Interestingly, the majority of the mutations found appear to have been fixed in their population, revealing a selective signal for nonsynonymous substitutions located within functionally significant genes. Thus, we suggest that the disease-associated variants underwent a positive selection process that might have contributed to the increased virulence of these mutated strains. However, by comparing each mutated region with publicly available GBS genomes as an outgroup, we also identified mutant alleles in samples obtained from carriage. Mutant variants found in GBS-positive milk samples from three mothers are compatible with a bidirectional transmission of GBS in which the infants contaminated their mothers through breastfeeding.

MATERIALS AND METHODS

Bacterial strains and culture conditions. A total of 47 GBS samples from 19 mother-child pairs were used in this study (Table 1; see Table S1 in the supplemental material). Biological samples were collected at 12 different hospitals throughout France by following routine clinical laboratory procedures in which an indeterminate number of GBS colonies were sampled and subsequently transferred to the Centre National de Référence des Streptocoques (CNR-Strep). These samples were cultured either on Todd-Hewitt (TH) agar plates at 37°C or on Columbia agar plates with 5% horse blood at 37°C with 5% CO_2 and stored in 20% glycerol–TH broth at −80°C.

Sequencing, *de novo* assembly, and epidemiological analysis. Chromosomal DNA was extracted with the DNeasy blood and tissue kit (Qiagen) by following the manufacturer's instructions. Genomes were sequenced by the Illumina HiSeq 2000 and MiSeq sequencing platforms with single-read runs of 101 and 150 nucleotides, respectively (see Fig. S1A in the supplemental material). Reads were filtered by the fqCleaner tool according to the following parameters. Each read was trimmed off to remove nonconfident bases at the 5′ and 3′ ends with a quality score below 20. All reads shorter than 30 bp were discarded. Reads with <80% confident bases were excluded, as were artifactual and duplicated reads. Genome sequences were assembled by the Velvet software (38) with an optimized k-mer value, a minimum coverage of 10, and a contig length of at least 200 bp. STs were determined by MLST by extracting the sequences of the seven genes of the GBS MLST system (*adhP*, *pheS*, *atr*, *glnA*, *sdhA*, *glcK*, and *tkt*) (24) and comparing them with the known STs from the GBS MLST web server (http://pubmlst.org/sagalactiae/) through the Center for Genomic Epidemiology web tool (http://www.genomicepidemiology.org/). Capsular types were determined by molecular serotyping (39) and confirmed by BLASTn similarity search of the nucleotide sequences of the 10 *cps* loci of the known GBS serotypes.

Variant detection. Within each pair of samples, several approaches were taken to ensure the most reliable detection of the existing genomic variation within the population and to circumvent the limitations of traditionally used bioinformatic pipelines (see Fig. S1 in the supplemental material). First, for each pair, both the child and mother filtered reads were mapped to the assembled genome of the mother sample by Burrows-Wheeler Aligner (BWA) (40). Variant calling was performed with Genome Analysis ToolKit (41) by taking into account the published recommendations (42, 43). Single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) were prefiltered and selected according to the following parameters adapted from Lieberman et al. (44): at least 15 reads aligning in both directions, at least a 100-bp distance from the contig boundaries of the reference assembly, an average base quality greater than 19, an average mapping quality above 34, and a *P* value above 0.05 with

TABLE 1 *S. agalactiae* samples used in this study

Strain	Pair	Hospital[a]	Location	Yr[b]	Time (days) after birth[c]	Disease	Origin	Serotype	MLST
Mother 1	1	A	Paris	2012	0	NA[d]	Vaginal fluid	III	ST17
Child 1		A	Paris	2012	0	EOD	Blood culture	III	ST17
Mother 2	2	B	Clamart	2012	110	NA	Milk	Ia	ST23
Child 2		B	Clamart	2012	104	LOD	Blood culture	Ia	ST23
Mother 3	3	A	Paris	2008	0	NA	Urine	Ia	ST23
Child 3		A	Paris	2008	1	EOD	Cerebrospinal fluid	Ia	ST23
Mother 4	4	A	Paris	2008	0	NA	Vaginal fluid	III	ST17
Child 4		A	Paris	2008	6	EOD	Blood culture	III	ST17
Mother 5-1	5	A	Paris	2008	4	NA	Vaginal fluid	III	ST17
Mother 5-2		A	Paris	2008	23	NA	Milk	III	ST17
Child 5		A	Paris	2008	22	LOD	Blood culture	III	ST17
Mother 6	6	C	Colombes	2008	1	NA	Urine	Ia	ST24
Child 6-1		C	Colombes	2008	1	EOD	Blood culture	Ia	ST24
Child 6-2		C	Colombes	2008	0	EOD	Gastric fluid	Ia	ST24
Mother 7	7	D	Chartres	2008	29	NA	Blood culture	V	ST1
Child 7		E	Montmorency	2008	18	LOD	Blood culture	III	ST17
Mother 8	8	A	Paris	2008	0	NA	Vaginal fluid	III	ST17
Child 8-1		A	Paris	2008	1	EOD	Cerebrospinal fluid	III	ST17
Child 8-2		A	Paris	2008	1	EOD	Blood culture	III	ST17
Mother 9	9	F	Colombes	2008	0	NA	Amniotic fluid	III	ST23
Child 9		F	Colombes	2008	1	EOD	Blood culture	III	ST23
Mother 10-1	10	A	Paris	2009	16	NA	Milk	Ia	ST23
Mother 10-2		A	Paris	2009	−3	NA	Vaginal fluid	Ia	ST23
Child 10		A	Paris	2009	16	LOD	Blood culture	Ia	ST23
Mother 11	11	A	Paris	2009	−1	NA	Vaginal fluid	III	ST23
Child 11-1		A	Paris	2009	21	LOD	Blood culture	III	ST17
Child 11-2		A	Paris	2009	35	LOD	Blood culture	III	ST17
Mother 12	12	G	Paris	2009	0	NA	Blood culture	III	ST27
Child 12-1		G	Paris	2009	0	EOD	Gastric fluid	III	ST27
Child 12-2		G	Paris	2009	0	EOD	Blood culture	III	ST27
Mother 13	13	H	Amiens	2010	7	NA	Milk	III	ST17
Child 13		H	Amiens	2010	5	EOD	Cerebrospinal fluid	III	ST17
Mother 14	14	I	Montpellier	2010	33	NA	Milk	III	ST17
Child 14		I	Montpellier	2010	29	LOD	Cerebrospinal fluid	III	ST17
Mother 15	15	C	Colombes	2010	−1	NA	Vaginal fluid	V	ST1
Child 15-1		C	Colombes	2010	0	EOD	Gastric fluid	V	ST1
Child 15-2		C	Colombes	2010	0	EOD	Blood culture	V	ST1
Mother 16	16	J	Frejus	2011	0	NA	Placenta	III	ST17
Child 16		J	Frejus	2011	0	EOD	Gastric fluid	III	ST17
Mother 17	17	C	Colombes	2011	0	NA	Placenta	III	ST17
Child 17		C	Colombes	2011	0	EOD	Blood culture	III	ST17
Mother 18	18	K	Le Kremlin Bicêtre	2011	0	NA	Blood culture	Ia[e]	ST24
Child 18		K	Le Kremlin Bicêtre	2011	0	EOD	Blood culture	Ia[e]	ST24

(Continued on following page)

TABLE 1 (Continued)

Strain	Pair	Hospital[a]	Location	Yr[b]	Time (days) after birth[c]	Disease	Origin	Serotype	MLST
Mother 19	19	L	Compiègne	2011	7	NA	Vaginal fluid	III	ST17
Child 19-1		L	Compiègne	2011	5	EOD	Cerebrospinal fluid	III	ST17
Child 19-2		L	Compiègne	2011	5	EOD	Blood culture	III	ST17
Child 19-3		L	Compiègne	2011	6	EOD	Urine	III	ST17

[a] One-letter code representing the hospital at which the sample was collected.
[b] Year of sampling.
[c] Time elapsed after the birth of the child from the corresponding pair.
[d] NA, nonapplicable.
[e] Deletion of the *cpsIad*, *cpsE*, *cpsF*, and *cpsG* genes from the *cps* operon was found.

Fisher's exact test supporting a null hypothesis that the variant frequency is the same for reads in both directions (see Fig. S1B in the supplemental material). After this initial filtering, if a variant was detected in <97% of the reads in the mapping of the child sample or >3% of the reads in the mapping of the mother population, it was suspected of corresponding to a polymorphic variant. Therefore, in these cases, an isolated colony from the putative mixed sample was sequenced and used as an isogenic control. False-positive polymorphisms were inferred as those found to be heterogeneous (with an allele frequency between 3% and 97%) in the mapping of the control sample against its assembled genome. Filtered SNPs were then classified as either fixed or polymorphic according to whether they were present in more or less than 97% of the reads, respectively, in at least one of the samples of each pair (Table 2; see Table S2 in the supplemental material). No frequency estimate was calculated for the indels because of a greater probability of mapping and sequencing errors.

To extend this analysis to the identification of genomic islands or other genetic material acquired by the GBS population in the child, the unmapped reads recovered from each pair were assembled by the Velvet

TABLE 2 Genomic mutations detected in the population of each pair of samples

Sample[a]	MLST	Type	Classification[b]	Locus ID[c]	Function[d]	BLAST[e]
Mother 2	ST23	SNP	Nonsense (Q7X), fixed	*gbs1596*	Intermediary metabolism	Glyoxalase
Mother 2	ST23	SNP	Synonymous, fixed	*gbs0198*	Intermediary metabolism	Polynucleotide phosphorylase, alpha chain
Mother 4	ST17	SNP	Missense (K88E), polymorphic	*SAG0163_02655*	Information pathways	Site-specific recombinase
Mother 9	ST23	Indel	Noncoding[h]	*gbs0476* (3′ UTR)	Unknown	Putative membrane protein
Child 9	ST23	SNP	Missense (R79H), polymorphic	*gbs1259*	Intermediary metabolism	Metallo-beta-lactamase superfamily protein
Mother 10-1	ST23	SNP	Noncoding, fixed	*gbs1986*	Cellular processes	ABC transporter
Child 11-2[f]	ST17	Indel	Frameshift	*gbs1427*	Information pathways	KH domain protein
Child 11-2[f]	ST17	SNP	Missense (A175D), fixed	*gbs1950*	Cellular processes	Phosphate ABC transporter
Mother 12	ST27	SNP	Noncoding, fixed	*gbs1946* (5′ UTR)	Cellular processes	Glucose-specific PTS enzyme IIABC
Child 12[g]	ST27	SNP	Missense (A49V), fixed	*gbs0668*	Intermediary metabolism	D-lactate dehydrogenase
Child 12[g]	ST27	SNP	Missense (G814S), fixed	*gbs1038*	Cellular processes	Hypothetical ABC transporter permease
Child 12-1	ST27	SNP	Noncoding, polymorphic	*gbs1946* (5′ UTR)	Cellular processes	Glucose-specific PTS enzyme IIABC
Child 12-1	ST27	Indel	In-frame deletion[h]	*gbs1377*	Intermediary metabolism	Homocysteine *S*-methyltransferase
Child 12-1	ST27	Indel	Large deletion[h]		Unknown	Phage terminase
Child 12-2	ST27	SNP	Noncoding, fixed	*gbs1672* (5′ UTR)	Information pathways	Two-component response regulator (CovR)
Child 12-2	ST27	SNP	Missense (R186K), fixed	*gbs0231*	Cellular processes	Efflux protein
Mother 14	ST17	SNP	Noncoding, fixed	*gbs0330* (5′ UTR)	Information pathways	Transcriptional regulator (MarR family)
Mother 14	ST17	SNP	Missense (G30V), fixed	*gbs1234*	Cellular processes	NeuD protein
Child 14	ST17	Indel	Duplication[h]	*gbs1958*	Information pathways	Transcriptional regulator (MerR family)
Mother 15	ST1	Indel	Noncoding[h]	*gbs0419* (5′ UTR)	Intermediary metabolism	GDXG lipolytic enzyme family protein
Mother 15	ST1	SNP	Missense (I183T), fixed	*gbs1230*	Intermediary metabolism	Glycerol-3-phosphate acyltransferase (PlsY)
Mother 15	ST1	SNP	Noncoding, fixed		Unknown	*Anaerococcus prevotii* plasmid
Mother 18	ST24	SNP	Synonymous, polymorphic	*gbs0189*	Cellular processes	Trehalose-specific PTS enzyme IIABC
Mother 18	ST24	SNP	Missense (P396S), polymorphic	*gbs1460*	Cellular processes	Polysaccharide biosynthesis protein
Mother 19	ST17	SNP	Missense (P80S), fixed	*gbs2047*	Information pathways	RecA protein
Child 19[g]	ST17	SNP	Nonsense (E96X), fixed	*gbs1159*	Intermediary metabolism	Phosphotransacetylase
Child 19[g]	ST17	SNP	Missense (A115V), fixed	*gbs1672*	Information pathways	Two-component response regulator (CovR)
Child 19[g]	ST17	Indel	Noncoding[h]	*GBSCOH1_0416*	Cellular processes	Rib protein

[a] Sample of the corresponding pair in which the mutant allele was most frequently detected.
[b] Classification of the mutation detected. For nonsynonymous substitutions, the amino acid change is shown in parentheses. A mutation was considered fixed if at least 97% of the filtered reads supported the mutation in at least one of the samples of the pair (see Table S2 in the supplemental material). No frequency estimate was calculated for indels.
[c] Locus ID of the mutated gene: *SAG0163_02655* corresponding to GBS MRI Z1-215 (GenBank accession no. NZ_ANET01000026), *GBSCOH1_0416* corresponding to GBS COH1 (GenBank accession no. HG939456), and the rest corresponding to GBS NEM316 (GenBank accession no. AL732656).
[d] Functional category of each affected gene based on the SagaList database of NEM316.
[e] Annotation based on BLASTn homology search.
[f] Mutations were called by mapping against child 11-1.
[g] Mutations were detected in all child samples of the corresponding pair.
[h] No homology with any annotated GBS genomes in the NCBI database.

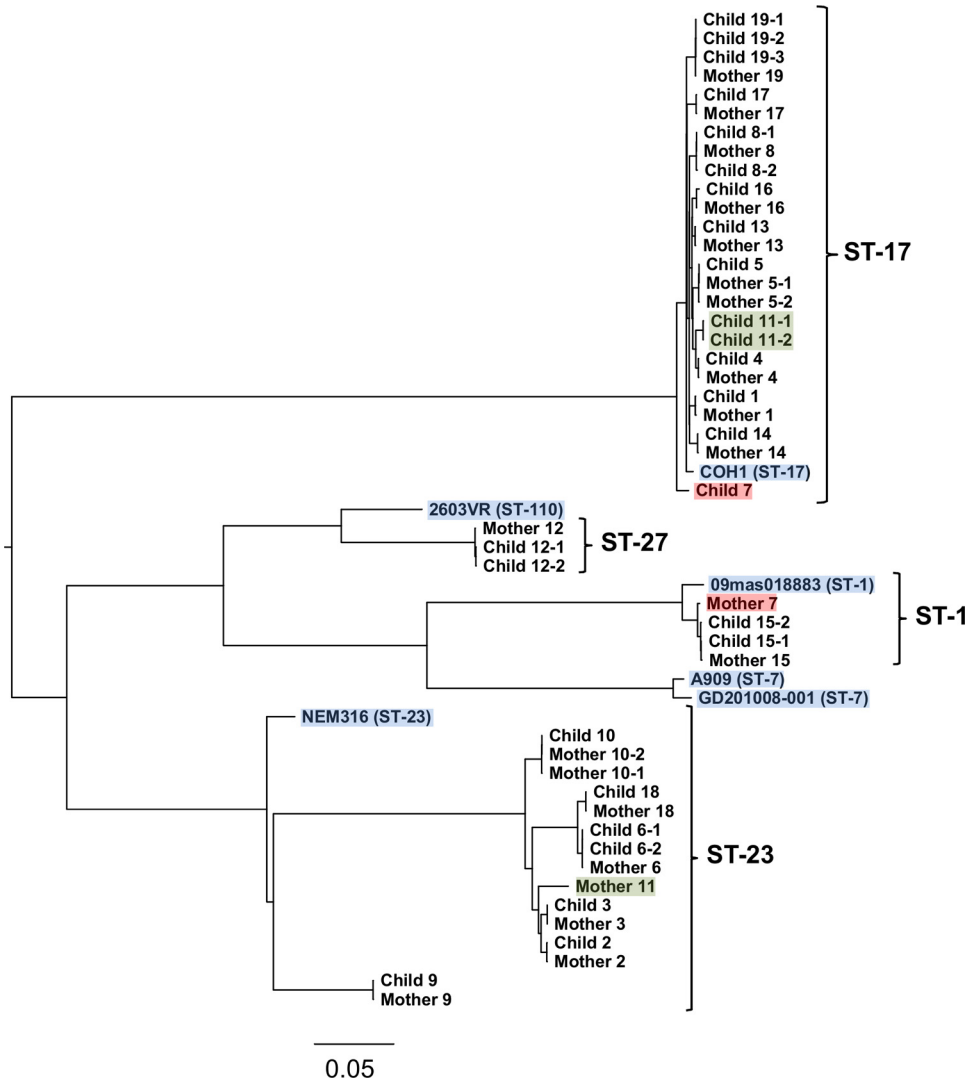


FIG 1 Whole-genome phylogeny of mother-child GBS pairs. Maximum-likelihood phylogenetic tree built with snpTree (49) based on genomic SNPs. The unrooted tree depicts the phylogenetic relationship of each strain in our study with the published complete GBS genomes in the NCBI database (highlighted in blue), i.e., 09mas018883 (GenBank accession no. HF952104), 2603V/R (GenBank accession no. AE009948), A909 (GenBank accession no. CP000114), COH1 (GenBank accession no. HG939456), GD201008-001 (GenBank accession no. CP003810), and NEM316 (GenBank accession no. AL732656). Strain names are as indicated in Table 1 (see also Table S1 in the supplemental material). Strains from pairs 7 and 11 are highlighted in red and green, respectively.

software (see Fig. S1B in the supplemental material). The same k-mer, coverage, and contig length values input for the assembly of the strains from the corresponding pair were used. Contigs assembled from the unmapped reads were blast searched against their own assembled genome to exclude false-positive matches.

The same strategy of read mapping and variant detection was used with the strains of each pair against their closest complete genome within the NCBI database (see Fig. S1C in the supplemental material). Pairs 1, 4, 5, 8, 13, 14, 16, 17, and 19 were mapped against the genome of GBS COH1 (ST17); pairs 2, 3, 6, 9, 10, and 18 were mapped against the genome of GBS NEM316 (ST23); pair 12 was mapped against the genome of GBS 2603V/R (ST110); and pair 15 was mapped against the genome of GBS 09mas018883 (ST1) (Fig. 1).

Moreover, the *breseq* computational pipeline (45) was used to identify the gain or loss of small and large sequence fragments between the mother and child populations sampled. Mapping was performed with the filtered sequencing reads from the child sample against the assembled genome of the mother sample in each pair (see Fig. S1D in the supplemental material). Large deletions or insertions with a coverage value below 10 and within repeated regions or genes with multiple copy numbers were not considered. Each mutated gene was blast searched against the 300 sequenced GBS genomes in the NCBI database available in April 2015. On the basis of the identification of the transcription start sites mapped on the genome of GBS NEM316 (46), the relative position of each SNP located within noncoding regions was ascertained. Functional categories of the mutated genes were inferred on the basis of the SagaList annotation created from the GBS NEM316 sequencing project (47, 48).

Phylogenetic inference. To determine the strain most closely related to each isolate, we built a phylogenetic tree by using snpTree 1.1 (49). With this software, each assembled and complete GBS genome was initially aligned against GBS NEM316 with the Nucmer application from the MUMmer software package (version 3.23), and genomic SNPs were detected with the show-snps option. A filtering process was subsequently applied to discard SNPs <100 bp apart and within 100 bp of the contigs' boundaries. Finally, the resulting genomic SNPs were used to build a maximum-likelihood tree (49) (Fig. 1).

We inferred the wild-type and mutated alleles of each variable position by reconstructing their ancestral sequence with the most closely related publicly available GBS genomes (see Fig. S2 in the supplemental material). Sequences were aligned with ClustalW, and maximum-likelihood phylogenetic trees were built with MEGA 6.06 (50) by using a general time-reversible substitution model with a gamma distribution rate across sites and a proportion of invariant sites. In this work, we assumed that the mutation event occurred from the wild-type allele to the mutant allele, even though we cannot formally exclude the very unlikely possibility of a reversion of the mutated variant.

RNA extraction and RT-qPCR. Bacterial strains were precultured overnight in TH broth at 37°C in standing cultures without agitation. For incubation in human blood, a modified form of the protocol of Oggioni et al. (51) was used. Briefly, bacterial dilutions were first grown to the exponential growth phase (optical density at 600 nm of 0.4 to 0.5) in 10 ml of TH broth. Subsequently, 5 ml of each culture was pelleted and stored at −80°C while the other 5 ml was pelleted, washed twice with phosphate-buffered saline, and resuspended in 5 ml of whole human blood from one donor. Blood cultures were then incubated for 1 h at 37°C under agitation. Afterwards, samples were subjected to low-speed centrifugation (200 × *g* for 5 min) to remove human cells and the supernatant containing the bacteria was then centrifuged at 5,000 × *g* for 10 min. Bacterial pellets obtained were stored at −80°C. Total RNAs were extracted with a phenol/TRIzol-based purification method as previously described (32). RNA quality was assessed by electrophoretic analysis with an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Palo Alto, CA), and reverse transcription (RT) was performed with the Superscript II indirect cDNA kit (Invitrogen, Life Technologies). Primer pairs were designed with Primer-BLAST from the NCBI website to obtain a predicted amplicon size of 100 to 200 bp (see Table S3 in the supplemental material) and quantitative PCR (qPCR) was carried out with SYBR green PCR kits (Applied Biosystems, Life Technologies). Relative gene expression was quantified by a standard-curve-based method in which regression analysis was performed by using serial dilutions of a positive amplification control. Calculated values given in arbitrary units were normalized with the expression of the housekeeping gene *gyrA*. Each assay was performed in triplicate with three independent cultures. A two-tailed *t* test was carried out to determine whether expression differences were statistically significant.

D-Lactate quantification and carbohydrate metabolism assays. To measure the concentration of D-lactate in the supernatant of each culture, we used the D-Lactate Colorimetric Assay kit (Sigma-Aldrich) according to the manufacturer's instructions. The API 50 CH test (bioMérieux) was used to compare the abilities of GBS strains from each target pair to ferment 49 different carbohydrates. Each strain was initially grown for 18 to 24 h on TH agar plates. Afterwards, a quantity of cells amounting to two bacterial loops was suspended in 10 ml of API 50 CHL medium. The culture suspensions were then loaded onto API 50 CH test strips in accordance with the manufacturer's instructions, and the tubes were covered with mineral oil. The test strips were incubated at 37°C, and results were recorded 24 and 48 h later.

Nucleotide sequence accession numbers. Sequencing reads from each run and the corresponding genome assemblies have been deposited in the EMBL nucleotide sequence database (http://www.ebi.ac.uk/ena) under study accession number PRJEB6691. For the accession numbers of the individual files, see Table S1 in the supplemental material.

RESULTS

Characterization and genotyping of GBS pairs. For this study, a total set of 47 GBS samples were obtained from 19 pairs of infected newborns and their mothers (Table 1; see Table S1 in the supplemental material). One to three samples were collected per individual, corresponding to various sites colonized by GBS at different time points relative to the child's birth. A total of 11 samples were collected from carriage-related sites in the mother (amniotic fluid, placenta, and vaginal fluid), in contrast to 10 from other contaminated samples (blood, milk, and urine). Conversely, 22 blood, cerebrospinal fluid, and urine samples were obtained from the newborns, together with only 4 carriage samples of gastric fluid. Of the 19 pairs analyzed, 13 corresponded to EOD cases and 6 corresponded to LOD, with strains isolated 16 to 110 days after the infant's birth (Table 1).

After whole-genome sequencing, CPSs and MLST profiles were determined by BLASTn similarity searches of the assembled genomes of the samples (Table 1). CPS III and, more specifically, hypervirulent clone ST17 were the most abundant lineages identified in our work, accounting for 31 (66%) and 25 (53%) strains, respectively. GBS samples obtained from LOD belonged exclusively to ST17 ($n = 4$) and ST23 ($n = 2$), while ST17 accounted for four (80%) of five strains responsible for meningitis. From the genotyping of the 19 sets of mother-child samples analyzed, 17 pairs (89%) corresponded to identical CPSs and STs, whereas strains from pairs 7 and 11, both obtained from LOD, belonged to distinct lineages. In these two pairs, the infants were infected by ST17 strains and their mothers were colonized by ST1 (pair 7) or ST23 (pair 11). No genomic comparison of the mother and child populations of these two pairs was performed, considering that they belong to distantly related lineages (Fig. 1 and Table 1). Even so, two samples were retrieved from the infant corresponding to pair 11, isolated from blood samples cultured at 21 (child 11-1) and 35 (child 11-2) days after birth. Since the two isolates belong to the same ST, we compared them by using the first population sampled as a reference to track any genomic changes occurring throughout the course of infection.

Comparative analysis reveals genomic variations within functionally relevant regions. To analyze the population diversity of each pair of GBS samples (Table 1; see Table S1 in the supplemental material), different strategies based on read mapping and *de novo* assembly were applied as described in Materials and Methods and as depicted in Fig. S1 in the supplemental material. The purpose was to identify genomic differences between mother and child GBS populations while taking into account the possibility of within-host variation. A phylogenetic tree was built with all 19 GBS pairs, together with the most closely related strains whose complete genome sequences are publicly available. A total of 6,122 genomic SNPs were used (Fig. 1). Even though pairs 4, 5, and 8 were all ST17 strains collected from the same hospital during the same year, their phylogenetic distribution suggests that these pairs are no closer to each other than to the other ST17 pairs collected in different time periods from different hospitals. Moreover, mother and child strains from pairs 7 and 11 are clearly shown to belong to divergent lineages (Fig. 1).

No genomic differences were detected in eight (42%) pairs (1, 3, 5, 6, 8, 13, 16, and 17), encompassing seven incidences of EOD and one of LOD. Whole-genome comparison of the nine genetically diverse pairs of samples (2, 4, 9, 10, 12, 14, 15, 18, and 19), as well as the two samples obtained from child 11, resulted in the identification of a total of 21 SNPs and seven indels, ranging from 1 to 1,132 bp (Table 2; see Table S2 in the supplemental material). Determination of the frequency of the reads supporting each of the SNPs detected within the population allowed the identification of 16 fixed mutations and five polymorphic sites present in less than 97% (between 3 and 91%) of the population sampled (see Table S2). BLASTn similarity search of the flanking regions of each mutated site against published genomes in the NCBI database allowed the characterization of each variant identified (Table

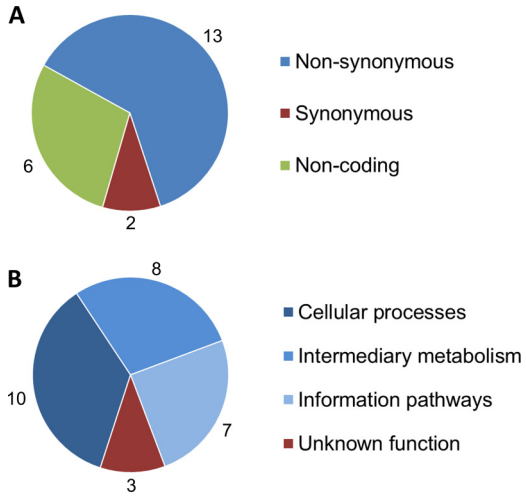


FIG 2 Characteristics of the mutations identified. (A) Classification of the SNPs identified in 10 of the 19 pairs of GBS samples analyzed by comparative genomics. (B) Functional category of each mutated region based on homology search of the flanking gene of each site, as annotated in the SagaList database of GBS NEM316. Both the type of mutation and the functional category are represented by different colors, as shown in the key. Each value depicted is the number of mutations identified belonging to each category.

2; see Table S2 in the supplemental material). Thirteen SNPs were classified as nonsynonymous (11 missense and 2 nonsense), but only 2 were classified as synonymous (Fig. 2A). Furthermore, one out-of-frame deletion (1 bp) was detected, together with one in-frame deletion (18 bp), a 46-bp out-of-frame duplication, and a large 1,132-bp deletion encompassing a phage terminase. Surprisingly, nine of the mutations (six SNPs and three indels) occurred in noncoding sequences, including five in 5′ untranslated regions (UTRs) and one in a 3′ UTR. Overall, on the basis of the functional classification of all the mutated regions, deduced from the SagaList database of GBS NEM316 (47, 48), 10 were related to cell envelope and cellular processes, 8 were related to intermediary metabolism, 7 were related to information pathways, and 3 were related to unknown functions (Fig. 2B).

In order to identify the most likely ancestral sequence of the common ancestor of each GBS pair and determine the mutated strain, we compared each variable locus with the 300 sequenced GBS genomes available in the NCBI database as an outgroup. These comprise carriage and invasive isolates of human origin, as well as strains from bovine, fish, and other animal hosts. We infer the wild-type allele as the one identical to all of the genetically closest genomes and the mutant variant as the one differing from these sequences (see Fig. S2 in the supplemental material). For all but four of the mutations we detected, the same variant was present in each of the 300 NCBI strains. A total of 14 mutations were deduced to have occurred within nine isolates obtained from the mother and 14 mutations within eight samples from the child (Table 2).

Focusing on the most relevant mutations identified in the GBS populations sampled, assessment of pair 2, from a case of LOD, revealed two mutations in the mother strains isolated from milk: one SNP within a gene coding for a glyoxalase (*gbs1596*, according to GBS NEM316 [48]), a detoxification system of reactive aldehydes, and another in a region encoding a polynucleotide phosphorylase (*gbs0198*). This enzyme is conserved among bacteria and has been shown to regulate the expression of virulence factors in *S. aureus*, *Francisella tularensis*, and *Yersinia* spp. (52–54).

In pair 12 (EOD), only one mutation was present in the mother GBS population, obtained from a blood culture, in the 5′ UTR of a glucose-specific phosphotransferase system (PTS) enzyme IIABC gene (*gbs1946* [*ptsG*]) that takes part in the phosphorylation and transport of glucose (55). In the gastric fluid sample from the child (child 12-1), six polymorphic variants were detected in various proportions (see Table S2 in the supplemental material), including an additional mutation in the 5′ UTR of *ptsG* (Table 2; see Table S2). In contrast, four mutations appear to have been positively selected specifically in the GBS population obtained from the blood (child 12-2). Interestingly, one of the mutations is located in the 5′ UTR of the gene encoding the response regulator of virulence CovR (*gbs1672* [*covR*]). Since the most common etiological pathway of EOD caused by GBS is aspiration or ingestion of GBS-contaminated fluids (4, 17, 18), our observations within this pair suggest that multiple GBS variants were ingested by the baby through the digestive tract but only one prevailed and was selected in the child's bloodstream.

Regarding pair 14 (LOD), a duplication of 46 bp in the MerR family of transcriptional regulators gene (*gbs1958*) was identified in the newborn sample. These regulators are responsible for the activation of genes associated with transport and multidrug resistance in response to various environmental stresses, as reported in *Bacillus subtilis* and other bacterial species (56). Conversely, two mutations were detected in the mother sample: one SNP in the 5′ UTR of a gene encoding a transcriptional regulator of the multiple antibiotic resistance regulator (MarR) family (*gbs0330*) and one additional missense SNP in *neuD*. MarR proteins have been described as responsible for controlling the expression of virulence factors and the bacterial response to antibiotic and oxidative stresses (57, 58). The other mutated gene, *neuD*, codes for an acetyltransferase with a role in the biosynthesis of capsular sialic acid and has been implicated in the development of meningitis in neonates (59).

Concerning pair 18 (EOD), two polymorphic mutations were detected, one of which is in a genomic region coding for trehalose-specific PTS enzyme IIABC (*gbs0189* [*trePI*]). Trehalose is known to contribute to the adaptation of bacteria to osmotic stress, while this particular PTS, responsible for trehalose phosphorylation and transport, is an upregulated target of CovR (60). The mutation was identified in approximately 90% of the population obtained specifically from the mother's blood sample (see Table S2 in the supplemental material). Strikingly, in pair 12, the mother strains also from a blood culture were mutated in *ptsG*, another gene involved in carbon source uptake.

Lastly, four mutations were found within the samples of pair 19 (EOD). In the three child populations, a nonsense mutation was detected in the phosphotransacetylase gene (*gbs1159* [*pta*]), which is involved in acetate metabolism and ATP production; a missense mutation in *covR*, altering the same system mutated in pair 12 with essential regulatory functions for virulence (31, 32); and a 1-bp insertion in the homopolymeric A tract affecting the promoter of the alpha-like Rib surface protein-encoding gene (*gbs0147* [*rib*]), a region known to have a significant impact on the expression of this immunogenic protein (61). In regard to the mother sample, a nonsynonymous mutation was found in the *recA* gene (*gbs2047*), coding for a multifunctional protein with a central role in DNA repair by homologous recombination (62).

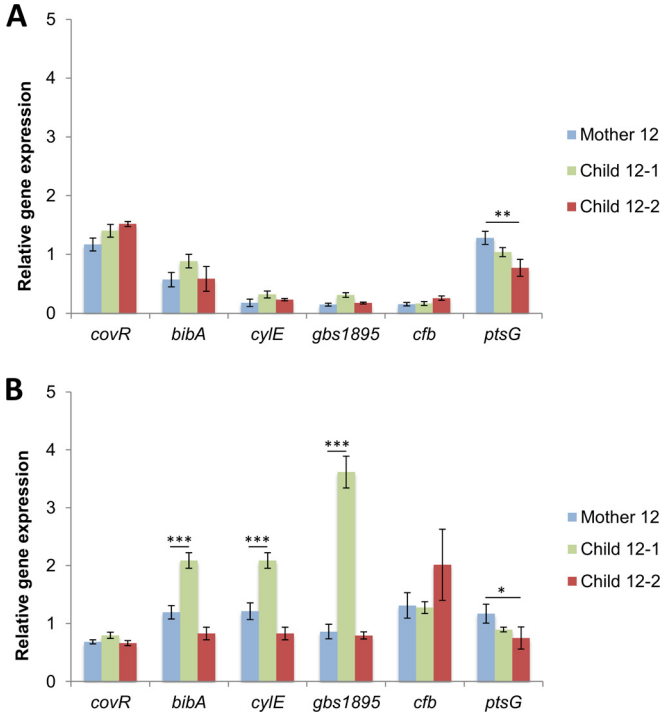


FIG 3 Relative expression of mutated genes (pair 12). RT-qPCR results obtained with GBS strains from pair 12 in TH medium (A) and after incubation for 1 h in human blood (B). The mother 12 isolate was mutated in the 5′ UTR of *gbs1946* (*ptsG*), and both child strains were mutated in *gbs0668* (D-lactate dehydrogenase) and *gbs1038* (permease). The isolate from child 12-1 tested presented additional mutations affecting *gbs1946* (*ptsG*), *gbs1377* (homocysteine *S*-methyltransferase), and a phage terminase, while the child 12-2 strain specifically was mutated in the 5′ UTR of *gbs1672* (*covR*) and *gbs0231* (putative transporter). Gene expression is represented after normalization to the housekeeping gene *gyrA*. Gene names are indicated below each graph. Experiments were performed in triplicate with three independent cultures. Error bars represent standard deviations. ***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$.

Almost all of the mutations identified are novel, since most of them were not present in any of the 300 GBS genomes available in the NCBI database as of April 2015, except for two SNPs and two indels. The two SNPs, identified in *neuD* and in the 5′ UTR of *ptsG* from mother 12, were each detected in one genome of human origin. Moreover, the one-nucleotide indel in the 5′ UTR of *gbs0419* from mother 15 was distributed among 122 strains from human, bovine, and fish hosts, while the 18-bp deletion in *gbs1377* was identified in two human GBS strains. Altogether, of the 14 mutations identified in our study in the child strains of five pairs, three fixed variants were associated with pathogenicity-related features in two pairs, comprising two different mutations in *covR* and one in the promoter region of the gene encoding the surface protein Rib.

Relative gene expression quantification showed distinct transcription levels of virulence-related genes. In light of the genomic comparison of the 19 pairs of GBS samples, we focused on fixed mutations identified in the child and relating to virulence-associated regions to determine any differences in the expression levels of the affected genes (pairs 12 and 19). In pair 12, with SNPs in the 5′ UTRs of *covRS* and *ptsG*, RT-qPCRs were performed with both mutated genes, as well as four major targets of CovR (Fig. 3; see Table S3 in the supplemental material). Given that CovR is primarily a virulence repressor, we chose as negatively regulated targets the *hvgA* (*bibA*) gene, which encodes a variable surface protein (63) shown to be a hypervirulent adhesin in the ST17 lineage (22), and *cylE*, from the *cyl* operon responsible for the synthesis of β-hemolysin (30). One additional target, coding for a hypothetical secreted protein (*gbs1895*) repressed by CovR, was also tested (32). Contrariwise, the *cfb* gene encoding the CAMP factor toxin (32) was selected as a positively regulated target of CovRS. Seeing that CovR is a major regulator dependent on the host environment, gene expression was determined both in TH medium and after 1 h of incubation in human blood. To compare the expression of the different strains from pair 12, one isolate each from mother 12 and child 12-2 (Table 1) was used, representing the genotype observed within 100% of the population obtained from blood (Table 2; see Table S2 in the supplemental material). Additionally, an isolate collected from the gastric fluid of child 12-1 was also tested and confirmed by sequencing to contain the four SNPs observed at a higher frequency in the original sample (*gbs0668*, *gbs1038*, and *gbs1946*), as well as the 18-bp deletion in *gbs1377* and the 1,132-bp deletion of a phage terminase (Table 2; see Table S2 in the supplemental material). Comparing gene expression levels between growth in TH and after 1 h of incubation in human blood showed that *covR* expression was lower in whole human blood, contrasting with the transcription of the CovR negatively regulated targets *bibA*, *cylE*, and *gbs1895*, which was significantly increased (Fig. 3). However, no difference was observed between expression levels of the mother strain and the child strain with the mutation in the 5′ UTR of *covR* (child 12-2). Surprisingly, there was a significant increase in the expression of *bibA*, *cylE*, and *gbs1895* in the strain isolated from gastric fluid (child 12-1) after incubation in human blood, hinting that it could be the result of an indirect effect of one of the other mutations found in this strain. This difference was confirmed to be significant for *gbs1895* following incubation with blood from a second donor, although a smaller effect was observed for *bibA* and *cylE* (see Fig. S3A in the supplemental material). In addition, to check if these expression differences could be associated with the nonsynonymous mutation in the D-lactate dehydrogenase, we quantified the amounts of D-lactate in the supernatants of all of the pair 12 cultures used for the RT-qPCR but found no significant differences (see Fig. S3B in the supplemental material). Lastly, the SNP in the 5′ UTR of *ptsG* found in mother 12 led to a small but significant expression increase solely in comparison to the strain from the child's blood (child 12-2), which was devoid of any mutations affecting this gene (Fig. 3).

As for pair 19, representative isolates from both the mother and the child were tested, with the child isolate containing a nonsynonymous SNP in the *covR* gene, as well as a 1-bp insertion in the promoter of the gene encoding the surface protein Rib. With this in mind, quantitative experiments were carried out with additional sets of primers for the *rib* gene (Fig. 4; see Table S3 in the supplemental material). Interestingly, only after incubation in human blood was there a significant increase in the expression of *hvgA*, *cylE*, and *gbs1895* in the mutated strain retrieved from the child, together with a significant reduction in the transcription of *covR* itself. This suggests that the amino acid change in the *covR* mutant affected the regulation of this system and caused the overexpression of several virulence-associated genes in the child strain. Finally, the *rib* gene in both the mother and child strains was significantly more strongly expressed in whole human blood than

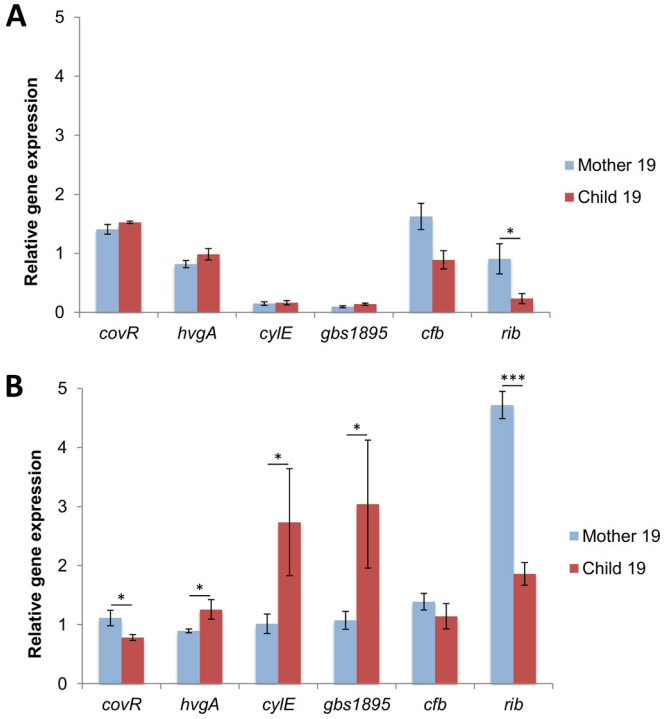


FIG 4 Relative expression of mutated genes (pair 19). RT-qPCR results obtained with GBS strains from pair 19 in TH medium (A) and after incubation for 1 h in human blood (B). The mother 19 isolate tested was mutated in *gbs2047* (*recA*), and the child strain was mutated in *gbs1159* (*pta*), *gbs1672* (*covR*), and the promoter of *GBSCOH1_0416* (*rib*). Gene expression is represented after normalization to the housekeeping gene *gyrA*. Gene names are indicated below each graph. Experiments were performed in triplicate with three independent cultures. Error bars represent standard deviations. ***, $P < 0.001$; *, $P < 0.05$.

in TH medium, but there was a consistent and significant reduction in expression in the child strain, compared to that in the mother sample, under both of the conditions tested (Fig. 4). The reduced expression of *rib* might have contributed to the mutant strain's escape from the antibodies transmitted by the mother, allowing GBS to replicate and spread within the infected infant.

Carbohydrate fermentation assays revealed no significant differences in metabolic activity. Because we identified mutated genes related to sugar transport and metabolism, we performed the API 50 CHL test on GBS pairs harboring genomic differences in genes involved in carbon source metabolism. This assay examines the ability of the bacteria to ferment 49 different carbohydrates. Phenotypic tests were performed in duplicate on strains from pairs 12, 15, 18, and 19, but no significant differences were found under the conditions tested (see Table S4 in the supplemental material).

DISCUSSION

Neonatal infections caused by GBS remain a significant health issue worldwide (1–4). In this work, we present for the first time a comparative genomic analysis of pairs of GBS samples obtained from infected infants and their mothers. Whole-genome comparison of 10 of 19 mother-child pairs revealed a total of 21 SNPs and seven indels (Table 2). In almost half (42%) of the GBS pairs, no differences were observed. Furthermore, the variants present in mother 15 were absent from the two samples from her child, while strains from pair 19 infecting the blood, cerebrospinal fluid, and urine of the newborn presented the same genomic makeup, distinct from its mother counterpart sample. This suggests that there is a low level of within-host GBS diversity, even if the sampling procedures might not have captured the complete genotypic diversity within each GBS population.

Despite the random nature of genomic mutagenesis, we detected a strong bias for particular types of mutations affecting genes with known functional implications (Fig. 2B). Indeed, only three mutations (11%) altered genes with unknown functions. Furthermore, of the 21 SNPs identified, only 5 appear to be polymorphic, suggesting that most of the mutations were likely selected and fixed in the population. The polymorphic variants could also have arisen from random genetic drift following the transmission of GBS between hosts, although we cannot exclude the possibility that some of these mutations occurred during subculturing in laboratory medium. Notably, most mutations corresponded to nonsynonymous substitutions, with a 13:2 ratio in relation to synonymous changes (Fig. 2A). Moreover, more than one-third of the mutations were located within noncoding sequences (Table 2), although they comprise only 10 to 12% of the GBS genome. Experimentally, their phenotypic impact was supported by the observed change in transcription levels resulting from a mutation in the 5′ UTR of *ptsG* in mother 12, as well as from a 1-bp insertion in the promoter region of the gene encoding the surface protein Rib in child 19 (Fig. 3 and 4). Altogether, these observations reinforce the significance of the mutations found, strongly suggesting that they have undergone a process of positive selection within their host.

Bearing in mind that colonization by GBS is acknowledged to be mostly asymptomatic, we show interesting data on the possible mechanisms further enhancing the virulence of this bacterium. Overall, 14 mutant variants were detected in GBS samples from infected newborns, showing a selective adaptation and evolution of GBS during neonatal disease. Particularly in pair 11, strains contaminating the child's bloodstream underwent two mutations in 14 days. In pair 19, the infant strains were mutated at three different loci after 5 to 6 days following the child's birth (Table 2). On the contrary, strains from the newborn obtained at the time of birth or 1 day later yielded the most similarities to the corresponding mother strains, as exemplified by pairs 1, 3, 6, 8, 9, 16, and 17, in which no differences were detected. Altogether, identical pairs made up more than half of the cases of EOD analyzed and only one of the six instances of LOD that were part of this study. Considering the immaturity of the newborn immune system at the time of birth, colonization by any GBS isolate transferred from the mother might be enough to lead to an infection state. Conversely, the development of the neonate's defenses throughout the first few weeks of life would be prone to elicit selective pressure on GBS strains with any fitness advantage.

From a total of 14 mutations identified in the strains from infected infants, two independently fixed mutations were found related to the two-component response regulator gene of the *covRS* locus, a central regulator of GBS virulence (32). This points toward the presence of selective pressures in favor of the alteration of CovR regulation during neonatal infections. In GAS, a broad range of nonsynonymous substitutions in the regulator gene (*covR*), as well as small indels disrupting the sensor gene (*covS*), have been described in invasive strains (64, 65). These mutations were shown to have a profound transcription effect, increasing the

expression of virulence-associated genes (65). Recently, a systematic screening for hyperpigmented GBS strains identified the presence of molecular alterations of *covR* in four isolates (66), but none of these strains were of neonatal origin. Here, we showed in GBS a transcriptional effect of a nonsynonymous mutation in the *covRS* locus after incubation in human blood (Fig. 4). We conclude that, similarly to GAS, mutations in the *covRS* locus might be selected during the development of GBS infections in neonates, but the effect of the mutations is considerably more subtle.

Another small indel in the child samples (pair 19) was pinpointed in the poly(A) tract upstream of the promoter of the gene encoding the alpha-like Rib protein. Short sequence repeat variations in this homopolymer tract, decreasing the expression of this family of surface proteins, were shown to be selected in a mouse model of immunization (61). Furthermore, smaller amounts of antibodies against alpha and Rib proteins, present in the host, have also been reported to be associated with a higher incidence of disease (67). Therefore, alpha-like proteins have been considered potential antigens for the development of a vaccine against GBS (67–69). Here, we observed a 4-fold increase in the expression of the *rib* gene following incubation in whole human blood, leading to a high level of expression of this immunogenic protein. However, the transcription levels of *rib* were significantly lower in the mutant strain colonizing the infant than in the mother strain, especially after incubation in blood (Fig. 4). Thus, the mutation in the poly(A) tract might have allowed GBS in the child to evade the protection provided by the antibodies transmitted from the mother.

Although the main focus of our work was characterizing GBS's transition from commensal to invasive after mother-to-child transmission, particular cases of infection in the mother were also part of this study, contributing to a better understanding of the pathology of GBS outside the context of neonatal disease. In this regard, two samples obtained from women diagnosed with septicemia (pairs 12 and 18) revealed parallel mutations in sugar transport systems, namely, altering glucose- and trehalose-specific PTS enzyme-encoding genes. Owing to the unique nutritional components in the mother's bloodstream, variations in metabolism-related mechanisms might provide a better adaptation of these clones and promote their progression to disease. Experimental results hinted at a functional effect of the mutation in the gene encoding a glucose-specific PTS enzyme (*ptsG*), showing different levels of expression in the mother (mother 12) and child (child 12-2) strains recovered from blood cultures (Fig. 3).

By combining the allelic profile with the available clinical information (i.e., the sampling date and body site from which the GBS strain was isolated), we are able to hypothesize the most likely direction of GBS transmission between the mother and child within each pair (Fig. 5). In pairs 7 and 11, however, the mother and child strains belong to distantly related lineages (Fig. 1 and Table 1). Genomic analyses showed no indication of the mixed presence of both clones in either the mother or the child samples. In pair 7, both strains were obtained solely from blood cultures, so it is unknown whether the mother's vaginal tract was colonized by the strain that contaminated her child. Conversely, the distinct strains from pair 11 were obtained from the vaginal tract of the mother and a blood culture of her child, suggesting that there was no mother-to-child transmission. Therefore, GBS colonization might have occurred through contamination by another carrier. Apart from these unique situations, in five pairs (4, 12, 15, 18, and

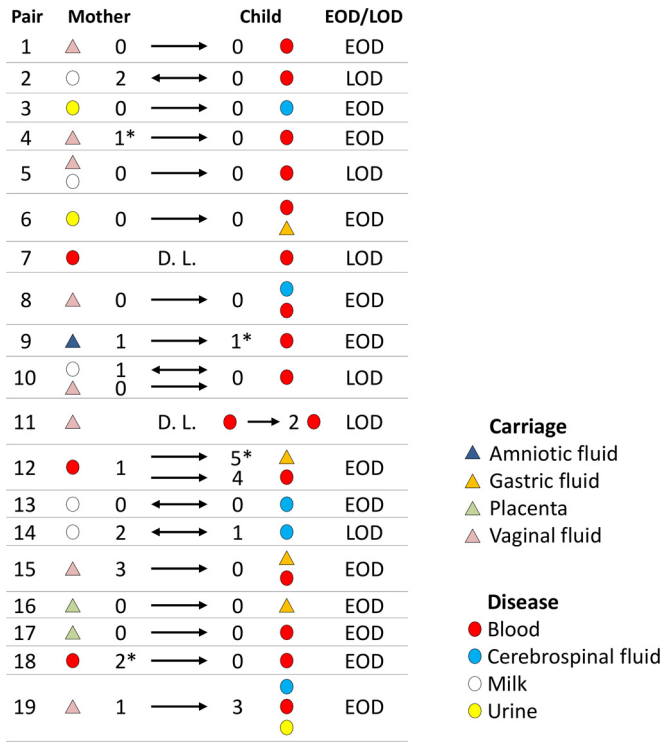


FIG 5 Transmission pathways between GBS samples. Graphical representation of the most likely transmission route hypothesized for each pair of GBS samples. The colonization site and whether it corresponds to a situation of carriage or disease are represented by different shapes and colors according to the key. The arrows indicate the presumed direction of transfer inferred from the clinical information available together with the number of mutations determined in the mother and child GBS samples. Polymorphic variants, present in <97% of the population sampled, are indicated by asterisks. When the child was diagnosed with EOD and sampled at the time of birth, even if the mutant strain was retrieved from the mother, we assumed mother-to-child transmission. Ambiguous transmission direction is indicated by a double-headed arrow. D. L. indicates that the two strains belong to distinct lineages, so the origin of the newborn GBS clone is unknown.

19), both the collection date (at birth or just after) and the origin of the sample (blood culture or vaginal fluid) are in favor of mother-to-child transmission of GBS (Fig. 5), which is acknowledged as the most frequent form of contamination. Intriguingly, three mutant GBS strains (pairs 2, 10, and 14) were collected from the mother's milk after the child's birth (110, 16, and 33 days, respectively). This indicates that the mutations identified in the GBS sample from the mother occurred after colonization of the infant by the wild-type variants, raising the possibility that the child contaminated the milk through breastfeeding (Fig. 5). In agreement with this hypothesis, previous studies have shown that the microbial composition of human milk may be shaped by multiple sources, including the baby's oral microbiome (70). Alternatively, it is also possible that the milk was colonized independently by a GBS strain from the mother herself. In fact, in pair 10, the vaginal sample from the mother was also positive for the same GBS strain infecting her child, whereas in pairs 2 and 14, no additional body sites of the mother were sampled. If the wild-type variant initially colonized the mother's milk, these strains might have mutated independently in the mammary gland after having been transmitted to the baby. To summarize, we provide new findings regarding

the pathogenesis and etiology of neonatal GBS infections by analyzing the *in vivo* diversity and evolution of clones in carriage and disease. Ultimately, this shows that the transient colonization of neonates by GBS elicits a particular adaptation and evolution of this species that might underlie its switch from commensal to pathogen.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (grant ANR-13-PRTS-0006-04), by the Labex IBEID, and by the Institut de Veille Sanitaire, INSERM, University Paris Descartes. Sequencing was performed at the Pasteur Genopole, a member of France Génomique (ANR10-IBNS-09-08). A.A. is a scholar in the Pasteur-Paris University (PPU) International Ph.D. program and received a stipend from the ANR-LabEx IBEID.

We thank Laurence Ma for her help in performing the Illumina sequencing and Pedro Escoll Guerrero for his help in carrying out the experiments with human blood. We also thank Carmen Buchrieser for critical reading of the manuscript and Dimitri Desvillechabrol, Pierre-Emmanuel Douarre, and Romain Guérillot for fruitful discussions.

REFERENCES

1. **Edmond KM, Kortsalioudaki C, Scott S, Schrag SJ, Zaidi AK, Cousens S, Heath PT.** 2012. Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. Lancet **379:**547–556. http://dx.doi.org/10.1016/S0140-6736(11)61651-6.
2. **Verani JR, McGee L, Schrag SJ.** 2010. Prevention of perinatal group B streptococcal disease: revised guidelines from CDC, 2010. MMWR Recomm Rep **59**(RR10):1–32. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5910a1.htm.
3. **Poyart C, Reglier-Poupet H, Tazi A, Billoet A, Dmytruk N, Bidet P, Bingen E, Raymond J, Trieu-Cuot P.** 2008. Invasive group B streptococcal infections in infants, France. Emerg Infect Dis **14:**1647–1649. http://dx.doi.org/10.3201/eid1410.080185.
4. **Edwards MS, Baker CJ.** 2005. Group B streptococcal infections, p 1091–1156. *In* Remington JS, Klein JO Wilson CB, Nizet V, Maldonado Y (ed), Infectious diseases of the fetus and newborn infant. Elsevier Saunders, Philadelphia, PA.
5. **Rosinski-Chupin I, Sauvage E, Mairey B, Mangenot S, Ma L, Da Cunha V, Rusniok C, Bouchier C, Barbe V, Glaser P.** 2013. Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage. BMC Genomics **14:**252. http://dx.doi.org/10.1186/1471-2164-14-252.
6. **Rato MG, Bexiga R, Florindo C, Cavaco LM, Vilela CL, Santos-Sanches I.** 2013. Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. Vet Microbiol **161:**286–294. http://dx.doi.org/10.1016/j.vetmic.2012.07.043.
7. **Kalmus P, Aasmae B, Karssin A, Orro T, Kask K.** 2011. Udder pathogens and their resistance to antimicrobial agents in dairy cows in Estonia. Acta Vet Scand **53:**4. http://dx.doi.org/10.1186/1751-0147-53-4.
8. **Wyder AB, Boss R, Naskova J, Kaufmann T, Steiner A, Graber HU.** 2011. *Streptococcus* spp. and related bacteria: their identification and their pathogenic potential for chronic mastitis—a molecular approach. Res Vet Sci **91:**349–357. http://dx.doi.org/10.1016/j.rvsc.2010.09.006.
9. **Keefe GP.** 1997. *Streptococcus agalactiae* mastitis: a review. Can Vet J **38:**429–437.
10. **Mian GF, Godoy DT, Leal CAG, Yuhara TY, Costa GM, Figueiredo HCP.** 2009. Aspects of the natural history and virulence of *S. agalactiae* infection in Nile tilapia. Vet Microbiol **136:**180–183. http://dx.doi.org/10.1016/j.vetmic.2008.10.016.
11. **Fluegge K, Siedler A, Heinrich B, Schulte-Moenting J, Moennig MJ, Bartels DB, Dammann O, von Kries R, Berner R, German Pediatric Surveillance Unit Study.** 2006. Incidence and clinical presentation of invasive neonatal group B streptococcal infections in Germany. Pediatrics **117:**e1139-45. http://dx.doi.org/10.1542/peds.2005-2481.
12. **Kalliola S, Vuopio-Varkila J, Takala AK, Eskola J.** 1999. Neonatal group B streptococcal disease in Finland: a ten-year nationwide study. Pediatr Infect Dis J **18:**806–810. http://dx.doi.org/10.1097/00006454-199909000-00012.
13. **Neto MT.** 2008. Group B streptococcal disease in Portuguese infants younger than 90 days. Arch Dis Child Fetal Neonatal Ed **93:**F90–F93.
14. **Zangwill KM, Schuchat A, Wenger JD.** 1992. Group B streptococcal disease in the United States, 1990: report from a multistate active surveillance system. MMWR CDC Surveill Summ **41:**25–32.
15. **Libster R, Edwards KM, Levent F, Edwards MS, Rench MA, Castagnini LA, Cooper T, Sparks RC, Baker CJ, Shah PE.** 2012. Long-term outcomes of group B streptococcal meningitis. Pediatrics **130:**e8-15. http://dx.doi.org/10.1542/peds.2011-3453.
16. **Bedford H, de Louvois J, Halket S, Peckham C, Hurley R, Harvey D.** 2001. Meningitis in infancy in England and Wales: follow up at age 5 years. BMJ **323:**533–536. http://dx.doi.org/10.1136/bmj.323.7312.533.
17. **Gibson RL, Nizet V, Rubens CE.** 1999. Group B streptococcal beta-hemolysin promotes injury of lung microvascular endothelial cells. Pediatr Res **45:**626–634. http://dx.doi.org/10.1203/00006450-199905010-00003.
18. **Nizet V, Kim KS, Stins M, Jonas M, Chi EY, Nguyen D, Rubens CE.** 1997. Invasion of brain microvascular endothelial cells by group B streptococci. Infect Immun **65:**5074–5081.
19. **Hood M, Janney A, Dameron G.** 1961. Beta hemolytic streptococcus group B associated with problems of the perinatal period. Am J Obstet Gynecol **82:**809–818.
20. **Baker CJ, Barrett FF.** 1974. Group B streptococcal infections in infants—importance of various serotypes. JAMA **230:**1158–1160.
21. **Boswihi SS, Udo EE, Al-Sweih N.** 2012. Serotypes and antibiotic resistance in group B streptococcus isolated from patients at the Maternity Hospital, Kuwait. J Med Microbiol **61:**126–131. http://dx.doi.org/10.1099/jmm.0.035477-0.
22. **Tazi A, Disson O, Bellais S, Bouaboud A, Dmytruk N, Dramsi S, Mistou MY, Khun H, Mechler C, Tardieux I, Trieu-Cuot P, Lecuit M, Poyart C.** 2010. The surface protein HvgA mediates group B streptococcus hypervirulence and meningeal tropism in neonates. J Exp Med **207:**2313–2322. http://dx.doi.org/10.1084/jem.20092594.
23. **Manning SD, Springman AC, Lehotzky E, Lewis MA, Whittam TS, Davies HD.** 2009. Multilocus sequence types associated with neonatal group B streptococcal sepsis and meningitis in Canada. J Clin Microbiol **47:**1143–1148. http://dx.doi.org/10.1128/JCM.01424-08.
24. **Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, Glaser P, Rusniok C, Crook DW, Harding RM, Bisharat N, Spratt BG.** 2003. Multilocus sequence typing system for group B *Streptococcus*. J Clin Microbiol **41:**2530–2536. http://dx.doi.org/10.1128/JCM.41.6.2530-2536.2003.
25. **Brochet M, Rusniok C, Couve E, Dramsi S, Poyart C, Trieu-Cuot P, Kunst F, Glaser P.** 2008. Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. Proc Natl Acad Sci U S A **105:**15961–15966. http://dx.doi.org/10.1073/pnas.0803654105.
26. **Da Cunha V, Davies MR, Douarre PE, Rosinski-Chupin I, Margarit I, Spinali S, Perkins T, Lechat P, Dmytruk N, Sauvage E, Ma L, Romi B, Tichit M, Lopez-Sanchez MJ, Descorps-Declere S, Souche E, Buchrieser C, Trieu-Cuot P, Moszer I, Clermont D, Maione D, Bouchier C, McMillan DJ, Parkhill J, Telford JL, Dougan G, Walker MJ, DEVANI Consortium, Holden MTG, Poyart C, Glaser P.** 2014. *Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. Nat Commun **5:**4544 http://dx.doi.org/10.1038/ncomms5544.
27. **Jones AL, Knoll KM, Rubens CE.** 2000. Identification of *Streptococcus agalactiae* virulence genes in the neonatal rat sepsis model using signature-tagged mutagenesis. Mol Microbiol **37:**1444–1455. http://dx.doi.org/10.1046/j.1365-2958.2000.02099.x.
28. **Chmouryguina I, Suvorov A, Ferrieri P, Cleary PP.** 1996. Conservation of the C5a peptidase genes in group A and B streptococci. Infect Immun **64:**2387–2390.
29. **Lindahl G, Stalhammar-Carlemalm M, Areschoug T.** 2005. Surface proteins of *Streptococcus agalactiae* and related proteins in other bacterial pathogens. Clin Microbiol Rev **18:**102–127. http://dx.doi.org/10.1128/CMR.18.1.102-127.2005.
30. **Spellerberg B, Pohl B, Haase G, Martin S, Weber-Heynemann J, Lutticken R.** 1999. Identification of genetic determinants for the hemolytic activity of *Streptococcus agalactiae* by ISS1 transposition. J Bacteriol **181:**3212–3219.
31. **Graham MR, Smoot LM, Migliaccio CAL, Virtaneva K, Sturdevant DE, Porcella SF, Federle MJ, Adams GJ, Scott JR, Musser JM.** 2002. Virulence control in group A streptococcus by a two-component gene regulatory system: global expression profiling and *in vivo* infection modeling.

Proc Natl Acad Sci U S A **99:**13855–13860. http://dx.doi.org/10.1073/pnas.202353699.

32. **Lamy MC, Zouine M, Fert J, Vergassola M, Couve E, Pellegrini E, Glaser P, Kunst F, Msadek T, Trieu-Cuot P, Poyart C.** 2004. CovS/CovR of group B streptococcus: a two-component global regulatory system involved in virulence. Mol Microbiol **54:**1250–1268. http://dx.doi.org/10.1111/j.1365-2958.2004.04365.x.
33. **Santi I, Grifantini R, Jiang SM, Brettoni C, Grandi G, Wessels MR, Soriani M.** 2009. CsrRS regulates group B *Streptococcus* virulence gene expression in response to environmental pH: a new perspective on vaccine development. J Bacteriol **191:**5387–5397. http://dx.doi.org/10.1128/JB.00370-09.
34. **Price EP, Sarovich DS, Mayo M, Tuanyok A, Drees KP, Kaestli M, Beckstrom-Sternberg SM, Babic-Sternberg JS, Kidd TJ, Bell SC, Keim P, Pearson T, Currie BJ.** 2013. Within-host evolution of *Burkholderia pseudomallei* over a twelve-year chronic carriage infection. mBio **4**(4): e00388-3.
35. **Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, Iqbal Z, Rimmer AJ, Cule M, Ip CLC, Didelot X, Harding RM, Donnelly P, Peto TE, Crook DW, Bowden R, Wilson DJ.** 2012. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci U S A **109:**4550–4555. http://dx.doi.org/10.1073/pnas.1113219109.
36. **Marvig RL, Sommer LM, Molin S, Johansen HK.** 2015. Convergent evolution and adaptation of Pseudomonas aeruginosa within patients with cystic fibrosis. Nat Genet **47:**57–64. http://dx.doi.org/10.1038/ng.3148.
37. **Paterson GK, Harrison EM, Murray GGR, Welch JJ, Warland JH, Holden MTG, Morgan FJE, Ba X, Koop G, Harris SR, Maskell DJ, Peacock SJ, Herrtage ME, Parkhill J, Holmes MA.** 2015. Capturing the cloud of diversity reveals complexity and heterogeneity of MRSA carriage, infection and transmission. Nat Commun **6:**6560–6560. http://dx.doi.org/10.1038/ncomms7560.
38. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
39. **Poyart C, Tazi A, Reglier-Poupet H, Billoet A, Tavares N, Raymond J, Trieu-Cuot P.** 2007. Multiplex PCR assay for rapid and accurate capsular typing of group B streptococci. J Clin Microbiol **45:**1985–1988. http://dx.doi.org/10.1128/JCM.00159-07.
40. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25:**1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.
41. **McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA.** 2010. The Genome Analysis ToolKit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20:**1297–1303. http://dx.doi.org/10.1101/gr.107524.110.
42. **Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA.** 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics **11**(1110)**:**11.10.1–11.10.33. http://dx.doi.org/10.1002/0471250953.bi1110s43.
43. **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ.** 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43:**491–498. http://dx.doi.org/10.1038/ng.806.
44. **Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, 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.
45. **Deatherage DE, Barrick JE.** 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using *breseq*. Methods Mol Biol **1151:**165–188. http://dx.doi.org/10.1007/978-1-4939-0554-6_12.
46. **Rosinski-Chupin I, Sauvage E, Sismeiro O, Villain A, Da Cunha V, Caliot M-E, Dillies M-A, Trieu-Cuot P, Bouloc P, Lartigue M-F, Glaser P.** 2015. Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen *Streptococcus agalactiae*. BMC Genomics **16:**419. http://dx.doi.org/10.1186/s12864-015-1583-4.
47. **Moszer I, Jones LM, Moreira S, Fabry C, Danchin A.** 2002. SubtiList: the reference database for the *Bacillus subtilis* genome. Nucleic Acids Res **30:**62–65. http://dx.doi.org/10.1093/nar/30.1.62.
48. **Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, Zouine M, Couve E, Lalioui L, Poyart C, Trieu-Cuot P, Kunst F.** 2002. Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol **45:**1499–1513. http://dx.doi.org/10.1046/j.1365-2958.2002.03126.x.
49. **Leekitcharoenphon P, Kaas RS, Thomsen MCF, Friis C, Rasmussen S, Aarestrup FM.** 2012. snpTree—a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics **13**(Suppl 7)**:**S6. http://dx.doi.org/10.1186/1471-2164-13-S7-S6.
50. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28:**2731–2739. http://dx.doi.org/10.1093/molbev/msr121.
51. **Oggioni MR, Trappetti C, Kadioglu A, Cassone M, Iannelli F, Ricci S, Andrew PW, Pozzi G.** 2006. Switch from planktonic to sessile life: a major event in pneumococcal pathogenesis. Mol Microbiol **61:**1196–1210. http://dx.doi.org/10.1111/j.1365-2958.2006.05310.x.
52. **Numata S, Nagata M, Mao H, Sekimizu K, Kaito C.** 2014. CvfA protein and polynucleotide phosphorylase act in an opposing manner to regulate *Staphylococcus aureus* virulence. J Biol Chem **289:**8420–8431. http://dx.doi.org/10.1074/jbc.M114.554329.
53. **Clements MO, Eriksson S, Thompson A, Lucchini S, Hinton JCD, Normark S, Rhen M.** 2002. Polynucleotide phosphorylase is a global regulator of virulence and persistency in *Salmonella enterica*. Proc Natl Acad Sci U S A **99:**8784–8789. http://dx.doi.org/10.1073/pnas.132047099.
54. **Rosenzweig JA, Chromy B, Echeverry A, Yang J, Adkins B, Plano GV, McCutchen-Maloney S, Schesser K.** 2007. Polynucleotide phosphorylase independently controls virulence factor expression levels and export in *Yersinia* spp. FEMS Microbiol Lett **270:**255–264. http://dx.doi.org/10.1111/j.1574-6968.2007.00689.x.
55. **Postma PW, Lengeler JW, Jacobson GR.** 1993. Phosphoenolpyruvate-carbohydrate phosphotransferase systems of bacteria. Microbiol Rev **57:**543–594.
56. **Brown NL, Stoyanov JV, Kidd SP, Hobman JL.** 2003. The MerR family of transcriptional regulators. FEMS Microbiol Rev **27:**145–163. http://dx.doi.org/10.1016/S0168-6445(03)00051-2.
57. **Ellison DW, Miller VL.** 2006. Regulation of virulence by members of the MarR/SlyA family. Curr Opin Microbiol **9:**153–159. http://dx.doi.org/10.1016/j.mib.2006.02.003.
58. **Wilkinson SP, Grove A.** 2006. Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. Curr Issues Mol Biol **8:**51–62.
59. **Pailhories H, Quentin R, Lartigue MF.** 2013. The transcription of the *neuD* gene is stronger in serotype III group B streptococci strains isolated from cerebrospinal fluid than in strains isolated from vagina. FEMS Microbiol Lett **349:**71–75.
60. **Elbein AD, Pan YT, Pastuszak I, Carroll D.** 2003. New insights on trehalose: a multifunctional molecule. Glycobiology **13:**17R-27R. http://dx.doi.org/10.1093/glycob/cwg047.
61. **Puopolo KM, Madoff LC.** 2003. Upstream short sequence repeats regulate expression of the alpha C protein of group B Streptococcus. Mol Microbiol **50:**977–991. http://dx.doi.org/10.1046/j.1365-2958.2003.03745.x.
62. **Roca AI, Cox MM.** 1990. The RecA protein—structure and function. Crit Rev Biochem Mol Biol **25:**415–456. http://dx.doi.org/10.3109/10409239009090617.
63. **Santi I, Scarselli M, Mariani M, Pezzicoli A, Masignani V, Taddei A, Grandi G, Telford JL, Soriani M.** 2007. BibA: a novel immunogenic bacterial adhesin contributing to group B Streptococcus survival in human blood. Mol Microbiol **63:**754–767.
64. **Horstmann N, Sahasrabhojane P, Suber B, Kumaraswami M, Olsen RJ, Flores A, Musser JM, Brennan RG, Shelburne SA, III.** 2011. Distinct single amino acid replacements in the control of virulence regulator protein differentially impact streptococcal pathogenesis. PLoS Pathog **7**(10)**:**e1002311. http://dx.doi.org/10.1371/journal.ppat.1002311.
65. **Cole JN, Barnett TC, Nizet V, Walker MJ.** 2011. Molecular insight into invasive group A streptococcal disease. Nat Rev Microbiol **9:**724–736. http://dx.doi.org/10.1038/nrmicro2648.

66. **Lupo A, Ruppen C, Hemphill A, Spellerberg B, Sendi P.** 2014. Phenotypic and molecular characterization of hyperpigmented group B streptococci. Int J Med Microbiol **304:**717–724. http://dx.doi.org/10.1016/j.ijmm.2014.05.003.
67. **Larsson C, Lindroth M, Nordin P, Stalhammar-Carlemalm M, Lindahl G, Krantz I.** 2006. Association between low concentrations of antibodies to protein alpha and Rib and invasive neonatal group B streptococcal infection. Arch Dis Child Fetal Neonatal Ed **91:**F403–F408. http://dx.doi.org/10.1136/adc.2005.090472.
68. **Kvam AI, Mavenyengwa RT, Radtke A, Maeland JA.** 2011. *Streptococcus agalactiae* alpha-like protein 1 possesses both cross-reacting and alp1-specific epitopes. Clin Vaccine Immunol **18:**1365–1370. http://dx.doi.org/10.1128/CVI.05005-11.
69. **Maeland JA, Bevanger L, Lyng RV.** 2004. Antigenic determinants of alpha-like proteins of *Streptococcus agalactiae*. Clin Diagn Lab Immunol **11:**1035–1039.
70. **Jeurink PV, van Bergenhenegouwen J, Jimenez E, Knippels LMJ, Fernandez L, Garssen J, Knol J, Rodriguez JM, Martin R.** 2013. Human milk: a source of more life than we imagine. Benef Microbes **4:**17–30. http://dx.doi.org/10.3920/BM2012.0040.

Table S1 – *Streptococcus agalactiae* samples used in this study – additional information.

Strain	Coverage[1]	Accession ID (reads)	Accession ID (assemblies)
Mother1	257	[EMBL:ERS500084]	[EMBL:ERS549891]
Child1	234	[EMBL:ERS500085]	[EMBL:ERS549892]
Mother2	314	[EMBL:ERS500086]	[EMBL:ERS549893]
Child2	212	[EMBL:ERS500087]	[EMBL:ERS549894]
Mother3	189	[EMBL:ERS500088]	[EMBL:ERS549895]
Child3	189	[EMBL:ERS500089]	[EMBL:ERS549896]
Mother4	231	[EMBL:ERS500090]	[EMBL:ERS549897]
Child4	151	[EMBL:ERS500091]	[EMBL:ERS549898]
Mother5-1	162	[EMBL:ERS500092]	[EMBL:ERS549899]
Mother5-2	167	[EMBL:ERS500093]	[EMBL:ERS549900]
Child5	181	[EMBL:ERS500094]	[EMBL:ERS549901]
Mother6	188	[EMBL:ERS500095]	[EMBL:ERS549902]
Child6-1	188	[EMBL:ERS500096]	[EMBL:ERS549903]
Child6-2	158	[EMBL:ERS500097]	[EMBL:ERS549904]
Mother7	203	[EMBL:ERS500098]	[EMBL:ERS549905]
Child7	163	[EMBL:ERS500099]	[EMBL:ERS549906]
Mother8	170	[EMBL:ERS500100]	[EMBL:ERS549907]
Child8-1	200	[EMBL:ERS500101]	[EMBL:ERS549908]
Child8-2	216	[EMBL:ERS500102]	[EMBL:ERS549909]
Mother9	241	[EMBL:ERS500103]	[EMBL:ERS549910]
Child9	154	[EMBL:ERS500104]	[EMBL:ERS549911]
Mother10-1	216	[EMBL:ERS500105]	[EMBL:ERS549912]
Mother10-2	210	[EMBL:ERS500106]	[EMBL:ERS549913]
Child10	202	[EMBL:ERS500107]	[EMBL:ERS549914]
Mother11	283	[EMBL:ERS500108]	[EMBL:ERS549915]
Child11-1	134	[EMBL:ERS500109]	[EMBL:ERS549916]
Child11-2	192	[EMBL:ERS500110]	[EMBL:ERS549917]
Mother12	217	[EMBL:ERS500111]	[EMBL:ERS549918]
Child12-1	268	[EMBL:ERS500112]	[EMBL:ERS549919]
Child12-2	94	[EMBL:ERS500113]	[EMBL:ERS549920]
Mother13	265	[EMBL:ERS500114]	[EMBL:ERS549921]
Child13	213	[EMBL:ERS500115]	[EMBL:ERS549922]
Mother14	258	[EMBL:ERS500116]	[EMBL:ERS549923]
Child14	221	[EMBL:ERS500117]	[EMBL:ERS549924]
Mother15	240	[EMBL:ERS500118]	[EMBL:ERS549925]
Child15-1	208	[EMBL:ERS500119]	[EMBL:ERS549926]
Child15-2	270	[EMBL:ERS500120]	[EMBL:ERS549927]
Mother16	166	[EMBL:ERS500121]	[EMBL:ERS549928]
Child16	72	[EMBL:ERS500122]	[EMBL:ERS549929]
Mother17	289	[EMBL:ERS500123]	[EMBL:ERS549930]
Child17	203	[EMBL:ERS500124]	[EMBL:ERS549931]
Mother18	236	[EMBL:ERS500125]	[EMBL:ERS549932]
Child18	218	[EMBL:ERS500126]	[EMBL:ERS549933]
Mother19	255	[EMBL:ERS500127]	[EMBL:ERS549934]
Child19-1	196	[EMBL:ERS500128]	[EMBL:ERS549935]
Child19-2	240	[EMBL:ERS500129]	[EMBL:ERS549936]
Child19-3	199	[EMBL:ERS500130]	[EMBL:ERS549937]

[1] - Coverage value indicated as fold-change in relation to the total size of the assembled genome

Table S2 - Genomic mutations detected in the population of each sample – additional information.

Sample[1]	Mutation[2]	% Mutant[3]				
		Mother (1)	Mother (2)	Child (1)	Child (2)	Child (3)
Mother 2	19C > T (Q7X)	100	-	0	-	-
Mother 2	642C > T	100	-	0	-	-
Mother 4	262A > G (K88E)	75,2	-	0	-	-
Mother 9	800_801ins (24 bp)			-		
Child 9	236G > A (R79H)	0	-	86,2	-	-
Mother 10-1	-266A > C	100	0	0	-	-
Child 11-2	213_214insA			-		
Child 11-2	524C > A (A175D)	0	-	0	100	-
Mother 12	-29A > G	100	-	13,9	0	-
Child 12	146C > T (A49V)	0	-	85,1	100	-
Child 12	2440G > A (G814S)	0	-	89,7	100	-
Child 12-1	72C > A	0	-	45,6	3,1	-
Child 12-1	603_620del (18 bp)			-		
Child 12-1	del (1132 bp)*			-		
Child 12-2	-132C > T	0	-	37,9	100	-
Child 12-2	557C > T (R186K)	0	-	26,2	100	-
Mother 14	-12G > T	100	-	0	-	-
Mother 14	89G > T (G30V)	100	-	0	-	-
Child 14	310_357dup (46 bp)			-		
Mother 15	-30delA			-		
Mother 15	548T > C (I183T)	100	-	0	0	-
Mother 15	G > A*	100	-	0	0	-
Mother 18	358A > C	90,5	-	0	-	-
Mother 18	1186C > T (P396S)	88,6	-	34	-	-
Mother 19	238C > T (P80S)	100	-	0	0	0
Child 19	286G > T (E96X)	4,7	-	100	100	100
Child 19	344C > T (A115V)	5,7	-	100	100	100
Child 19	-104_-103insA			-		

[1] – Sample in which the mutant allele was most frequently detected; [2] – Representation of the mutation identified. Numbering denotes the position of the mutation in relation to the +1 nucleotide of the translational initiation codon of the affected gene. For non-synonymous SNPs and INDELs larger than one bp, the amino acid change and the size of the inserted/deleted fragment, respectively, is indicated in parenthesis; [3] – Percentage of the mutant variant identified in the sampled population from each individual of the corresponding pair. For pairs with multiple samples per mother or child, columns are ordered from left to right in accordance to the order of the samples indicated in Table 1 and S1. A hyphen indicates that no frequency estimate was performed. * - The location of the mutation is not indicated given that no homology was found with any annotated GBS genome.

Table S3 - Primer pairs used to perform the RT-qPCRs.

Primer	Sequence	Tm (ºC)	GC%	Length (bp)	Gene
CovR_F	ACGTGAATTTGATTTGTTGAATGTC	58	32	105	*covR*
CovR_R	GTCTCTGCTGCCACATCGTA	60	55		
hvbiba_F	AGTTCTGATGAGGTTTGCCTT	57	43	174	*hvgA*/*biba*
hvbiba_R	TACCATCAACGGGTGAAGCC	60	55		
cylE_F	CCAGACGGTAGGCCTTTAACT	59	52	210	*cylE*
cylE_R	AGTGATTGCCTGTCCACTACG	60	52		
1895_F	AACTTGCTATGTGCGTTGGC	60	50	124	*gbs1895*
1895_R	TACGCTCTCACAGCTGCAT	59	53		
CAMP_F	AGTGACAACTCCACAAGTGGTAA	60	43	191	*cfb*
CAMP_R	GGTTGGCACGCAATGAAGTC	60	55		
GluPTS_F	CAGGTCAACCTGTAGCAGCA	60	55	146	*ptsG*
GluPTS_R	AGCAAAGCCATCACCCATCA	60	50		
Rib_F	ACCCTAAATGGGACGAGGGA	60	55	161	*rib*
Rib_R	GGCATCTGGGATTCGAGGTA	59	55		

Table S4 - API 50 CHL results.[1]

		Pair12				Pair15				Pair18				Pair19			
		Mother		Child*		Mother		Child		Mother		Child		Mother		Child	
		24h	48h	24h	48h	24h	48h	24h	48h	24h	48h	24h	48h	24h	48h	24h	48h
0		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1	Glycerol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	Grythritol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	D-arabinose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	L-arabinose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	D-ribose	+	+	+	+	-	+/-	+	+/-	+	+/-	+	+/-	+	+	+	+
6	D-xylose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	L-xylose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	D-adonitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	Methylxylopyranoside	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	D-galactose	+	+	+	+	+	+	+	+	+	+/-	+	+/-	-	-	-	-
11	D-glucose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
12	D-fructose	+	+/-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
13	D-Mannose	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
14	L-sorbose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	L-rhamnose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	Dulcitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	Inositol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	D-manitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	D-sorbitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	Methyl-aD-mannopyranoside	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	Methyl-aD-glucopyranoside	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	N-acetylglucosamine	+	+	+	+/-	+	+	+	+	+	+	+	+	+	+	+	+
23	Amygdalin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	Arbutin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	Esculin ferric citrate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	Salicilin	+/-	-	+/-	-	+	+	+	+	+/-	+/-	+/-	+/-	-	-	-	-
27	D-celobiose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	D-maltose	+	+/-	+	+	+	+	+	+	+	+/-	+	+/-	+	+	+	+/-
29	D-lactose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
30	D-melibiose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
31	D-sucrose	+/-	+/-	+	+/-	+	+	+	+	+	+/-	+	+	+	+	+	+
32	D-trehalose	+/-	+/-	+/-	+/-	+/-	+/-	+/-	+/-	-	-	-	-	+/-	+/-	+/-	+/-
33	Inulin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34	D-melezitose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35	D-raffinose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	Starch	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
37	Glycogen	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	Xylitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
39	Gentiobiose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
40	D-turanose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41	D-lyxose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
42	D-tagatose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
43	D-fucose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
44	L-fucose	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
45	D-arabitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
46	L-arabitol	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
47	Potassium gluconate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
48	Potassium 2-ketogluconate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
49	Potassium 5-ketogluconate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

[1] – Results were recorded after 24 and 48h of incubation at 37ºC. A positive reaction, corresponding to a color change from blue to yellow is indicated with "+", to green with "+/-" and negative reactions with "-". * - An isolated strain from Child 12-2 was used for this assay.

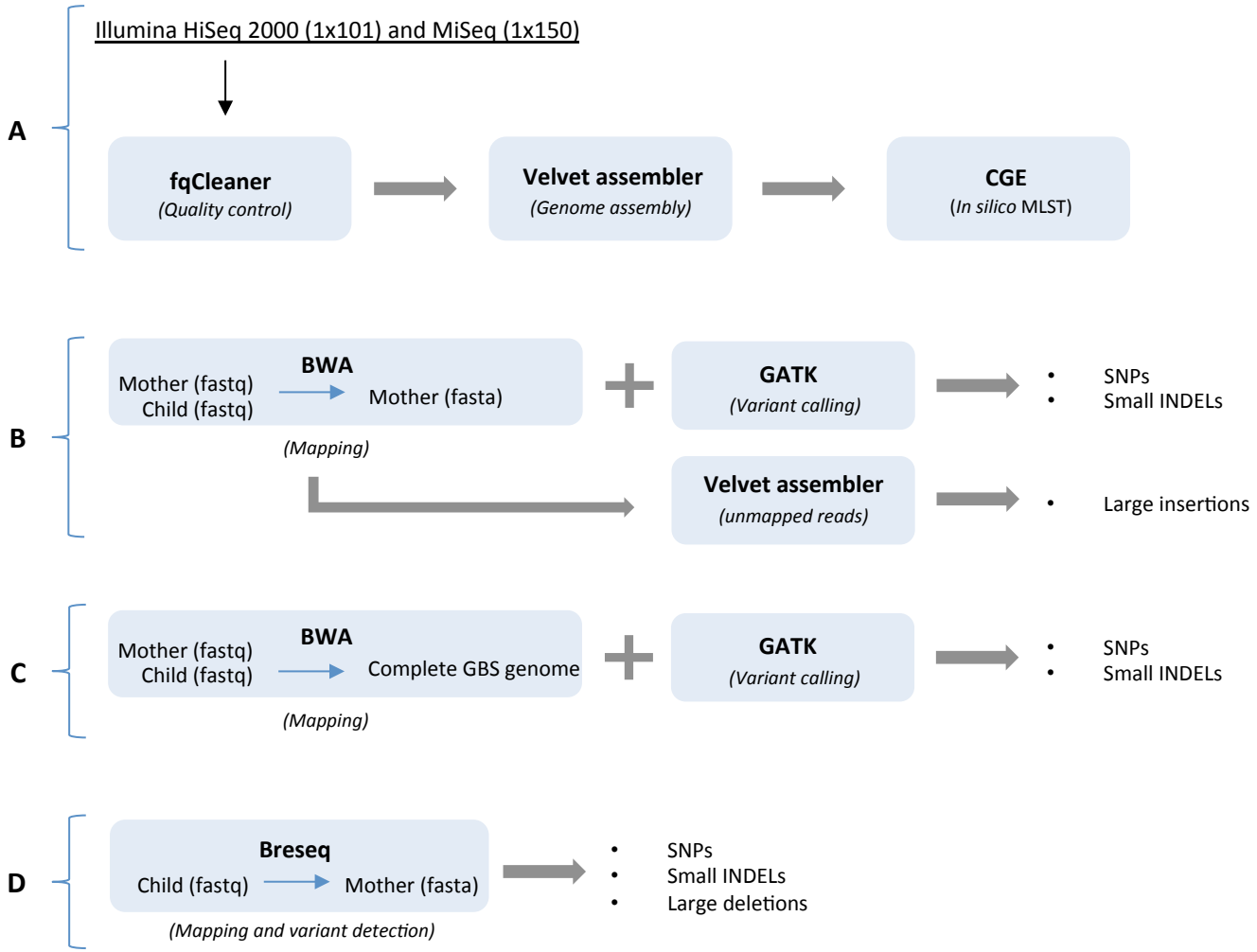


Figure S1 – Bioinformatics pipelines.
Strategy used for the genomic comparison of each set of mother-child GBS populations, as described in the "Methods" section. (A) Sequencing, quality control of reads, genome assembly and genotyping. (B) Mapping of each child sample against the assembled genome of the mother GBS. (C) Mapping of both samples from each pair to their closest-related complete genome of GBS. (D) Comparison of strains from each pair using the breseq computational tool (45).

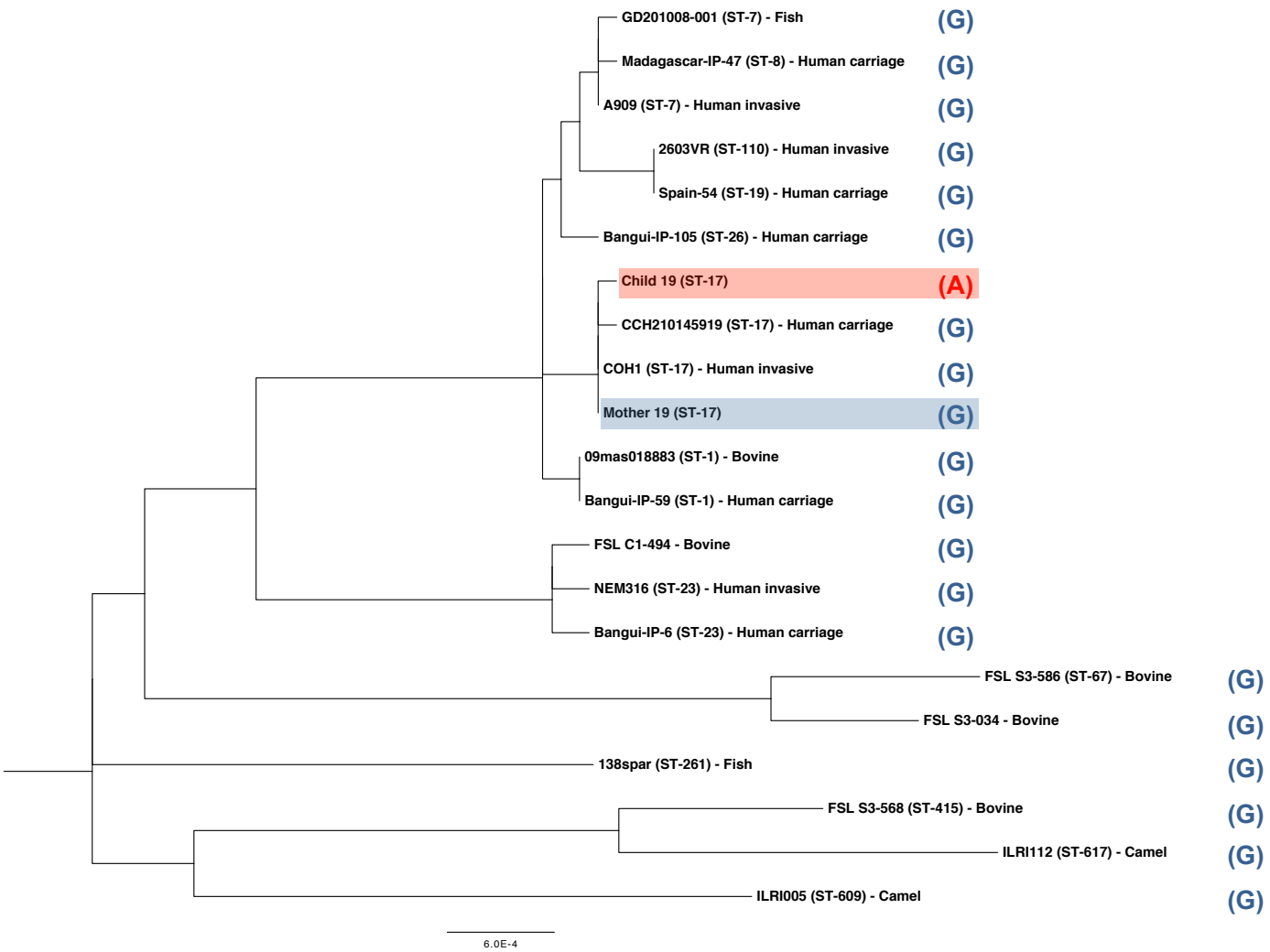


Figure S2 – Phylogenetic strategy to infer the direction of mutation.
Maximum-likelihood tree of a 10-kb sequence fragment encompassing the *covR* gene as an example to illustrate the methodology used to distinguish between the ancestral and mutant variants of each mutated region. Tree depicts the phylogeny of GBS pair 19 together with sequenced GBS genomes of various origins, indicated next to each strain. The ancestral (G) or mutant variant (A) identified in each genome is highlighted in blue and red, respectively. Here, the mutant variant was only detected in the child sample. Separate analyses using this strategy were performed for all the other mutations.

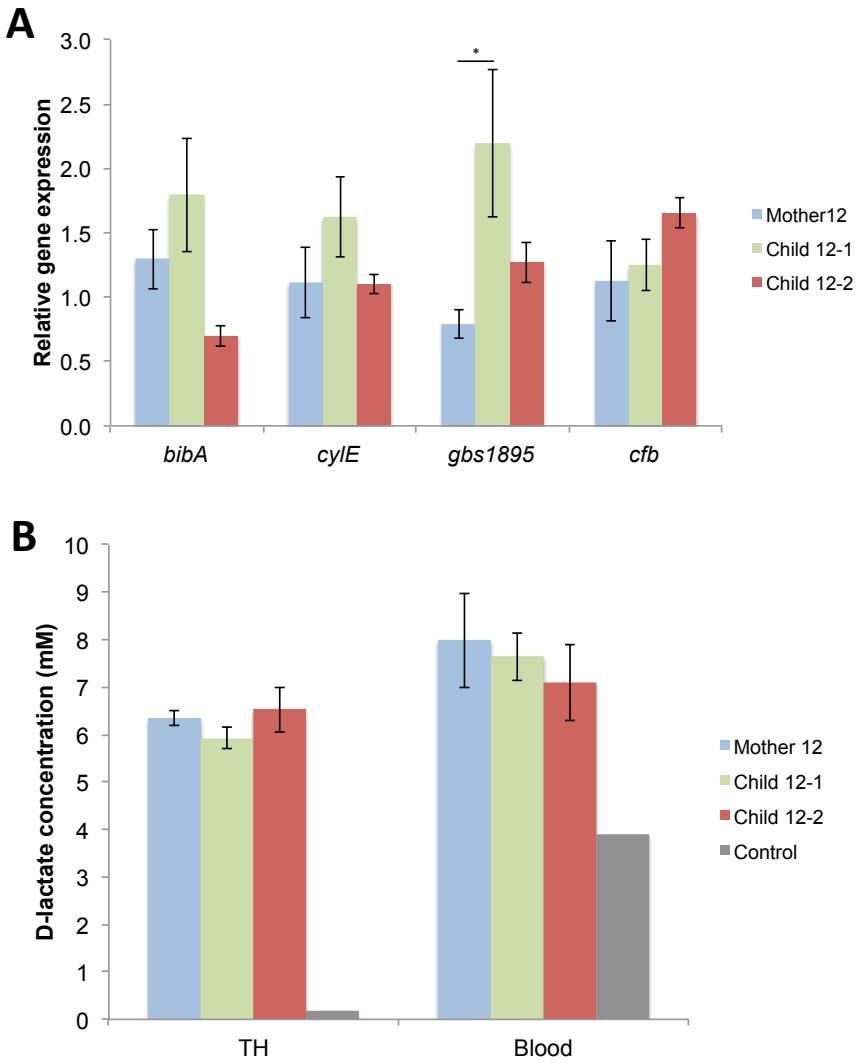


Figure S3 – Additional experimental analyses of pair 12.
(A) RT-qPCR results obtained with three cultures of GBS strains from pair 12 after incubation for 1h in human blood from a second donor. (B) D-lactate concentration measured in the supernatant of the pair 12 cultures, grown in TH and whole human blood, which were used for the quantitative PCR. Control samples represent D-lactate concentration in the growth medium without GBS. Error bars represent standard deviation (SD) +/-. *$p < 0.05$

Manuscript nº1

Parallel evolution of the hypervirulent clonal complex 17 of group B *Streptococcus* during carriage and disease

Alexandre Almeida[1, 2, 3], Isabelle Rosinski-Chupin[1, 2], Céline Plainvert[4, 5, 6, 7, 8], Pierre-Emanuel Douarre[1, 2], Maria J. Borrego[9], Claire Poyart[4, 5, 6, 7, 8] and Philippe Glaser[1, 2, *]

[1]Institut Pasteur, Unité Ecologie et Evolution de la Résistance aux Antibiotiques, Paris, France; [2]CNRS UMR 3525, Paris, France; [3]Université Pierre et Marie Curie, Paris, France; [4]Service de Bactériologie, Centre National de Référence des Streptocoques, Groupe Hospitalier Paris Centre Cochin-Hôtel Dieu-Broca, Assistance Publique Hôpitaux de Paris, France; [5]DHU "Risques et Grossesse", Assistance Publique Hôpitaux de Paris; [6]INSERM, U1016, Paris, France; [7]CNRS (UMR 8104), Paris, France; [8]Université Paris Descartes, Sorbonne Paris Cité, Paris, France, [9]National Institute of Health, Lisbon, Portugal.

[*]Corresponding author:
Philippe Glaser,
Institut Pasteur, 28 Rue du Dr Roux, 75724 Paris Cedex 15,
Tel + 33 (0)1 45 68 89 96 E-mail: pglaser@pasteur.fr

Abstract

Group B *Streptococcus* (GBS) is a commensal of the gastrointestinal and genitourinary tracts, while a prevailing cause of neonatal disease worldwide. Of the various clonal complexes (CCs), CC17 is over-represented in GBS-infected newborns, particularly in late-onset disease. Here, we report a comprehensive analysis of 612 CC17 genomes collected worldwide to investigate the genetic traits behind their successful adaptation to humans, and the underlying differences between carriage and clinical strains. We reveal that modification of pathways related to metabolism, adhesion, regulation and immune evasion represent various alternative adaptive strategies of CC17 strains in the human host. The most distinctive features of disease-specific isolates were frequent mutations in the virulence-associated CovRS and its regulator Stk1. Strikingly, convergent adaptation of the *neuD* gene involved in capsular sialylation was detected in the context of late-onset disease, evidencing a possible clinical association. Phase variation of the surface protein Rib also distinguishes disease from carriage isolates, revealing the presence of different immune selective pressures. We propose that clinically associated variants acquired among CC17 strains, combined with their unique genetic repertoire, determine their ability to cause disease in humans.

Introduction

During the mid-20th century, *Streptococcus agalactiae* (group B *Streptococcus*, GBS) emerged as one of the main etiological agents of neonatal disease worldwide. Neonatal infections are classified into early- and late-onset disease (EOD and LOD), according to whether they occur before or after the first week of life. EOD is usually manifested in the form of pneumonia or sepsis, whereas LOD frequently progresses to meningitis[1]. In spite of current prophylactic measures, GBS continues to be a prevalent cause of neonatal morbidity and mortality, especially in LOD, with an average incidence rate of 0.53 infections per 1000 live births[2].

Of the ten capsular (CPS) types, and various clonal complexes (CC) defined by multilocus sequence typing (MLST), it has been recurrently shown that there is a strong association between serotype III strains, and CC17 in particular, with neonatal disease[3,4]. Therefore, this hypervirulent clone has been the focus of genetic and functional studies aimed at identifying its unique virulent traits. One of the most remarkable findings was that CC17 strains harbour a specific adhesin, known as the hypervirulent GBS adhesin (HvgA), that facilitates crossing of the blood-brain barrier[5]. Moreover, unique variants of the serine-rich repeat protein (Srr2), and its SecA2/Y2 secretion system, have also been shown to promote the adhesion of CC17 to human epithelial cells[6]. When analysing the diversity of the alpha-like family of surface proteins in GBS, it was also observed that CC17 strains only carry a particular variant known as Rib[7]. Notably, this surface protein has been reported as an important contributor to the virulence potential of GBS[8], and to elicit protective immunity[9]. Despite its virulent properties, the CC17 lineage is also found in individuals with no clinical symptoms, raising the question on whether there are strain-specific features differentiating disease- from carriage-related isolates[4,10]. Furthermore, it is also unclear why CC17 strains still prevail in the context of LOD, in spite of existing control and preventive measures.

Here, we performed a detailed analysis of 612 GBS genomes belonging to CC17, spanning multiple countries and isolation dates, collected from adults and newborns representative of different clinical states. We provide a greater understanding of the hypervirulence of CC17 by highlighting the evolutionary changes under natural selection

in the human host. The expansion of this clone was driven by mutations recurrently affecting metabolism-related genes, as well as those implicated in adhesion, regulation and immune evasion. We also show that disease-specific isolates acquire unique mutation patterns that may determine their distinctive propensity to cause disease in humans.

Results

Geographic distribution and clinical association

To obtain a detailed overview of the genomic diversity of CC17, we compiled all the publicly available genomes belonging to CC17 as of August 2016[4,10,11,12,13], together with a new panel of 45 strains, totalling a set of 612 GBS sequences (Supplementary Table 1). This comprises strains collected between 1955 and 2016 from Africa (*n* = 359), Europe (*n* = 131), North America (*n* = 95) and Australia (*n* = 3), as well as others from unknown origins (*n* = 24). For the majority (*n* = 572), we had access to clinical information about each colonized individual. A total of 306 strains were obtained from asymptomatic carriers, and 266 isolates from infected patients. Of the strains obtained from disease, 42% (*n* = 111) were from LOD, 20% (*n* = 52) from EOD and 15% (*n* = 41) from invasive disease in adults. As for the biological origin of the 612 isolates here studied, 44% (*n* = 268) were collected from the gastrointestinal and genitourinary tracts, 24% (*n* = 144) were retrieved from blood cultures, 5% (*n* = 33) from the cerebrospinal fluid and 27% (*n* = 167) were from other or unknown origins.

Intercontinental transmission and worldwide dissemination

First, to assess the structure of the CC17 population, we built a maximum-likelihood phylogenetic tree of the 612 CC17 genomes (Fig. 1), based on the core and recombination-free alignment of 12 431 single nucleotide polymorphisms (SNPs). We found that one of the novel strains included in this work (B83, isolated in 1970; Table 1) diverged before the main expansion of CC17 and the acquisition of the tetracycline resistance determinant *tetM* — reported as the main driver for the expansion of human GBS clones during the 1950s[10] — and thus represents a reliable outgroup for further analysis (Fig. 1). We did not detect any additional recombination events apart from the two previously described regions present in strains from cluster IV[10,14] (Fig. 1). The phylogenetic tree shows that the CC17 population is structured into four main clusters, with most isolates represented within two key lineages (clades I and II). Carriage and disease-related isolates are generally intermixed, with strains responsible for disease found across the four main clones. Likewise, isolates from LOD are similarly dispersed, suggesting that they diverged independently from distinct genomic backgrounds. Interestingly, there is no well-defined correlation between the phylogenetic distribution

and location of origin. Based on our sampling, transmission from Europe to North America seems to have occurred most frequently, with an average of 23 transitions predicted from the phylogeny (Fig. 2a). Also to a noticeable extent were exchanges between Africa and Europe, and from North America to Europe (Fig. 2a). This shows that CC17 strains are not geographically constrained, having disseminated worldwide through repeated travels across different continents following the expansion of the four main clones around the 1950s (Fig. 2).

Parallel adaptive evolution in the CC17 population

The accumulation rate and distribution of genomic mutations fixed in a population provide a record of evolutionary pressures driving adaptation. In order to detect genes with a mutational bias, we compared the expected number of mutations acquired per gene under a neutral Poisson model of evolution, with that observed in the CC17 population (Fig. 3). A total of 152 genes acquired significantly more mutations than expected ($P < 0.05$, exact Poisson test; Supplementary Table 2). Globally, pathways involved in nucleotide and amino acid metabolism were frequently affected ($P < 0.05$, Fisher's exact test; Fig. 3b), with major virulence-associated factors listed among the most mutated genes (Supplementary Table 2). CovS, the histidine kinase of the CovRS two-component regulator of virulence, acquired 25 independent mutations and had a statistically significant mutational bias after correcting for multiple testing (Supplementary Table 2). In line with this, the serine/threonine protein Stk1, which regulates CovRS, was also highly affected by accumulating a total of 24 mutations. Besides *covRS*-related regions, other notable genes particularly mutated included the *dltD* gene, part of the *dlt* operon responsible for D-alanylation of the cell wall; the fibrinogen-binding protein FbsA; the penicillin-binding protein DacA; and three genes from the *cps* operon (*cpsD*, *cpsE* and *cpsG*) involved in the biosynthesis of the capsule. Notably, the serine-rich repeat protein Srr2, unique to CC17 strains, was among the most significantly mutated genes, with 51 independent substitutions. We then looked for additional evidence of positive selection by searching for genes that accumulated more amino acid-modifying mutations than expected, using the canonical test of natural selection (dN/dS). Genes coding for FbsB (GBSCOH1_RS04170), the transcriptional regulator of the Pilus 1 genomic island (PI-1; GBSCOH1_RS03235) and a protein similar to Zoocin A (GBSCOH1_RS00315) presented a statistically significant signal of

diversifying selection (dN/dS > 1; $P < 0.05$) with acquisition of 18, 20 and 13 amino acid changes, respectively, relative to no synonymous substitutions. Both the FbsB and PI-1 proteins are implicated in cell adhesion, while Zoocin A is a streptococcolytic endopeptidase described in *Streptococcus equi* subsp. *zooepidemicus* as exhibiting penicillin-binding properties and weak β-lactamase activity[15]. Conversely, Srr2, despite being a mutational hotspot, was under strong purifying selection (dN/dS = 0.28, $P < 0.001$), meaning that most nonsynonymous mutations have been purged, presumably to preserve the functional integrity of the gene.

A similar evolutionary analysis was performed to look for recurrent mutations within non-coding sequences, and we identified a group of 86 regions with a mutational bias ($P < 0.05$, exact Poisson test; Supplementary Table 3). Among the most highly mutated regions were sequences upstream two CovR-regulated targets: the hypervirulent GBS adhesin HvgA (GBSCOH1_RS09645) and the hypothetical secreted protein GBSCOH1_RS08830. The upstream region of *cspA* (*GBSCOH1_RS02410*), a serine protease with the ability to cleave fibrinogen and inactivate specific chemokines, was also recurrently mutated. Importantly, 14 mutations were identified upstream the *fbsA* and *srr2* genes, together with 21 mutations upstream *fbsB* (Supplementary Table 3), further emphasizing a potential fitness advantage in the regulation of these systems.

Mutations leading to truncation or disruption of the coding sequence potentially have the strongest functional impact, as the structure and function of the protein product is more likely to be affected. We identified a total of 434 nonsense mutations and 696 frameshifts acquired by the CC17 isolates. Of the 667 genes affected, 14 acquired more than five inactivating mutations (Supplementary Table 4). Of note were eight mutations in the C5a peptidase, a protein usually conserved among human GBS isolates. Given the low number of nonsense and frameshift mutations accumulated per gene, we further gauged their relative importance by determining whether they were acquired in more divergent lineages representative of a longer evolutionary period since their fixation within the CC17 population. The most ancestral nonsense mutation, acquired by the common ancestor of clades I and II (Fig. 1), was in an ABC transporter ATP-binding protein (GBSCOH1_RS03295) with a domain similar to DrrA of the daunorubicin efflux system.

Genomic differentiation between carriage- and disease-related isolates

From the core-genome analysis of the CC17 population (Fig. 1), we reasoned that the terminal branches corresponding to the most recent evolutionary changes underlie the distinction between colonizing and infectious CC17 strains, and might represent a possible genetic basis for the pathogenesis of disease-related isolates. To address this hypothesis, we devised a strategy to extract the mutations exclusively acquired by carriage or disease isolates (see Methods section). A total of 5824 mutations were uniquely identified among isolates collected from carriers, while 6381 substitutions were exclusively detected in strains from infected patients (Supplementary Table 1). Strikingly, of the disease-specific mutations, we identified nine homoplasic substitutions affecting genes directly associated with virulence (Table 1). Homoplasy, i.e. the acquisition of the same mutation by independent lineages, is the strongest evidence of convergent adaptation as a response to similar evolutionary pressures. This remarkable signal of adaptive evolution was observed in regions related to cell adhesion (*fbsA*, PI-1 and *rib*), immune evasion (*neuD*, *neuC*, *cpsE* and *cpsL*) and virulence regulation (*covS*; Table 1).

After correlating the number of mutations acquired per gene between carriage and disease isolates, we searched for outlier genes that had a mutation pattern more associated with one of the two clinical states (Fig. 4). Variation within 15 genes differed significantly ($P < 0.05$, Bonferroni-adjusted outlier test) between carriage and infection scenarios (Table 2 and Fig. 4). The gene with mutations most associated with disease-causing strains codes for an ABC transporter permease (GBSCOH1_RS04975) and is located alongside another encoding an ATP-binding protein with similarities to ABC exporter systems. Other genes with mutations enriched in infection encode the serine/threonine protein kinase Stk1 and the CovS two-component sensor histidine kinase (Table 2). This indicates that both genes are among the most highly mutated in the CC17 population because of a mutational bias predominantly within clinical isolates. The CovRS system is known to regulate the haemolytic activity of GBS and the production of an orange pigment through the *cyl* operon[16,17]. By looking at the degree of pigmentation in a panel of 18 strains — nine mutants and nine others not affected in genes potentially affecting the expression of CovRS (*covR*, *covS*, *abx1* and *stk1*) — we observed extreme levels of pigmentation (≤ 0.09 and ≥ 0.97; Supplementary Fig. 1)

across different strains that acquired nonsynonymous substitutions in *covS* or *stk1*. Moreover, two of the least pigmented strains acquired independent SNPs affecting the same amino acid within *covS* (Trp297; Supplementary Table 1). Hence, we speculate that these mutations might have an impact on the phenotype of CC17 strains through modulation of CovRS activity. In contrast, we found four genes more significantly mutated in carriage-associated isolates (Table 2). This was most evident in the gene coding for the ScpB C5a peptidase, which acquired 21 mutations in carriage isolates, compared to nine mutations in disease-related strains, suggesting that modifications within this protein are preferentially selected during colonization.

Analysing the unique genetic repertoire of CC17

The surge of complete GBS genomes currently available provides the opportunity to gain a better understanding of the genetic makeup of this species. Complementing the SNP-based evolutionary analysis of the CC17 population, we compared the gene content of a diverse set of GBS genomes to identify regions that might be associated with the hypervirulent properties of CC17. By analysing the pan-genome across 32 complete GBS genomes belonging to 17 different sequence types (STs), we identified a total of 1553 core genes (present in ≥ 95% of the strains) and 2590 accessory genes. Of those from the accessory part, 80 genes were acquired specifically by CC17 genomes and up to one other ST, while a group of 23 genes are present in most CCs but absent from CC17 in particular (Fig. 5 and Supplementary Table 5). A remaining set of 27 genes is distributed among a few different clones alongside CC17, reflecting a more ambiguous history of genetic gain and loss (Fig. 5 and Supplementary Table 5). Looking at the functions potentially gained by horizontal exchange, we confirmed the acquisition of *hvgA*, *srr2* and *rib* by CC17, together with several genes coding for proteins with unknown function (Supplementary Table 5). Of those, we identified putative cell wall surface proteins, proteases and lipoproteins. This suggests that beyond the virulent traits already characterized in CC17, other genes were also exclusively selected in this lineage, and might contribute to its evolutionary success in the human host.

As antibiotic resistance is known to be involved in the successful dissemination of epidemic clones, we probed the presence of antibiotic resistance determinants in the CC17 genomes (Supplementary Fig. 2). The *tetM* gene conferring tetracycline resistance

is widespread across most (96%) of the CC17 isolates, which emphasizes both its crucial role in the dissemination of these strains and its stable integration (Supplementary Fig. 2). Besides tetracycline resistance, resistant traits against macrolides and aminoglycosides were the most abundant, but to a much smaller degree (Supplementary Fig. 2). These resistant genes are also phylogenetically constrained, having been acquired only by specific clades within the CC17 population. This suggests that they might have been independently selected in particular contexts but were not determinant to the successful spread and hypervirulence of the CC17 lineage.

Phase variation in the surface protein Rib

Diversity within a population not only stems from nucleotide-level mutations (SNPs or indels), but might also result from the presence of gene copy number variation (CNV), which ultimately affects fitness and evolutionary outcomes. After normalizing the coverage distribution with that of the reference strain COH1 isolated from disease, we identified 128 genes that displayed at least double the coverage in at least one strain. After excluding those with unknown functions or related to mobile genetic elements, we ended up with 16 functionally relevant genes (Fig. 6). Although the number of copies of these 16 candidates varied across the phylogeny, that of the gene coding for the alpha-like surface protein Rib was markedly increased in various isolates (Fig. 6). In fact, we found that the coverage estimated for *rib* was a direct correlation of the number of tandem repeating motifs following the unique N-terminal part. There was no association between its coverage and the phylogenetic structure of the CC17 population, reflecting a flexible adaptability typical of phase variation. However, by comparing carriage-related isolates with disease-associated strains, we observed a significant reduction ($P < 0.001$, two-tailed *t* test) in the number of repeat units in Rib among strains from infection (Fig. 6b).

Investigating the persistence of GBS in neonates

To gain insights into a potential genetic basis for the persistence of CC17 in the newborn, we characterized the adaptive changes inferred from longitudinal samples collected one-month apart from three different newborns with GBS-positive blood cultures (Supplementary Fig. 3). The first isolates were obtained within the first week of life, representative of EOD syndrome, while the second set of samples was collected within

the LOD time-window. In all cases, isolates obtained from both time points were genotypically related, signifying that the relapsed infection of the bloodstream was the result of GBS strains deriving from the ones originally infecting the newborn. After the one-month period, we detected the acquisition of one point mutation per sample (Supplementary Fig. 3). These corresponded to one nonsynonymous substitution in the gene encoding an oligopeptide ABC transporter similar to OppA (GBSCOH1_RS04855), another in the *neuD* gene (GBSCOH1_RS05705), and a synonymous mutation in a gene with unknown function (GBSCOH1_RS04720). The *opp* operon has been shown to play an important role in the pathogenesis of *Streptococcus pyogenes*[18], while *neuD*, whose transcription has been described as being higher in strains from the cerebrospinal fluid[19], is involved in capsular sialylation. Remarkably, the missense mutation in *neuD* (Thr67Ala) is one of the homoplasic sites independently mutated in an unrelated CC17 strain (Table 1). In this case, the other mutation was acquired by a strain associated with meningitis (K38783; Supplementary Table 1), supporting a possible link between this amino acid replacement in *neuD* with the incidence of LOD and meningitis. Also notably was that for two of the pairs of strains, the number of repeats in the gene coding for the Rib surface protein was higher in the original EOD isolate (Supplementary Fig. 3), further evidencing the presence of evolutionary pressures in the human host selecting for the reduction of Rib repeats during infection.

Discussion

The hypervirulent CC17 is the most prevalent cause of GBS neonatal disease worldwide, but the reasons underlying its persistence remain poorly understood. In this work, we leveraged the genomic data obtained from 612 CC17 strains to track the parallel evolution of multiple clinical isolates, and to characterize the distinguishing traits between carriage- and disease-specific lineages.

Genomic analyses of a wide range of GBS strains had shown that most of the major human CCs were selected in the 1950s after acquisition of the tetracycline resistance determinant *tetM*[10]. Moreover, their evolution was shown to have been primarily mediated by the exchange of large chromosomal regions[10,20]. In the case of CC17, limited diversity within the core genome suggests that recombination did not play a crucial role in the diversification of this clone. Therefore, aside from the acquisition of tetracycline resistance, any additional adaptive traits were introduced mainly through incremental and vertically acquired mutations. We identified a recurrent bias in the frequency and type of genomic mutations affecting genes with pivotal roles in GBS, namely for nucleotide and amino acid metabolism, adhesion, immune evasion and regulation (Fig. 3 and Supplementary Table 2). Likewise, the mutation pattern between carriage- and disease-associated isolates was most distinct among some of the same regions with an overall mutational bias in the CC17 population (Table 2 and Supplementary Table 2). This means that there are various adaptive strategies successful in the human host that may additionally have a potential clinical association. The presence of multiple targets of adaptation also underscores the multifactorial pathogenesis of GBS infections, in that disease outcome is determined by the intricate relationship between host- and pathogen-specific properties.

Of the SNPs detected in genes involved in cell adhesion, those present in the PI-1 genomic island may illustrate a general mechanism of adaptation in GBS. Indeed, mutations in the PI-1 transcriptional regulator were also predicted to have been positively selected in the ST1 lineage[21]. By contrast, diversifying selection in the *fbs* genes has not been reported in other GBS clones, implying a unique advantage to CC17-specific strains. Importantly, both *fbsA* and *fbsB* have been described as having a

synergistic effect for the binding ability of GBS to fibrinogen specifically within the CC17 genomic background[22].

The alpha-like surface protein Rib, while also implicated in adhesion and invasion[8], has highly immunogenic properties and has been considered a potential vaccine target against GBS[9]. Interestingly, we identified a strong correlation between the number of repeated domains in Rib with the clinical state of the GBS-colonized individual (Fig. 6 and Supplementary Fig. 3). The immunogenicity of Rib is inversely correlated with the number of repeats[23]. Therefore, we can conclude that strains with a smaller-sized Rib may be able to more readily evade the immune response of the host, and be strongly selected in the course of infection, as was previously observed in a mouse model of infection[24]. Similarly, we had formerly reported that a reduction in the expression of Rib could present a selective advantage in neonatal disease following maternal transmission[11].

One of the most notable virulence factors known to contribute to evasion of the host immune response is the capsule polysaccharide. We identified recurrent polymorphisms in the *neu* and *cps* operons that control the biosynthesis and immune evasion properties of the capsule. Notably, we highlight a possible clinical association between a specific nonsynonymous mutation in *neuD* with patients diagnosed with LOD and meningitis (Table 1 and Supplementary Fig. 3), two intrinsically related syndromes.

Controlling the expression of many virulence-associated genes is the two-component CovRS system[25]. We identified frequent substitutions in the CovS sensor histidine kinase predominantly among disease-associated strains (Tables 1 and 2). Moreover, a mutational bias was detected in the serine/threonine kinase Stk1, and in the upstream regions of the CovR-regulated *hvgA* and *GBSCOH1_RS08830* genes. Stk1 mediates CovRS activity through phosphorylation of the CovR response regulator[26], which is able to bind to the promoter of several virulence-associated genes (e.g. *hvgA*), repressing their transcription. Therefore, CovRS represents a key target for adaptation, as slight genomic changes can have dramatic phenotypic repercussions[27,28]. We thus predict that the abundance of mutations acquired in genes affecting the regulation of CovRS is indicative of parallel adaptive evolution and of their global impact on the host interaction of GBS.

Although we stress the importance of genomic variation within traits for which there is extensive knowledge, our work also paves the way for further characterization of genes not previously implicated in virulence. Such examples are the regions coding for the Zoocin A protein and an ABC exporter permease. The former has only been studied in *S. zooepidemicus* and was among the three genes with a statistically significant bias of nonsynonymous substitutions, while the latter was one of the most distinctly mutated genes in disease-associated isolates (Table 2). Pan-genome analyses of the GBS population also uncovered the acquisition by CC17 strains of a number of genes with uncharacterized functions that may play a contributing role to their hypervirulence (Fig. 5 and Supplementary Table 5).

This work underscores the value of using whole-genome and population approaches to track how microbial species adapt to their environment, and to ultimately deduce how genomic variation might influence their ability to cause disease. We present a thorough analysis of the adaptive evolution of CC17 and investigate the genetic traits behind its persistence and pathogenesis in the human host. We provide a greater understanding of the evolutionary pressures acting on CC17 *in vivo* and of its parallel evolution in a clinical setting. Identifying the genes associated with pathogenic phenotypes will potentially contribute to more informed treatment decisions and steer new therapeutic directions for the control of GBS neonatal infections.

Methods

Bacterial strains

A total of 612 GBS genomes were analysed in this work (Supplementary Table 1). Among those, 567 genomes were obtained by surveying public databases for all the available GBS sequences belonging to CC17. In brief, an *in silico* MLST analysis was performed on all the accessible GBS genomes, as of August 2016, using either SRST2[29] on the raw sequencing reads, or a BLAST-based python script on the assemblies. For the purpose of this work, strains with no more than one allelic difference in regards to the MLST profile of ST17 were considered CC17. A total of 547 genomes were selected and retrieved from the published studies of Almeida et al.[11] (*n* = 25), Da Cunha et al.[10] (*n* = 79), Rosini et al.[12] (*n* = 18), Seale et al.[4] (*n* = 333) and Teatero et al.[13] (*n* = 92), together with 20 additional genomes deposited in the NCBI database (Supplementary Table 1). A set of 45 strains were also added and sequenced in this study, provided by the Centre National de Référence des Streptocoques (CNR-Strep) in France, the Collection of Institut Pasteur (CIP) and the Paediatric Hospital of Luanda[30] (Supplementary Table 1).

Whole-genome sequencing, assembly and pan-genome analysis

Chromosomal DNA extraction of the novel set of 45 isolates was performed with the DNeasy blood and tissue kit (Qiagen). Libraries were prepared with the Nextera XT protocol, and genomes were sequenced using the Illumina HiSeq 2500 platform with paired-end read runs of ~ 150 bp. Reads were filtered for quality and then assembled using either Velvet[31] or SPAdes[32]. Strain BM110 (Supplementary Table 1) was additionally selected for single molecule real-time sequencing (PacBio RS II system). PacBio subreads were assembled with both Canu[33] and the RS_HGAP_Assembly.3 protocol from the SMRT analysis toolkit v2.3, while consensus accuracy was further polished using Quiver[34], as was previously described[35]. Genome assemblies were annotated with Prokka[36] using a custom database comprising reference sequences (RefSeq) of GBS and other streptococci. Pan-genome analysis of 32 available complete genomes of GBS, belonging to various CCs, were carried out using Roary[37]. Genes with a specific association with CC17 genomes were identified with the Scoary script (https://github.com/AdmiralenOla/Scoary). Detection of antibiotic resistance genes was performed with SRST2[29] on the sequencing reads, and with the Large Scale BLAST

score ratio (LS-BSR) pipeline[38] on the genome assemblies, using the ResFinder database[39]. Each antibiotic resistance gene was considered present if detected in both the raw sequencing data and the assembled genome.

Sequencing reads from the 45 newly sequenced strains, as well as the complete assembly of strain BM110, have been deposited in the EMBL nucleotide sequence database (http://www.ebi.ac.uk/ena) under the study accession number PRJEB18603.

Genome mapping, variant calling and phylogenetic analyses

We used Burrows-Wheeler Aligner (BWA)[40] to map the sequencing reads of each CC17 genome against the complete reference sequence of COH1. Copy number variation of each gene was deduced with the R package CNOGpro[41], based on the BWA mapping of each strain after normalization with that of simulated reads from the reference genome of COH1. Variant calling was performed with Genome Analysis ToolKit (GATK) v3.4.0[42] according to the published recommendations[43,44].

Phylogeny of all CC17 strains was inferred from the polymorphic positions detected in the variant calling workflows, while the phylogenetic tree comprising the 32 complete GBS genomes was based on the core-genome alignment obtained with Parsnp[45]. Recombinant sites and accessory genes — missing from more than 1% of the isolates — were removed following identification with Gubbins[46] and the filter_BSR_variome.py script from the LS-BSR pipeline[38], respectively. Maximum-likelihood (ML) phylogenies were built with RAxML[47], with a General Time-Reversible (GTR) substitution model with a gamma-distributed rate across sites combined with an ascertainment bias.

To investigate the temporal evolution of CC17, the Bayesian phylogenetic software BEAST v2.3.1[48] was used. The evolutionary rate of this population was calibrated with the corresponding sampling date of each strain (Supplementary Table 1) as previously described[10]. To characterize the phylogeographic distribution of the CC17 strains, and infer possible events of intercontinental transmission, we then used the make.simmap tool[49] from the phytools R package[50]. Discrete ancestral traits, matching one of the geographical locations of the samples, were predicted for each node of the CC17 phylogeny through modelling of 100 simulations. The resulting number of

intercontinental transitions between nodes and from node to tip was calculated using the count.simmap function[49].

Variant annotation, parallel evolution and mutation classification

To predict the impact of the SNPs detected within coding sequences, snpEff[51] was used to classify each point mutation as either nonsynonymous or synonymous. As for indels, only frameshift mutations were taken into account. Without recombination and assuming a constant mutation rate across genes, the number of substitutions per gene under neutral evolution can be modelled as a Poisson distribution. Therefore, a signal of parallel evolution was inferred from a statistically significant increase in substitution rate over that expected under a null hypothesis of neutral evolution, as previously described[52]. Multiple testing correction was performed with a false discovery rate (FDR) of 10%. Genes with a mutational bias were classified into functional categories using the eggNOG database v4.5[53]. We then used a Fisher's exact test to assess the statistical significance of the functions affected in relation to their overall proportion in the COH1 reference strain. For dN/dS calculations, the observed spectrum of N and S mutations per gene was normalized by the expected ratio of nonsynonymous to synonymous sites obtained through simulation of all possible nucleotide substitutions. Values above 1 are indicative of positive selection and statistical significance was assessed using the binomial test.

Variants associated with the available metadata (Supplementary Table 1) were extracted using VCFtools[54]. Mutations were classified into carriage- or disease-related based on whether they were exclusively present in GBS isolates collected either from infected individuals or from asymptomatic carriers. This involved removing mutations that arose in ancestral lineages common to both carriage and disease strains. Subsequently, for each locus of the genome of COH1, the total number of mutations classified into each clinical state was calculated. To investigate the genes most differentially mutated, a linear model of correlation was built between carriage and disease mutation patterns, and the outlier genes were detected using a Bonferroni-adjusted outlier test.

GBS pigmentation

In order to assess the degree of pigmentation of each GBS strain, bacterial cells were cultured overnight in TH broth at 37ºC in standing cultures. Subsequently, 7 ul of each strain were spotted on Granada Agar plates (bioMérieux) and grown overnight at 37ºC in a CO_2 environment. Pictures were taken with plates against a black background, and contrast was adjusted to more easily discriminate weak from strong pigment producers. The level of pigmentation was measured by the colour intensity of each spot, quantified within a circle of the same area size using imageJ (https://imagej.nih.gov/ij). Values were then normalized against the sample with the highest intensity in each test to generate a ratio from 0 to 1. Experiments were performed in quadruplicate, with the resulting average and standard deviation represented in Supplementary Fig. 1.

Acknowledgements

This work was supported by a project of ANR LabEx IBEID and ANR-13-PRTS-0006-04. Sequencing was performed at the Pasteur Genopole, a member of France Génomique (ANR10-IBNS-09-08). A.A. is a scholar in the Pasteur-Paris University (PPU) International PhD program and received a stipend from the ANR LabEx IBEID. The authors thank Laurence Ma for her help in performing the Illumina sequencing.

Author contributions

A.A. performed the bioinformatics analyses and the experimental tests, analysed the data and wrote the manuscript. I.R-C. and P-E.D. performed bioinformatics analyses and revised the manuscript. C.Plainvert and M.J.B. provided isolates and revised the manuscript. C.Poyart selected and provided isolates, and revised the manuscript. P.G. supervised the work, participated in the design of the bioinformatics pipelines and experiments, helped analyse data and wrote the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing financial interests.

References

1. Le Doare K, Heath PT. An overview of global GBS epidemiology. *Vaccine* **31**, D7-D12 (2013).
2. Edmond KM*, et al.* Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. *Lancet* **379**, 547-556 (2012).
3. Bisharat N*, et al.* Population structure of group B streptococcus from a low-incidence region for invasive neonatal disease. *Microbiol-Sgm* **151**, 1875-1881 (2005).
4. Seale AC*, et al.* Maternal colonization with *Streptococcus agalactiae* and associated stillbirth and neonatal disease in coastal Kenya. *Nat Microbiol* **1**, 16067-16067 (2016).
5. Tazi A*, et al.* The surface protein HvgA mediates group B *Streptococcus* hypervirulence and meningeal tropism in neonates. *J Exp Med* **207**, 2313-2322 (2010).
6. Six A*, et al.* Srr2, a multifaceted adhesin expressed by ST-17 hypervirulent group B *Streptococcus* involved in binding to both fibrinogen and plasminogen. *Mol Microbiol* **97**, 1209-1222 (2015).
7. Brochet M*, et al.* Genomic diversity and evolution within the species *Streptococcus agalactiae*. *Microbes Infect* **8**, 1227-1243 (2006).
8. Baron MJ, Filman DJ, Prophete GA, Hogle JM, Madoff LC. Identification of a glycosaminoglycan binding region of the alpha C protein that mediates entry of group B streptococci into host cells. *J Biol Chem* **282**, 10526-10536 (2007).
9. Stalhammarcarlemalm M, Stenberg L, Lindahl G. Protein Rib - a novel group-B streptococcal cell-surface protein that confers protective immunity and is expressed by most strains causing invasive infections. *J Exp Med* **177**, 1593-1603 (1993).
10. Da Cunha V*, et al. Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nat Commun* **5**, 11 (2014).
11. Almeida A*, et al.* Whole-genome comparison uncovers genomic mutations between group B streptococci sampled from infected newborns and their mothers. *J Bacteriol* **197**, 3354-3366 (2015).
12. Rosini R*, et al.* Genomic analysis reveals the molecular basis for capsule loss in

the group B *Streptococcus* population. *PLOS One* **10**, (2015).

13. Teatero S*, et al.* Clonal Complex 17 group B *Streptococcus* strains causing invasive disease in neonates and adults originate from the same genetic pool. *Sci Rep* **6**, (2016).
14. Bellais S*, et al.* Capsular switching in group B *Streptococcus* CC17 hypervirulent clone: a future challenge for polysaccharide vaccine development. *J Infect Dis* **206**, 1745-1752 (2012).
15. Heath LS*, et al.* The streptococcolytic enzyme zoocin A is a penicillin-binding protein. *FEMS Microbiol Lett* **236**, 205-211 (2004).
16. Lupo A, Ruppen C, Hemphill A, Spellerberg B, Sendi P. Phenotypic and molecular characterization of hyperpigmented group B streptococci. *Zentralbl Bakteriol* **304**, 717-724 (2014).
17. Spellerberg B, Martin S, Brandt C, Lutticken R. The *cyl* genes of *Streptococcus agalactiae* are involved in the production of pigment. *FEMS Microbiol Lett* **188**, 125-128 (2000).
18. Wang CH*, et al.* Effects of oligopeptide permease in group A streptococcal infection. *Infect Immun* **73**, 2881-2890 (2005).
19. Pailhories H, Quentin R, Lartigue M-F. The transcription of the *neuD* gene is stronger in serotype III group B streptococci strains isolated from cerebrospinal fluid than in strains isolated from vagina. *FEMS Microbiol Lett* **349**, 71-75 (2013).
20. Brochet M*, et al.* Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **105**, 15961-15966 (2008).
21. Flores AR*, et al.* Sequence type 1 group B *Streptococcus*, an emerging cause of invasive disease in adults, evolves by small genetic changes. *Proc Natl Acad Sci U S A* **112**, 6431-6436 (2015).
22. Al Safadi R*, et al.* Two-component system RgfA/C activates the *fbsB* gene encoding major fibrinogen-binding protein in highly virulent CC17 clone group B *Streptococcus*. *PLOS One* **6**, (2011).
23. Gravekamp C, Kasper DL, Michel JL, Kling DE, Carey V, Madoff LC. Immunogenicity and protective efficacy of the alpha C protein of group B streptococci are inversely related to the number of repeats. *Infect Immun* **65**, 5216-5221 (1997).

24. Madoff LC, Michel JL, Gong EW, Kling DE, Kasper DL. Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein. *Proc Natl Acad Sci U S A* **93**, 4131-4136 (1996).
25. Lamy MC*, et al.* CovS/CovR of group B *Streptococcus*: a two-component global regulatory system involved in virulence. *Mol Microbiol* **54**, 1250-1268 (2004).
26. Rajagopal L, Vo A, Silvestroni A, Rubens CE. Regulation of cytotoxin expression by converging eukaryotic-type and two-component signalling mechanisms in *Streptococcus agalactiae*. *Mol Microbiol* **62**, 941-957 (2006).
27. Di Palo B*, et al.* Adaptive response of group B *Streptococcus* to high glucose conditions: new insights on the covrs regulation network. *PLOS One* **8**, 11 (2013).
28. Park SE, Jiang SM, Wessels MR. CsrRS and environmental pH regulate group B *Streptococcus* adherence to human epithelial cells and extracellular matrix. *Infect Immun* **80**, 3975-3984 (2012).
29. Inouye M*, et al.* SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6**, (2014).
30. Florindo C*, et al.* Molecular epidemiology of group B streptococcal meningitis in children beyond the neonatal period from Angola. *J Med Microbiol* **60**, 1276-1280 (2011).
31. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).
32. Bankevich A*, et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477 (2012).
33. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**, 623-+ (2015).
34. Chin C-S*, et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-+ (2013).
35. Almeida A*, et al.* Persistence of a dominant bovine lineage of group B *Streptococcus* reveals genomic signatures of host adaptation. *Environ Microbiol*, n/a-n/a (2016).
36. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
37. Page AJ*, et al.* Roary: rapid large-scale prokaryote pan genome analysis.

Bioinformatics **31**, 3691-3693 (2015).

38. Sahl JW, Caporaso JG, Rasko DA, Keim P. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *Peerj* **2**, (2014).
39. Zankari E*, et al.* Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**, 2640-2644 (2012).
40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
41. Brynildsrud O, Snipen L-G, Bohlin J. CNOGpro: detection and quantification of CNVs in prokaryotic whole-genome sequencing data. *Bioinformatics* **31**, 1708-1715 (2015).
42. McKenna A*, et al.* The Genome Analysis ToolKit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
43. DePristo MA*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-+ (2011).
44. Van der Auwera GA*, et al.* From fastQ data to high-confidence variant calls: the Genome Analysis ToolKit best practices pipeline. In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. (2013).
45. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* **15**, (2014).
46. Croucher NJ*, et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, (2015).
47. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
48. Bouckaert R*, et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Comput Biol* **10**, (2014).
49. Bollback JP. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**, (2006).
50. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* **3**, 217-223 (2012).
51. Cingolani P*, et al.* A program for annotating and predicting the effects of single

nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6**, 80-92 (2012).

52. Hedge J, Wilson DJ. Practical approaches for detecting selection in microbial genomes. *PLOS Comput Biol* **12**, (2016).
53. Huerta-Cepas J, *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-D293 (2016).
54. Danecek P, *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

Table 1. Disease-associated homoplasic mutations.

Position[1]	Locus	Product	Mutation	Effect[2]	Total[3]
449849	GBSCOH1_RS02505	Alpha-like surface protein Rib	C > T	Upstream[4]	8
594584	GBSCOH1_RS03240	PI-1 backbone protein	C > T	Upstream[4]	17
1020604	GBSCOH1_RS05180	Fibronectin-binding protein FbsA	A > T	Asp120Glu	13
1121025	GBSCOH1_RS05705	NeuD protein	T > C	Thr67Ala	7
1121393	GBSCOH1_RS05710	NeuC protein	T > G	Thr328Pro	6
1124067	GBSCOH1_RS05720	Capsular polysaccharide transporter CpsL	C > A	Trp270Cys	10
1131143	GBSCOH1_RS05755	Galactosyl transferase CpsE	C > T	Gly65Ser	17
1539816	GBSCOH1_RS07705	CovS two-component sensor histidine kinase	T > G[5]	Glu337Ser	23
1539817	GBSCOH1_RS07705	CovS two-component sensor histidine kinase	C > A[5]		

[1]Nucleotide position of the mutation in relation to the genome of strain COH1.

[2]Resulting amino acid change.

[3]Number of mutations additionally detected in the corresponding gene.

[4]The mutation is located upstream the corresponding locus.

[5]Sequential mutations acquired by the same isolates.

Table 2. Genes most distinctly mutated between carriage and disease isolates.

Locus	Product	Disease[1]	Carriage[1]	*P* value[2]
GBSCOH1_RS00285	Phosphoribosylformylglycinamidine synthase	19	10	0.0004
GBSCOH1_RS01850	Serine/threonine protein kinase Stk1	15	6	0.0091
GBSCOH1_RS04925	23S rRNA methyltransferase	11	1	0.0317
GBSCOH1_RS04975	Permease	16	4	2.40×10^{-5}
GBSCOH1_RS05095	Cell division protein FtsK	26	19	3.74×10^{-5}
GBSCOH1_RS06600	Srr2 cell wall anchor	16	7	0.0044
GBSCOH1_RS07085	Ribonuclease R	13	4	0.0359
GBSCOH1_RS07650	Amidase	13	4	0.0359
GBSCOH1_RS07705	CovS two-component sensor histidine kinase	14	5	0.0183
GBSCOH1_RS08845	DNA polymerase III subunit alpha	15	6	0.0091
GBSCOH1_RS04480	Type II CRISPR RNA-guided endonuclease Cas9	13	19	3.88×10^{-5}
GBSCOH1_RS05130	Carbamoyl phosphate synthase large subunit	10	17	0.0001
GBSCOH1_RS06095	Peptidase C5	9	21	9.55×10^{-12}
GBSCOH1_RS08245	X-prolyl-dipeptidyl aminopeptidase	5	14	0.0002
GBSCOH1_RS08960	Hypothetical protein	7	14	0.0042

[1]Number of mutations exclusively acquired by strains associated with each clinical state

[6]Bonferroni-adjusted *P* values obtained with the outlier test.

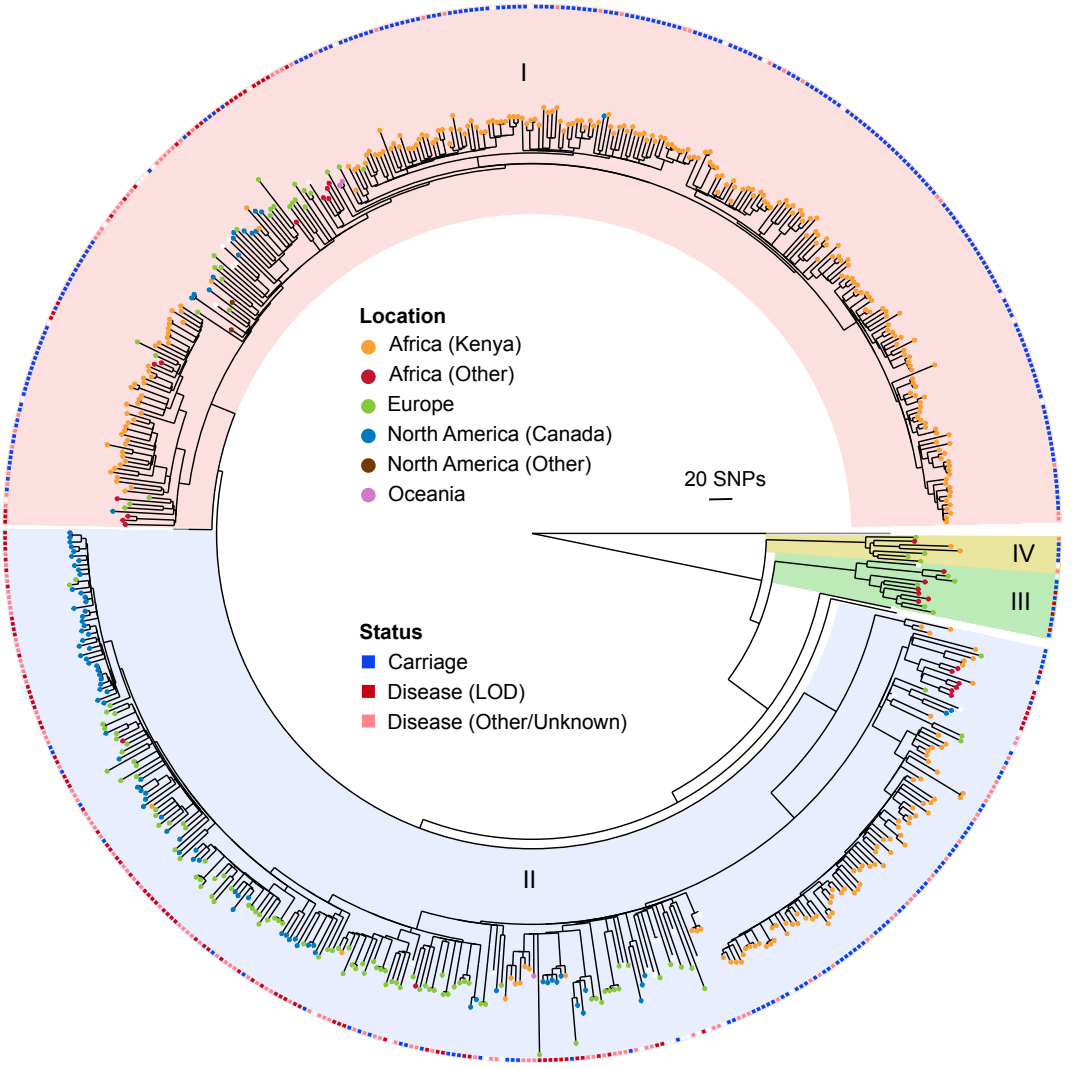


Figure 1. Core-genome phylogeny of the CC17 population.
Phylogenetic tree of 612 CC17 genomes, built with RAxML[47] based on the core and recombination-free alignment of 12 431 SNPs along a 1.48 Mb sequence. Isolates are colour-coded according to the geographical origin and clinical state, as indicated in the figure key. The four main distinct clades in the phylogeny are differentially coloured, and labelled with roman numerals.

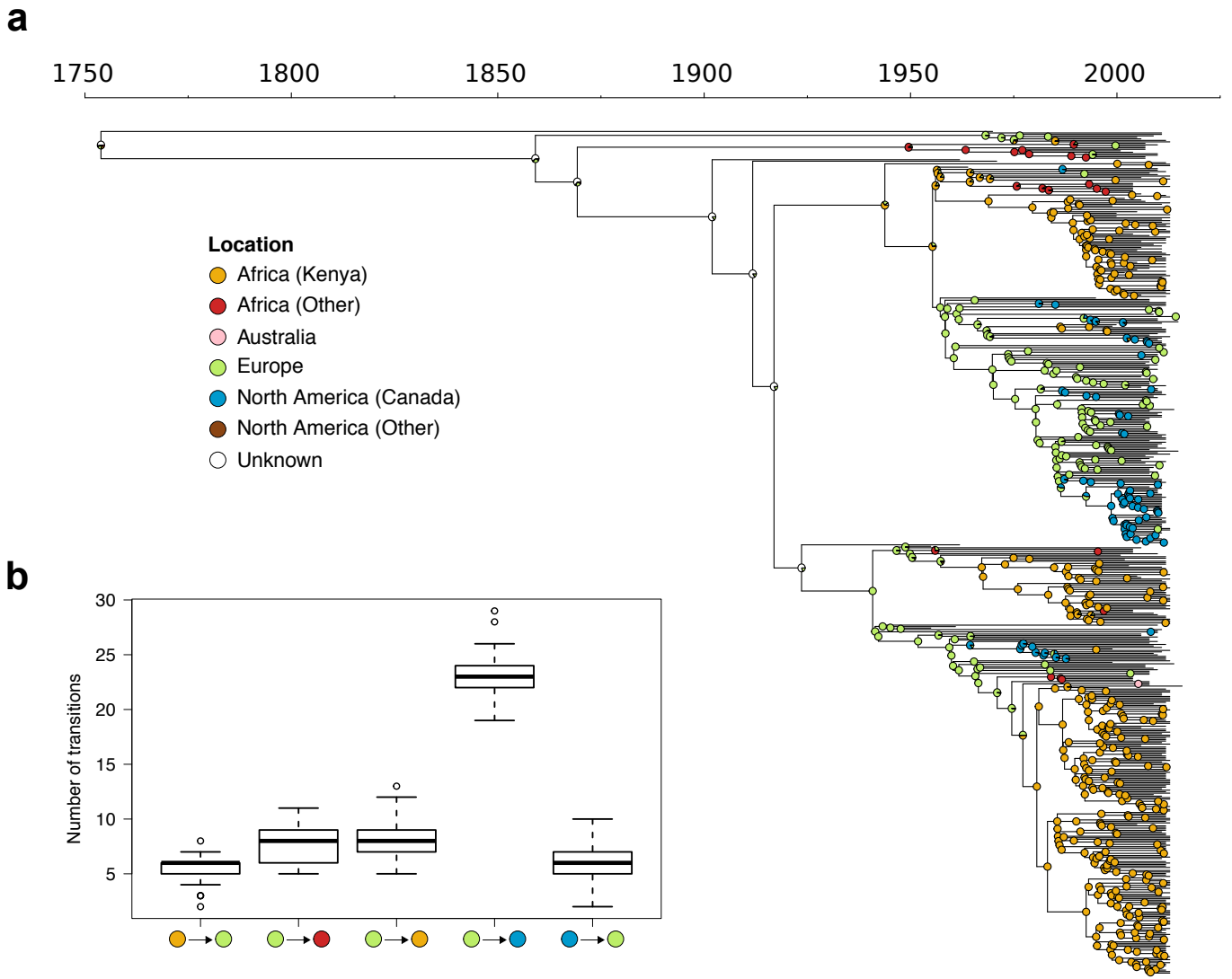


Figure 2. Intercontinental transmission events and ancestral reconstruction of each geographical origin. **a**, Plot of the most likely ancestral state (location) predicted for each node of the core-genome phylogeny of CC17 using the make.simmap tool[49] from the phytools R package[50]. **b**, The five most frequent intercontinental transmission events inferred with the count.simmap function[49] by counting the number of transitions between the locations predicted for each node in the phylogeny and the ones at the tips of the tree. Coloured-circles depict different locations according to the figure key.

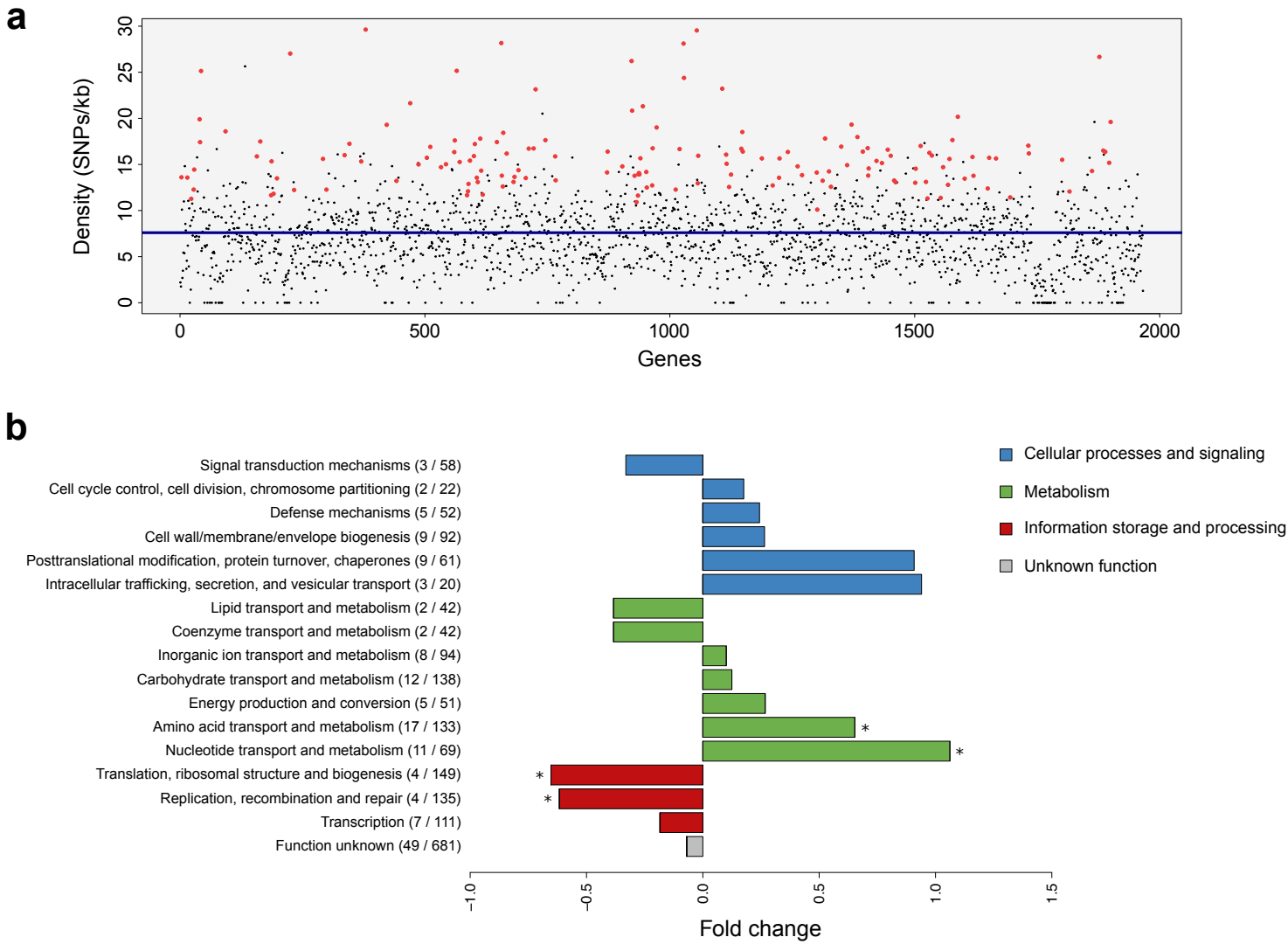


Figure 3. Parallel evolution of coding sequences above neutral expectation.
a, Mutation frequency observed per gene in the CC17 population (black) in relation to the expected substitution rate across the COH1 reference genome (blue line). Red dots correspond to the 152 genes with a statistically significant ($P < 0.05$, exact Poisson test) mutational bias compared to a neutral model of evolution (Supplementary Table 2). **b**, Functional classification of the 152 genes, based on the eggNOG database. Fold change corresponds to the proportional difference of the number of genes among those with a mutational bias, compared to the COH1 coding sequences (indicated in parenthesis). Statistical significance was assessed with a Fisher's exact test. *$P < 0.05$

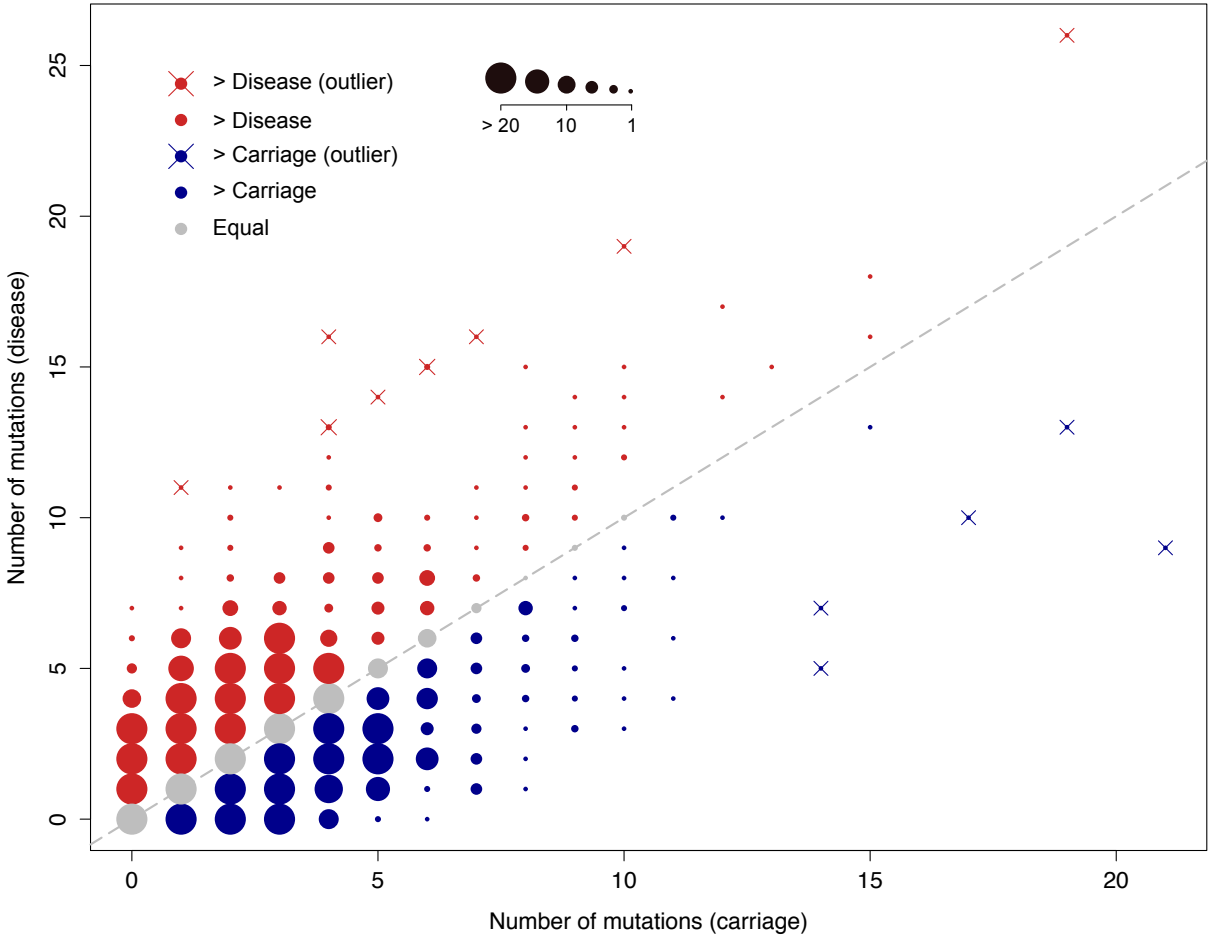


Figure 4. Mutation frequencies per gene between carriage- and disease-associated isolates. Plot of a linear model assessing the correlation between the mutations acquired per gene for each clinical status. Red, grey and blue coloured dots depict genes that were more or equally mutated in strains associated with disease or carriage according to the figure key. The size of each data point is proportional to the number of genes found, as indicated in the figure key. Grey dashed line depicts the separation between carriage- and disease-associated genes. Outlier genes were detected with a Bonferroni-adjusted outlier test and only those with $P < 0.05$ are represented.

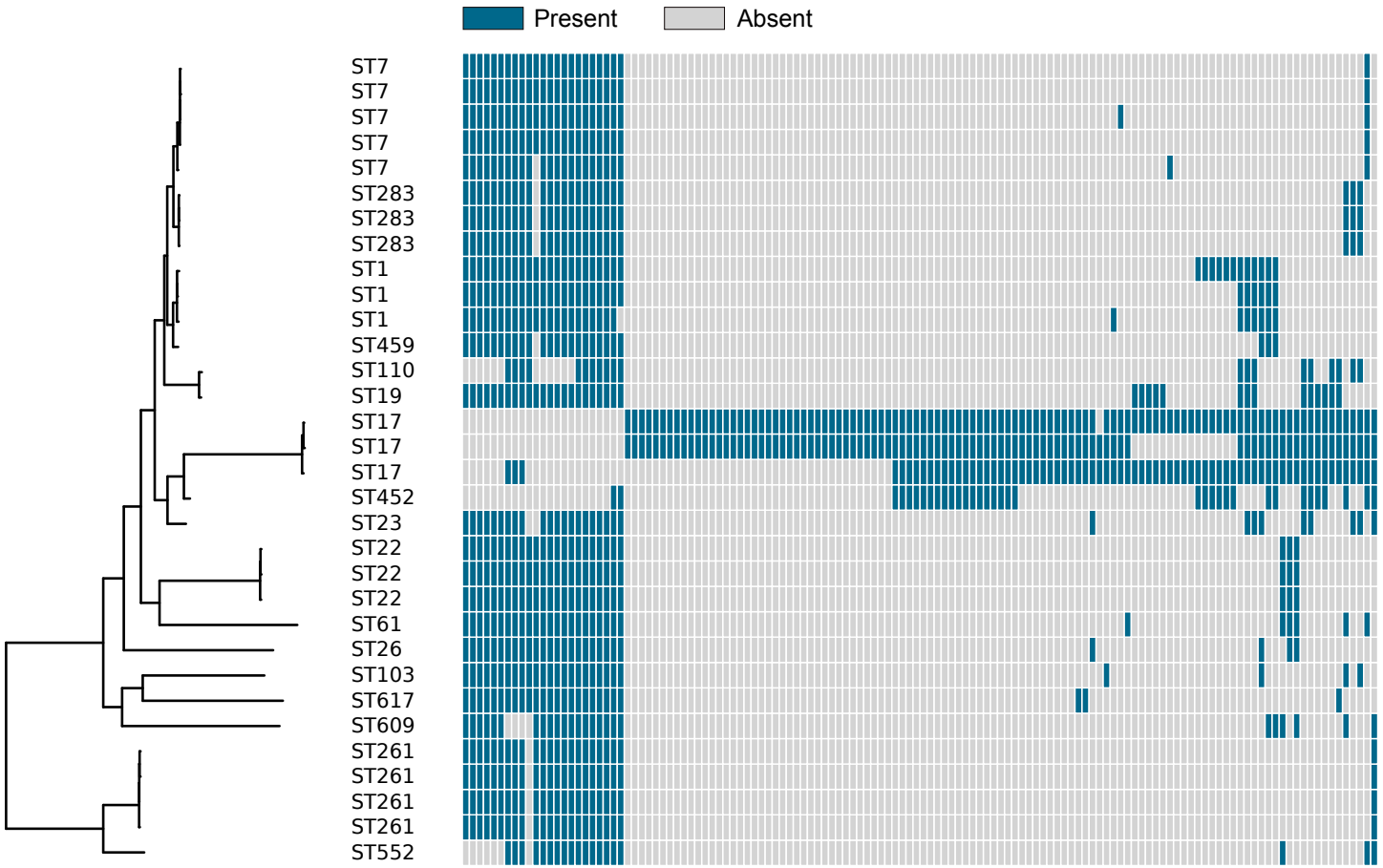


Figure 5. Phylogenetic distribution of genes associated with CC17.
Plot of the 130 genes significantly associated with the CC17 lineage, against the phylogeny of 32 complete GBS genomes. The association between gene content and CC17 was inferred with Scoary (https://github.com/AdmiralenOla/Scoary), and the phylogeny was built from the core-genome alignment of the assembled genomes using Parsnp[45].

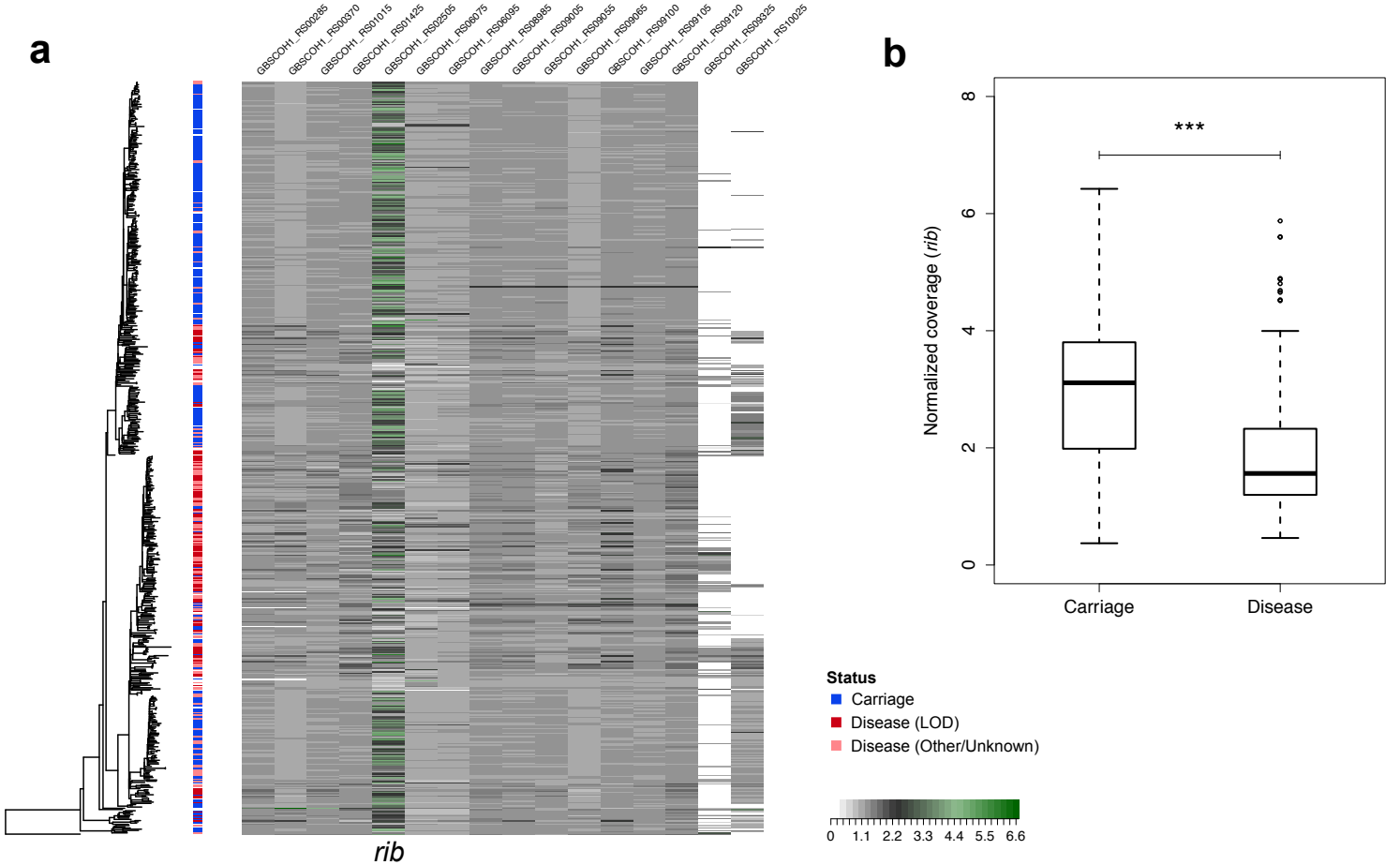


Figure 6. Coverage variation of the most replicated genes.
a, Heatmap depicting the coverage variation of 16 genes across the CC17 phylogeny. Values correspond to the sequencing coverage normalized with that obtained with the self-mapped reference genome of COH1. **b**, Normalized coverage of the gene coding for the alpha-like surface protein Rib in carriage- and disease-specific strains. Significance values were calculated from a two-tailed *t* test. ***$P < 0.001$.

Supplementary Table 1. GBS strains analysed in this work belonging to CC17

Sample	Continent	Country	Year	Host	Status[1]	Disease	Origin[2]	ST[3]	Source
Child1	Europe	France	2012	Human	Disease	EOD	Blood	ST17	Almeida, et al. (2015)
Child11-1	Europe	France	2009	Human	Disease	LOD	Blood	ST17	Almeida, et al. (2015)
Child11-2	Europe	France	2009	Human	Disease	LOD	Blood	ST17	Almeida, et al. (2015)
Child13	Europe	France	2010	Human	Disease	EOD	CSF	ST17	Almeida, et al. (2015)
Child14	Europe	France	2010	Human	Disease	LOD	CSF	ST17	Almeida, et al. (2015)
Child16	Europe	France	2011	Human	Carriage	Unknown	Gastric fluid	ST17	Almeida, et al. (2015)
Child17	Europe	France	2011	Human	Disease	EOD	Blood	ST17	Almeida, et al. (2015)
Child19-1	Europe	France	2011	Human	Disease	EOD	CSF	ST17	Almeida, et al. (2015)
Child19-2	Europe	France	2011	Human	Disease	EOD	Blood	ST17	Almeida, et al. (2015)
Child19-3	Europe	France	2011	Human	Disease	EOD	Urine	ST17	Almeida, et al. (2015)
Child4	Europe	France	2008	Human	Disease	EOD	Blood	ST17	Almeida, et al. (2015)
Child5	Europe	France	2008	Human	Disease	LOD	Blood	ST17	Almeida, et al. (2015)
Child7	Europe	France	2008	Human	Disease	LOD	Blood	ST17	Almeida, et al. (2015)
Child8-1	Europe	France	2008	Human	Disease	EOD	CSF	ST17	Almeida, et al. (2015)
Child8-2	Europe	France	2008	Human	Disease	EOD	Blood	ST17	Almeida, et al. (2015)
Mother1	Europe	France	2012	Human	Carriage	Unknown	Vaginal fluid	ST17	Almeida, et al. (2015)
Mother13	Europe	France	2010	Human	Disease	Other	Milk	ST17	Almeida, et al. (2015)
Mother14	Europe	France	2010	Human	Disease	Other	Milk	ST17	Almeida, et al. (2015)
Mother16	Europe	France	2011	Human	Carriage	Unknown	Placenta	ST17	Almeida, et al. (2015)
Mother17	Europe	France	2011	Human	Carriage	Unknown	Placenta	ST17	Almeida, et al. (2015)
Mother19	Europe	France	2011	Human	Carriage	Unknown	Vaginal fluid	ST17	Almeida, et al. (2015)
Mother4	Europe	France	2008	Human	Carriage	Unknown	Vaginal fluid	ST17	Almeida, et al. (2015)
Mother5-1	Europe	France	2008	Human	Carriage	Unknown	Vaginal fluid	ST17	Almeida, et al. (2015)
Mother5-2	Europe	France	2008	Human	Disease	Other	Milk	ST17	Almeida, et al. (2015)
Mother8	Europe	France	2008	Human	Carriage	Unknown	Vaginal fluid	ST17	Almeida, et al. (2015)
B55	Unknown	Unknown	1962	Unknown	Unknown	Unknown	Unknown	ST17	CIP
B56	Unknown	Unknown	1962	Unknown	Unknown	Unknown	Unknown	ST482	CIP
B74	Unknown	Unknown	1971	Unknown	Unknown	Unknown	Unknown	ST125	CIP
B83	Unknown	Unknown	1970	Unknown	Unknown	Unknown	Unknown	ST17	CIP
B85	Unknown	Unknown	1964	Unknown	Unknown	Unknown	Unknown	ST17	CIP
6630_5_10	Europe	Denmark	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_5_21	Europe	Denmark	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_5_24	Europe	Bulgaria	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_5_3	Europe	Belgium	2008	Human	Disease	EOD	Unknown	ST315	Da Cunha, et al. (2014)
6630_5_5	Europe	Denmark	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_5_8	Europe	Denmark	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_6_19	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_6_21	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_6_22	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_6_24	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_10	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_12	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_13	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST467	Da Cunha, et al. (2014)
6630_7_14	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_17	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_18	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_19	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_2	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_3	Europe	Germany	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_7	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST496	Da Cunha, et al. (2014)
6630_7_8	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_7_9	Europe	Italy	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_8_14	Europe	Great Britain	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_8_15	Europe	Great Britain	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6630_8_16	Europe	Great Britain	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6753_5_2	Australia	Australia	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6753_5_23	Australia	Australia	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
6753_5_28	Australia	Aus_Indig	2008	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
Bangui-IP-112	Africa	Central African Republic	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Bangui-IP-29	Africa	Central African Republic	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Bangui-IP-61	Africa	Central African Republic	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
BM110	North America	USA	<1989	Human	Disease	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH206800353	Europe	France	2006	Human	Disease	LOD	CSF	ST17	Da Cunha, et al. (2014)
CCH206800391	Europe	France	2006	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH207800343	Europe	France	2007	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH207800974	Europe	France	2007	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH208800031	Europe	France	2008	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH208800147	Europe	France	2008	Human	Disease	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH208800438	Europe	France	2008	Human	Disease	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH208800481	Europe	France	2008	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH208800879	Europe	France	2008	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH209800039	Europe	France	2009	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH209800160	Europe	France	2009	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH209800361	Europe	France	2009	Human	Disease	Unknown	Unknown	ST291	Da Cunha, et al. (2014)
CCH209800522	Europe	France	2009	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH209801071	Europe	France	2009	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH210107206	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)

CCH210125551	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210126535	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210145919	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210150530	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210152823	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210155112	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210160764	Europe	France	2010	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH210169151	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210169294	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210172343	Europe	France	2010	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CCH210800096	Europe	France	2010	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH210800140	Europe	France	2010	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH210800593	Europe	France	2008	Human	Disease	LOD	CSF	ST17	Da Cunha, et al. (2014)
CCH210800688	Europe	France	2010	Human	Disease	LOD	Unknown	ST17	Da Cunha, et al. (2014)
CCH211800247	Europe	France	2011	Human	Disease	Unknown	Knee prothesis	ST291	Da Cunha, et al. (2014)
CCH211800398	Europe	France	2011	Human	Carriage	Unknown	Gastric fluid	ST291	Da Cunha, et al. (2014)
cliniq918	Europe	France	2009	Human	Disease	LOD	CSF	ST17	Da Cunha, et al. (2014)
CRBIP22-120	North America	USA	1961	Human	Disease	Unknown	CSF	ST17	Da Cunha, et al. (2014)
CRBIP22-94	Europe	France	1955	Human	Disease	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CRBIP22-98	Europe	France	1955	Human	Disease	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
CZ183	Europe	Czech republic	Unknown	Bovine	Unknown	Unknown	Unknown	ST355	Da Cunha, et al. (2014)
Dakar-IP-107	Africa	Central African Republic	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Dakar-IP-8	Africa	Senegal	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Dakar-IP-98	Africa	Senegal	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Madagascar-IP-20	Africa	Madagascar	2006	Human	Carriage	Unknown	Unknown	ST291	Da Cunha, et al. (2014)
Madagascar-IP-33	Africa	Madagascar	2006	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Madagascar-IP-71	Africa	Madagascar	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
Madagascar-IP-84	Africa	Madagascar	2007	Human	Carriage	Unknown	Unknown	ST17	Da Cunha, et al. (2014)
NEM1857	Europe	France	2001	Human	Disease	LOD	CSF	ST252	Da Cunha, et al. (2014)
NEM318	Europe	France	1995	Human	Disease	LOD	CSF	ST17	Da Cunha, et al. (2014)
Spain-82	Europe	Spain	2005	Human	Unknown	Unknown	Unknown	ST291	Da Cunha, et al. (2014)
WC3	Europe	UK	2001	Human	Disease	EOD	Unknown	ST17	Da Cunha, et al. (2014)
BSU96	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
CCUG17336	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
CCUG37736	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
CCUG44186	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
CCUG49086	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
CCUG49087	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
COH1	Unknown	Unknown	Unknown	Human	Disease	Unknown	Unknown	ST17	NCBI
FSLS3-102	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST31	NCBI
GB00097	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00111	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST32	NCBI
GB00112	North America	Canada	1999	Human	Carriage	Unknown	Unknown	ST17	NCBI
GB00115	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00557	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00654	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00891	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00940	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
GB00963	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
LMG15085	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
LMG15094	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
LMG15095	Unknown	Unknown	Unknown	Human	Unknown	Unknown	Unknown	ST17	NCBI
299693	Europe	Unknown	Unknown	Human	Disease	Adult	Unknown	ST17	Rosini, et al. (2015)
379415	Europe	Unknown	Unknown	Human	Disease	Adult	Unknown	ST17	Rosini, et al. (2015)
BE-PW-080	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST484	Rosini, et al. (2015)
BG-PW-074	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
DE-NI-024	Europe	Unknown	Unknown	Human	Disease	LOD	Unknown	ST17	Rosini, et al. (2015)
DE-NI-035	Europe	Unknown	Unknown	Human	Disease	LOD	Unknown	ST17	Rosini, et al. (2015)
ES-NI-001	Europe	Unknown	Unknown	Human	Disease	EOD	Unknown	ST17	Rosini, et al. (2015)
ES-NI-003	Europe	Unknown	Unknown	Human	Disease	LOD	Unknown	ST17	Rosini, et al. (2015)
ES-NI-010	Europe	Unknown	Unknown	Human	Disease	LOD	Unknown	ST17	Rosini, et al. (2015)
ES-PW-034	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
ES-PW-088	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
ES-PW-181	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
IT-NI-001	Europe	Unknown	Unknown	Human	Disease	LOD	Unknown	ST17	Rosini, et al. (2015)
IT-PW-0063	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
IT-PW-0094	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
ML30419	Europe	Unknown	Unknown	Human	Carriage	Unknown	Unknown	ST17	Rosini, et al. (2015)
SH0248	Europe	Unknown	Unknown	Human	Disease	Adult	Unknown	ST17	Rosini, et al. (2015)
SH3115	Europe	Unknown	Unknown	Human	Disease	Adult	Unknown	ST17	Rosini, et al. (2015)
K28931	Africa	Kenya	2008	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K31176	Africa	Kenya	2008	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K33534	Africa	Kenya	2008	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K3605	Africa	Kenya	1998	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K33921	Africa	Kenya	2008	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K4340	Africa	Kenya	1999	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K36829	Africa	Kenya	2009	Human	Disease	Unknown	Blood	ST484	Seale, et al. (2016)
K38783	Africa	Kenya	2009	Human	Disease	Unknown	CSF	ST17	Seale, et al. (2016)
K39258	Africa	Kenya	2009	Human	Disease	Unknown	CSF	ST484	Seale, et al. (2016)

K4958	Africa	Kenya	1999	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K5136	Africa	Kenya	1999	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K6229	Africa	Kenya	2000	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K6234	Africa	Kenya	2000	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K6245	Africa	Kenya	2000	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K6625	Africa	Kenya	2000	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K6666	Africa	Kenya	2000	Human	Disease	Unknown	Blood	ST484	Seale, et al. (2016)
K45678	Africa	Kenya	2010	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K47340	Africa	Kenya	2011	Human	Disease	Unknown	Blood	ST484	Seale, et al. (2016)
K7229	Africa	Kenya	2001	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K48051	Africa	Kenya	2011	Human	Disease	Unknown	CSF	ST17	Seale, et al. (2016)
K48612	Africa	Kenya	2011	Human	Disease	Unknown	CSF	ST484	Seale, et al. (2016)
K50468	Africa	Kenya	N/A	Human	Disease	Unknown	Blood	ST792	Seale, et al. (2016)
K8309	Africa	Kenya	2002	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K8555	Africa	Kenya	2002	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K9182	Africa	Kenya	2002	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K55813	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K55826	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K56640	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K10268	Africa	Kenya	2003	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K57085	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K57810	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K59563	Africa	Kenya	2012	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K11385	Africa	Kenya	2003	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K13758	Africa	Kenya	2004	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K14206	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K16534	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K16755	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K17110	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K17308	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K17389	Africa	Kenya	2005	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K17897	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST147	Seale, et al. (2016)
K17900	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K18723	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST484	Seale, et al. (2016)
K19212	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K19244	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K19924	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST484	Seale, et al. (2016)
K20019	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K20332	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K20349	Africa	Kenya	2006	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K23965	Africa	Kenya	2007	Human	Disease	Unknown	Blood	ST17	Seale, et al. (2016)
K49738	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K49740	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K49831	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K49933	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K49965	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K49970	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K49981	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50057	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K50140	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50142	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K50146	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K50223	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50414	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50441	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50466	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST788	Seale, et al. (2016)
K50709	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K50724	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50791	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50792	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K50877	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST784	Seale, et al. (2016)
K50972	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51091	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51201	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51260	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51348	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K51572	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K51578	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51590	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K51608	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51617	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51738	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51742	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51745	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K51818	Africa	Kenya	2011	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52049	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52065	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K52315	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52346	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)

K52367	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52486	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52623	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52646	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52647	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52773	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K52862	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55286	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST291	Seale, et al. (2016)
K55292	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55320	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55416	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55493	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55547	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55593	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55594	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K55603	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55631	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K55655	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K55881	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56031	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56061	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56062	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56064	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56071	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56183	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56185	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56195	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K56218	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56228	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56231	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56235	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST788	Seale, et al. (2016)
K56245	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56246	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56369	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56371	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56372	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST792	Seale, et al. (2016)
K56508	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56615	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56645	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56664	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56665	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56721	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56794	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56796	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56809	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56814	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56857	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56866	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56889	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56949	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56981	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST790	Seale, et al. (2016)
K56988	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56989	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56991	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K56995	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K57049	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57089	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57335	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57349	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57350	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57351	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57402	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K57477	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57513	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57538	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K57543	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57749	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K57751	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K57806	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57809	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST791	Seale, et al. (2016)
K57843	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57910	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57965	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58104	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58136	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58211	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K58243	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58435	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58491	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)

K58493	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58551	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K58574	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K58634	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58694	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58711	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K58758	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K58760	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58795	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K58805	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59095	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59097	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59113	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59126	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59183	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59200	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59245	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59410	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59423	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59425	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59426	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59439	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59498	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59522	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59564	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59566	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K59590	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59605	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59645	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59648	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59673	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59710	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59711	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59712	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST795	Seale, et al. (2016)
K59729	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59731	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59733	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59747	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59749	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59753	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59772	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59778	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59791	Africa	Kenya	2012	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60291	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60432	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60564	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60565	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K60566	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60574	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60663	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60670	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60715	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K60719	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60720	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60756	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60788	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60790	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60792	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60793	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K60797	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60812	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60830	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60831	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60843	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST800	Seale, et al. (2016)
K60885	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60887	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60939	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K60957	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61006	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61016	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61018	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61053	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61063	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61068	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61078	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61079	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61164	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61177	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61251	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)

K61252	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61276	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61280	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61282	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61305	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61331	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61335	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61336	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61362	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61407	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61415	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61431	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST790	Seale, et al. (2016)
K61435	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61461	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61514	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61516	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61521	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61572	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61586	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61590	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61594	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61608	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61611	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61620	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61639	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61640	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61688	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST109	Seale, et al. (2016)
K61704	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61731	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61763	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61767	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K61832	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61840	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61841	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61844	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61845	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K61994	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62025	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62032	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62042	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62043	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62079	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62080	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62110	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62142	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62143	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62145	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST801	Seale, et al. (2016)
K62153	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62154	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62156	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62168	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62250	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62313	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62348	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62349	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62393	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62418	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62459	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62464	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62474	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62475	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K62476	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62554	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62558	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST291	Seale, et al. (2016)
K62569	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62582	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K62583	Africa	Kenya	2013	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K31131	Africa	Kenya	N/A	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K57911	Africa	Kenya	N/A	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59069	Africa	Kenya	N/A	Human	Carriage	Unknown	Recto-vaginal	ST484	Seale, et al. (2016)
K59185	Africa	Kenya	N/A	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K59565	Africa	Kenya	N/A	Human	Carriage	Unknown	Recto-vaginal	ST17	Seale, et al. (2016)
K56485	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST786	Seale, et al. (2016)
K56941	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K57063	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K57353	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K57354	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K58245	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K58759	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST484	Seale, et al. (2016)

K58784	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K59216	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K59497	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K59732	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K59750	Africa	Kenya	2012	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST484	Seale, et al. (2016)
K60691	Africa	Kenya	2013	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST484	Seale, et al. (2016)
K61004	Africa	Kenya	2013	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST17	Seale, et al. (2016)
K61520	Africa	Kenya	2013	Human	Unknown	Unknown	Ear/Nose/Umbilicus	ST484	Seale, et al. (2016)
608	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST450	Florindo, et al. (2011)
681	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
1810	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST450	Florindo, et al. (2011)
2211	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST17	Florindo, et al. (2011)
2410	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST450	Florindo, et al. (2011)
2536	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
2596	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST287	Florindo, et al. (2011)
3501	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST450	Florindo, et al. (2011)
4512	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
5234	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST17	Florindo, et al. (2011)
5590	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
5659	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
5717	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST109	Florindo, et al. (2011)
5995	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST17	Florindo, et al. (2011)
6129	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST287	Florindo, et al. (2011)
6521	Africa	Angola	2004/2005	Human	Disease	LOD	Unknown	ST451	Florindo, et al. (2011)
CH_7	Europe	France	2007	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_8	Europe	France	2011	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_9	Europe	France	2010	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_10	Europe	France	2009	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_11	Europe	France	2008	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_12	Europe	France	2007	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_13	Europe	France	2012	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_14	Europe	France	2014	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_15	Europe	France	2014	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_16	Europe	France	2015	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_17	Europe	France	2010	Human	Disease	EOD	Blood	ST17	CNR-Strep
CH_18	Europe	France	2010	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_19	Europe	France	2016	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_20	Europe	France	2015	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_21	Europe	France	2009	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_22	Europe	France	2012	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_23	Europe	France	2015	Human	Disease	EOD	Blood	ST17	CNR-Strep
CH_24	Europe	France	2015	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_25	Europe	France	2013	Human	Disease	EOD	Blood	ST17	CNR-Strep
CH_26	Europe	France	2013	Human	Disease	LOD	Blood	ST17	CNR-Strep
CH_27	Europe	France	2013	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_3	Europe	France	2006	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_4	Europe	France	2006	Human	Disease	LOD	CSF	ST17	CNR-Strep
CH_5	Europe	France	2006	Human	Disease	LOD	CSF	ST17	CNR-Strep
NGBS003	North America	Canada	2009	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS026	North America	Canada	2009	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS031	North America	Canada	2009	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS034	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS050	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS069	North America	Canada	2010	Human	Disease	LOD	CSF	ST17	Teatero, et al. (2016)
NGBS079	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS080	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS082	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS089	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS106	North America	Canada	2010	Human	Disease	LOD	CSF	ST17	Teatero, et al. (2016)
NGBS108	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS126	North America	Canada	2010	Human	Disease	Adult	Other	ST17	Teatero, et al. (2016)
NGBS128	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS129	North America	Canada	2010	Human	Disease	Adult	Blood	ST31	Teatero, et al. (2016)
NGBS141	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS143	North America	Canada	2010	Human	Disease	Adult	Other	ST17	Teatero, et al. (2016)
NGBS147	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS149	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS152	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS160	North America	Canada	2010	Human	Disease	Adult	Other	ST31	Teatero, et al. (2016)
NGBS169	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS186	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS205	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS220	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS222	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS238	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS239	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS250	North America	Canada	2011	Human	Disease	LOD	CSF	ST17	Teatero, et al. (2016)
NGBS271	North America	Canada	2011	Human	Disease	LOD	CSF	ST17	Teatero, et al. (2016)
NGBS277	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)

NGBS282	North America	Canada	2010	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS291	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS296	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS297	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS299	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS306	North America	Canada	2011	Human	Disease	Adult	Blood	ST482	Teatero, et al. (2016)
NGBS312	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS317	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS318	North America	Canada	2011	Human	Disease	Adult	Blood	ST290	Teatero, et al. (2016)
NGBS327	North America	Canada	2011	Human	Disease	Adult	Blood	ST484	Teatero, et al. (2016)
NGBS343	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS345	North America	Canada	2011	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS356	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS361	North America	Canada	2011	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS362	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS368	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS370	North America	Canada	2011	Human	Disease	LOD	Blood	ST482	Teatero, et al. (2016)
NGBS374	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS377	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS386	North America	Canada	2011	Human	Disease	Adult	Synovial fluid	ST17	Teatero, et al. (2016)
NGBS398	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS403	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS417	North America	Canada	2011	Human	Disease	Adult	Synovial fluid	ST17	Teatero, et al. (2016)
NGBS421	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS422	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS431	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS440	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS456	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS464	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS469	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS470	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS483	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS485	North America	Canada	2011	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS486	North America	Canada	2011	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS500	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS501	North America	Canada	2012	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS502	North America	Canada	2012	Human	Disease	Adult	Blood	ST95	Teatero, et al. (2016)
NGBS515	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS516	North America	Canada	2012	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS531	North America	Canada	2012	Human	Disease	Adult	Blood	ST148	Teatero, et al. (2016)
NGBS534	North America	Canada	2012	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS551	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS576	North America	Canada	2012	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS583	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS593	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS594	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS596	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS607	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS608	North America	Canada	2012	Human	Disease	Adult	Unknown	ST17	Teatero, et al. (2016)
NGBS609	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS613	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS614	North America	Canada	2012	Human	Disease	LOD	CSF	ST148	Teatero, et al. (2016)
NGBS618	North America	Canada	2012	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS622	North America	Canada	2012	Human	Disease	LOD	Blood	ST148	Teatero, et al. (2016)
NGBS625	North America	Canada	2012	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS627	North America	Canada	2012	Human	Disease	Adult	Blood	ST17	Teatero, et al. (2016)
NGBS632	North America	Canada	2012	Human	Disease	Adult	Unknown	ST17	Teatero, et al. (2016)
NGBS636	North America	Canada	2010	Human	Disease	LOD	Blood	ST17	Teatero, et al. (2016)
NGBS641	North America	Canada	2010	Human	Disease	EOD	Blood	ST17	Teatero, et al. (2016)
NGBS644	North America	Canada	2011	Human	Disease	Adult	Synovial fluid	ST17	Teatero, et al. (2016)
NGBS650	North America	Canada	2011	Human	Disease	LOD	CSF	ST17	Teatero, et al. (2016)

[1]Clinical status (carriage or disease).

[2]Biological tissue from which the sample was taken. CSF = cerebrospinal fluid

[3]Sequence type inferred from the *in silico* MLST.

Supplementary Table 2. Genes with a statistically significant mutational bias.

Locus	Product	Length (bp)	NS[1]	S[2]	dN/dS[3]	Density[4]	Isolates[5]	*P* value[6]
GBSCOH1_RS00015	diacylglycerol kinase	882	10	2	1.34	13.61	165	0.0410
GBSCOH1_RS00245	CHAP domain-containing protein	1326	14	4	1.10	13.57	28	0.0153
GBSCOH1_RS00285	phosphoribosylformylglycinamidine synthase	3726	26	16	0.48	11.27	378	0.0095
GBSCOH1_RS00310	inosine monophosphate cyclohydrolase	1548	11	8	0.41	12.27	38	0.0317
GBSCOH1_RS00315	hypothetical protein	900	13	0	13.00	14.44	18	0.0230
GBSCOH1_RS00370	RpiR family transcriptional regulator	804	13	3	1.14	19.90	31	0.0006
GBSCOH1_RS00375	phosphoribosylamine--glycine ligase	1263	14	8	0.53	17.42	30	0.0004*
GBSCOH1_RS00385	5-(carboxyamino)imidazole ribonucleotide synthase	1074	13	14	0.28	25.14	328	1.52×10^{-7}*
GBSCOH1_RS00715	D-alanyl-D-alanine carboxypeptidase DacA	753	12	2	1.57	18.59	13	0.0024
GBSCOH1_RS01065	hypothetical protein	504	8	0	8.00	15.87	15	0.0416
GBSCOH1_RS01210	hypothetical protein	1629	14	5	0.76	11.66	339	0.0480
GBSCOH1_RS01215	ABC transporter permease	978	12	3	1.26	15.34	342	0.0095
GBSCOH1_RS01235	PTS beta-glucoside transporter subunit EIIBCA	2031	20	4	1.54	11.82	104	0.0259
GBSCOH1_RS01270	oxidoreductase	1260	12	5	0.66	13.49	23	0.0189
GBSCOH1_RS01520	hypothetical protein	1635	12	8	0.42	12.23	34	0.0290
GBSCOH1_RS01815	hypothetical protein	705	9	2	1.26	15.60	330	0.0214
GBSCOH1_RS01850	Stk1 serine/threonine protein kinase	1956	17	7	0.73	12.27	37	0.0177
GBSCOH1_RS02060	membrane protein	1062	15	2	2.12	16.01	341	0.0040
GBSCOH1_RS02110	GNAT family acetyltransferase	522	6	3	0.51	17.24	220	0.0204
GBSCOH1_RS02235	YSIRK signal domain/LPXTG anchor domain surface protein	1566	17	7	0.65	15.33	46	0.0013
GBSCOH1_RS02280	hypothetical protein	135	3	1	0.62	29.63	5	0.0206
GBSCOH1_RS02555	GNAT family acetyltransferase	570	8	3	0.72	19.30	156	0.0051
GBSCOH1_RS02655	multidrug ABC transporter ATP-binding protein	1740	13	10	0.36	13.22	65	0.0092
GBSCOH1_RS02795	hypothetical protein	231	3	2	0.40	21.65	88	0.0333
GBSCOH1_RS02880	dihydroorotate oxidase	933	9	5	0.49	15.01	320	0.0141
GBSCOH1_RS02965	hypothetical protein	636	8	2	1.10	15.72	17	0.0262
GBSCOH1_RS03000	hypothetical protein	414	4	3	0.37	16.91	7	0.0414
GBSCOH1_RS03115	transposase	612	8	1	1.88	14.71	16	0.0476
GBSCOH1_RS03165	endonuclease	1065	12	4	0.79	15.02	26	0.0091
GBSCOH1_RS03250	class C sortase	918	12	3	1.16	16.34	22	0.0055
GBSCOH1_RS03255	class C sortase	852	8	7	0.34	17.61	41	0.0028
GBSCOH1_RS03275	amidase	159	4	0	4.00	25.16	330	0.0345
GBSCOH1_RS03305	multidrug ABC transporter permease	786	8	4	0.54	15.27	53	0.0195
GBSCOH1_RS03380	peptidase S8	2658	23	8	0.80	11.66	533	0.0152
GBSCOH1_RS03390	hypothetical protein	2895	26	9	0.77	12.09	172	0.0064
GBSCOH1_RS03395	peptidase M13	1086	12	2	1.54	12.89	316	0.0422
GBSCOH1_RS03410	hypothetical protein	519	4	4	0.27	15.41	8	0.0478
GBSCOH1_RS03450	hypothetical protein	1446	19	4	1.31	15.91	27	0.0010
GBSCOH1_RS03460	cytochrome o ubiquinol oxidase	639	9	2	1.33	17.21	32	0.0113
GBSCOH1_RS03480	LysR family transcriptional regulator	885	8	4	0.53	13.56	37	0.0419
GBSCOH1_RS03490	lactate dehydrogenase	993	11	2	1.58	13.09	40	0.0443
GBSCOH1_RS03515	2-dehydro-3-deoxyphosphooctonate aldolase	618	10	1	2.96	17.80	30	0.0090
GBSCOH1_RS03525	mannonate dehydratase	1047	10	5	0.55	14.33	20	0.0165
GBSCOH1_RS03540	beta-hexosamidase	1791	18	3	1.65	11.73	28	0.0376
GBSCOH1_RS03685	SprT family protein	459	8	0	8.00	17.43	8	0.0263
GBSCOH1_RS03730	hypothetical protein	213	3	3	0.25	28.17	14	0.0064
GBSCOH1_RS03735	membrane protein	798	8	3	0.76	13.78	17	0.0453
GBSCOH1_RS03745	5-amino-6-(5-phosphoribosylamino)uracil reductase	1110	12	2	1.66	12.61	18	0.0489
GBSCOH1_RS03750	riboflavin synthase subunit alpha	651	10	2	1.43	18.43	10	0.0050
GBSCOH1_RS03785	epimerase	618	8	2	1.10	16.18	19	0.0221
GBSCOH1_RS03855	MFS transporter	1221	13	3	1.32	13.10	41	0.0280
GBSCOH1_RS03865	cell surface protein	1538	16	5	0.91	13.65	91	0.0089
GBSCOH1_RS03905	methyltransferase	765	8	3	0.72	14.38	27	0.0354
GBSCOH1_RS03980	gluconate:proton symporter	1257	14	3	1.56	13.52	18	0.0186
GBSCOH1_RS04010	glutathione-dependent reductase	957	13	3	1.14	16.72	86	0.0035
GBSCOH1_RS04060	membrane protein	717	9	3	0.89	16.74	18	0.0103
GBSCOH1_RS04080	transcriptional regulator	216	4	1	1.18	23.15	12	0.0260
GBSCOH1_RS04180	hypothetical protein	567	9	1	2.77	17.64	11	0.0131
GBSCOH1_RS04280	glucose-1-phosphate adenylyltransferase	1134	14	4	0.93	15.87	127	0.0034
GBSCOH1_RS04285	starch synthase	1431	15	4	0.93	13.28	34	0.0159
GBSCOH1_RS04815	O-sialoglycoprotein endopeptidase	849	8	4	0.55	14.13	12	0.0323
GBSCOH1_RS04970	CAAX amino protease	879	10	3	0.96	14.79	22	0.0195
GBSCOH1_RS05065	hypothetical protein	267	5	2	0.65	26.22	26	0.0049
GBSCOH1_RS05070	hypothetical protein	384	8	0	8.00	20.83	8	0.0102
GBSCOH1_RS05095	cell division protein FtsK	4428	45	16	0.75	13.78	461	1.43×10^{-5}*
GBSCOH1_RS05115	phage infection protein	2475	21	6	0.92	10.91	596	0.0440
GBSCOH1_RS05130	carbamoyl phosphate synthase large subunit	3183	31	6	1.48	11.62	101	0.0092
GBSCOH1_RS05135	carbamoyl phosphate synthase small subunit	1077	12	3	1.09	13.93	21	0.0206
GBSCOH1_RS05140	aspartate carbamoyltransferase	924	11	2	1.44	14.07	85	0.0276
GBSCOH1_RS05145	dihydroorotase	1293	14	4	0.99	13.92	29	0.0122
GBSCOH1_RS05155	orotidine 5'-phosphate decarboxylase	702	6	5	0.33	15.67	598	0.0208
GBSCOH1_RS05180	FbsA fibronectin-binding protein	657	9	5	0.50	21.31	21	0.0007
GBSCOH1_RS05215	luciferase	987	9	5	0.49	14.18	22	0.0216
GBSCOH1_RS05220	NADH-dependent flavin oxidoreductase	1200	8	7	0.33	12.50	27	0.0454
GBSCOH1_RS05275	serine hydroxymethyltransferase	1257	12	4	0.85	12.73	605	0.0349
GBSCOH1_RS05280	threonylcarbamoyl-AMP synthase	597	4	6	0.18	16.75	11	0.0179
GBSCOH1_RS05320	formate transporter	789	10	5	0.55	19.01	15	0.0014
GBSCOH1_RS05515	choline transporter	1548	16	3	1.66	12.27	51	0.0317
GBSCOH1_RS05555	hypothetical protein	660	8	3	0.61	16.67	40	0.0140
GBSCOH1_RS05595	hypothetical protein	249	4	3	0.44	28.11	24	0.0034
GBSCOH1_RS05600	membrane protein	246	3	3	0.33	24.39	23	0.0123
GBSCOH1_RS05745	beta-1,4-galactosyltransferase CpsG	474	12	2	1.49	29.54	18	2.53×10^{-5}*
GBSCOH1_RS05755	galactosyl transferase CpsE	1389	16	2	2.08	12.96	27	0.0228
GBSCOH1_RS05760	tyrosine-protein kinase CpsD	690	10	1	2.82	15.94	19	0.0187
GBSCOH1_RS06005	NAD(P)H nitroreductase	603	14	0	14.00	23.22	79	0.0003*
GBSCOH1_RS06045	phage tail protein	498	7	1	1.86	16.06	13	0.0393
GBSCOH1_RS06050	Lj965 prophage superinfection immunity protein	597	6	3	0.49	15.08	333	0.0420

GBSCOH1_RS06075	histidine triad (HIT) protein	2469	24	7	0.92	12.56	31	0.0059
GBSCOH1_RS06095	peptidase C5	3453	35	13	0.77	13.90	60	0.0001*
GBSCOH1_RS06200	amino acid transporter	1377	16	7	0.72	16.70	92	0.0005
GBSCOH1_RS06210	hypothetical protein	324	5	1	1.38	18.52	9	0.0395
GBSCOH1_RS06215	TetR family transcriptional regulator	549	8	1	2.04	16.39	16	0.0269
GBSCOH1_RS06410	peptidoglycan branched peptide synthesis protein	1086	14	3	1.23	15.65	24	0.0050
GBSCOH1_RS06530	metallophosphatase	1179	13	2	1.72	12.72	25	0.0401
GBSCOH1_RS06595	accessory Sec system protein translocase subunit SecY2	1254	13	4	0.92	13.56	31	0.0182
GBSCOH1_RS06600	Srr2 cell wall anchor	3258	24	27	0.28	15.65	351	2.57×10^{-6}*
GBSCOH1_RS06685	iron ABC transporter ATP-binding protein	795	8	5	0.43	16.35	33	0.0093
GBSCOH1_RS06785	glycosyl transferase	1149	11	6	0.49	14.80	32	0.0085
GBSCOH1_RS06830	hypothetical protein	1734	18	6	0.73	13.84	361	0.0047
GBSCOH1_RS06980	peptidase	1062	10	5	0.54	14.12	36	0.0185
GBSCOH1_RS06985	peptidase	4158	35	7	1.37	10.10	209	0.0437
GBSCOH1_RS07040	amidase	2040	19	8	0.67	13.24	47	0.0050
GBSCOH1_RS07065	lactoylglutathione lyase	393	4	3	0.34	17.81	23	0.0328
GBSCOH1_RS07105	ABC transporter ATP-binding protein	702	6	4	0.40	14.25	22	0.0456
GBSCOH1_RS07125	transglutaminase	1590	15	5	0.79	12.58	32	0.0226
GBSCOH1_RS07230	peptide ABC transporter permease	945	11	5	0.65	16.93	20	0.0031
GBSCOH1_RS07290	peptidyl-prolyl cis-trans isomerase	804	9	3	0.79	14.93	66	0.0227
GBSCOH1_RS07335	hypothetical protein	414	6	2	0.74	19.32	15	0.0154
GBSCOH1_RS07395	molecular chaperone	612	6	5	0.30	17.97	19	0.0084
GBSCOH1_RS07450	acetyltransferase	549	8	1	2.04	16.39	10	0.0269
GBSCOH1_RS07500	branched-chain amino acid ABC transporter permease	954	15	1	4.84	16.77	40	0.0034
GBSCOH1_RS07505	branched-chain amino acid ABC transporter permease	870	7	5	0.45	13.79	34	0.0377
GBSCOH1_RS07510	branched-chain amino acid ABC transporter substrate-binding protein	1167	14	3	1.41	14.57	34	0.0097
GBSCOH1_RS07590	non-canonical purine NTP pyrophosphatase	975	8	7	0.30	15.38	85	0.0093
GBSCOH1_RS07645	ABC transporter substrate-binding protein	858	10	3	0.90	15.15	16	0.0164
GBSCOH1_RS07705	CovS two-component sensor histidine kinase	1506	20	5	1.05	16.60	29	0.0004*
GBSCOH1_RS07735	V-type Na+ -ATPase subunit J	1380	19	3	1.93	15.94	315	0.0013
GBSCOH1_RS07770	alcohol dehydrogenase	1056	9	5	0.55	13.26	22	0.0349
GBSCOH1_RS07785	peptidase M20	1377	13	5	0.74	13.07	27	0.0212
GBSCOH1_RS07985	haloacid dehalogenase	1383	14	4	0.98	13.02	35	0.0220
GBSCOH1_RS07990	L-asparaginase	963	9	5	0.52	14.54	21	0.0180
GBSCOH1_RS08040	fructokinase	882	12	3	1.15	17.01	22	0.0039
GBSCOH1_RS08095	Xaa-Pro aminopeptidase	1068	11	3	1.04	13.11	605	0.0377
GBSCOH1_RS08110	excinuclease ABC subunit A	2829	23	9	0.75	11.31	38	0.0202
GBSCOH1_RS08130	Single-stranded DNA-binding protein 1	492	4	4	0.28	16.26	8	0.0371
GBSCOH1_RS08155	acid phosphatase	501	5	3	0.47	15.97	8	0.0405
GBSCOH1_RS08245	X-prolyl-dipeptidyl aminopeptidase	2286	18	8	0.59	11.37	45	0.0315
GBSCOH1_RS08270	cytochrome c oxidase assembly protein	1020	11	4	0.82	14.71	22	0.0134
GBSCOH1_RS08320	pyridine nucleotide-disulfide oxidoreductase	1173	10	5	0.57	12.79	28	0.0387
GBSCOH1_RS08335	amino acid lyase	1026	13	3	1.20	15.59	33	0.0065
GBSCOH1_RS08365	amidase	510	7	2	1.00	17.65	10	0.0179
GBSCOH1_RS08420	3'-5' exoribonuclease	942	13	6	0.54	20.17	32	0.0002*
GBSCOH1_RS08490	D-alanyl-lipoteichoic acid biosynthesis protein DltD	1263	15	2	1.94	13.46	21	0.0193
GBSCOH1_RS08570	carbohydrate kinase	1518	15	9	0.47	15.81	267	0.0009
GBSCOH1_RS08580	PTS galactitol transporter subunit IIC	1452	14	6	0.72	13.77	26	0.0096
GBSCOH1_RS08725	alkyl hydroperoxide reductase subunit F	1533	7	12	0.17	12.39	100	0.0292
GBSCOH1_RS08745	peptidase S66	954	11	4	0.77	15.72	23	0.0077
GBSCOH1_RS08815	hypothetical protein	639	6	4	0.40	15.65	12	0.0269
GBSCOH1_RS08960	hypothetical protein	2451	22	6	0.96	11.42	337	0.0253
GBSCOH1_RS09145	thymidylate kinase	587	7	3	0.59	17.04	14	0.0162
GBSCOH1_RS09150	16S rRNA (uracil(1498)-N(3))-methyltransferase	741	8	4	0.56	16.19	36	0.0130
GBSCOH1_RS09500	ABC transporter ATP-binding protein	903	11	3	1.02	15.50	17	0.0110
GBSCOH1_RS09575	5-methyltetrahydropteroyltriglutamate-- homocysteine S-methyltransferase	2238	21	6	0.98	12.06	39	0.0152
GBSCOH1_RS09805	damage-inducible protein A	1260	10	8	0.37	14.29	44	0.0096
GBSCOH1_RS09880	50S ribosomal protein L33	150	1	3	0.09	26.67	21	0.0288
GBSCOH1_RS09920	hypothetical protein	606	8	2	0.97	16.50	12	0.0196
GBSCOH1_RS09940	DNA-binding response regulator	672	8	3	0.73	16.37	15	0.0157
GBSCOH1_RS09980	glycine/betaine ABC transporter permease	1515	21	2	3.11	15.18	33	0.0018
GBSCOH1_RS09995	membrane protein	357	7	0	7.00	19.61	13	0.0210
GBSCOH1_RS10350	competence protein CglA	972	15	2	1.99	17.49	336	0.0017
GBSCOH1_RS10355	transcriptional regulator	222	3	3	0.27	27.03	8	0.0077
GBSCOH1_RS10360	hypothetical protein	488	7	1	1.83	16.39	15	0.0356

[1]Number of nonsynomyous mutations.

[2]Number of synonymous mutations.

[3]Normalized ratio of nonsynonymous to synonymous substitutions. In the absence of synonymous mutations, dN/dS is indicated as the number of nonsynonymous changes.

[4]Number of SNPs per kb.

[5]Number of isolates polymorphic in relation to the COH1 reference sequence.

[6]*P* values obtained with the exact Poisson test. Those with an asterisk indicate $P < 0.05$ after multiple testing correction with a false discovery rate (FDR) of 10%.

Supplementary Table 3. Non-coding regions with a statistically significant mutational bias.

Region	Upstream locus	Strand	Downstream locus	Strand	Mutations	Length (bp)	Density[1]	Isolates[2]	*P* value[3]
IR1	GBSCOH1_RS00010	+	GBSCOH1_RS00015	+	3	69	43.48	5	0.0283
IR2	GBSCOH1_RS00275	+	GBSCOH1_RS00280	+	4	123	32.52	5	0.0302
IR3	GBSCOH1_RS00370	-	GBSCOH1_RS00375	+	9	281	32.03	11	0.0017
IR4	GBSCOH1_RS00425	+	GBSCOH1_RS00430	+	6	268	22.39	13	0.0436
IR5	GBSCOH1_RS00825	+	GBSCOH1_RS00830	-	3	39	76.92	2	0.0063
IR6	GBSCOH1_RS00925	-	GBSCOH1_RS00930	-	9	341	26.39	11	0.0058
IR7	GBSCOH1_RS01080	+	GBSCOH1_RS01085	+	10	523	19.12	34	0.0291
IR8	GBSCOH1_RS01670	+	GBSCOH1_RS01675	-	5	207	24.15	9	0.0481
IR9	GBSCOH1_RS01765	+	GBSCOH1_RS01770	+	5	120	41.67	8	0.0061
IR10	GBSCOH1_RS01905	+	GBSCOH1_RS01910	+	6	180	33.33	73	0.0079
IR11	GBSCOH1_RS01940	+	GBSCOH1_RS01945	+	4	76	52.63	6	0.0062
IR12	GBSCOH1_RS02095	+	GBSCOH1_RS02100	+	11	102	107.84	84	6.67×10^{-9}*
IR13	GBSCOH1_RS02190	-	GBSCOH1_RS02195	+	10	162	61.73	120	4.73×10^{-6}*
IR14	GBSCOH1_RS02255	+	GBSCOH1_RS02260	+	4	108	37.04	607	0.0200
IR15	GBSCOH1_RS02405	+	GBSCOH1_RS02410	+	10	188	53.19	47	1.68×10^{-5}*
IR16	GBSCOH1_RS02415	+	GBSCOH1_RS02420	+	10	553	18.08	156	0.0399
IR17	GBSCOH1_RS02500	-	GBSCOH1_RS02505	+	7	246	28.46	27	0.0097
IR18	GBSCOH1_RS02505	+	GBSCOH1_RS02510	+	14	711	19.69	99	0.0090
IR19	GBSCOH1_RS02670	-	GBSCOH1_RS02675	+	8	124	64.52	76	3.05×10^{-5}*
IR20	GBSCOH1_RS03020	+	GBSCOH1_RS03025	+	5	141	35.46	6	0.0116
IR21	GBSCOH1_RS03025	+	GBSCOH1_RS03030	-	4	112	35.71	3	0.0225
IR22	GBSCOH1_RS03215	+	GBSCOH1_RS03220	-	6	120	50.00	6	0.0011
IR23	GBSCOH1_RS03415	+	GBSCOH1_RS03420	+	2	34	58.82	3	0.0415
IR24	GBSCOH1_RS03490	+	GBSCOH1_RS03495	+	13	207	62.80	35	1.55×10^{-7}*
IR25	GBSCOH1_RS03510	+	GBSCOH1_RS03515	+	4	116	34.48	35	0.0252
IR26	GBSCOH1_RS03715	-	GBSCOH1_RS03720	+	7	342	20.47	41	0.0458
IR27	GBSCOH1_RS03850	+	GBSCOH1_RS03855	+	6	186	32.26	400	0.0092
IR28	GBSCOH1_RS03870	+	GBSCOH1_RS03875	+	3	85	35.29	31	0.0475
IR29	GBSCOH1_RS03980	+	GBSCOH1_RS03985	+	4	100	40.00	11	0.0156
IR30	GBSCOH1_RS03995	-	GBSCOH1_RS04000	+	9	152	59.21	10	1.93×10^{-5}*
IR31	GBSCOH1_RS04165	+	GBSCOH1_RS04170	+	21	258	81.40	34	2.40×10^{-13}*
IR32	GBSCOH1_RS04195	+	GBSCOH1_RS04200	+	3	83	36.14	4	0.0448
IR33	GBSCOH1_RS04410	+	GBSCOH1_RS04415	+	4	97	41.24	6	0.0141
IR34	GBSCOH1_RS04495	+	GBSCOH1_RS04505	+	18	1088	16.54	21	0.0176
IR35	GBSCOH1_RS04550	+	GBSCOH1_RS04555	-	2	36	55.56	2	0.0460
IR36	GBSCOH1_RS04970	-	GBSCOH1_RS04975	-	10	105	95.24	15	1.00×10^{-7}*
IR37	GBSCOH1_RS05010	-	GBSCOH1_RS05015	-	9	239	37.66	250	0.0006*
IR38	GBSCOH1_RS05030	-	GBSCOH1_RS05035	-	6	201	29.85	6	0.0130
IR39	GBSCOH1_RS05060	-	GBSCOH1_RS05065	-	4	107	37.38	17	0.0194
IR40	GBSCOH1_RS05090	-	GBSCOH1_RS05095	-	1	4	250.00	20	0.0370
IR41	GBSCOH1_RS05160	-	GBSCOH1_RS05165	-	3	39	76.92	3	0.0063
IR42	GBSCOH1_RS05180	-	GBSCOH1_RS05185	-	14	688	20.35	26	0.0069
IR43	GBSCOH1_RS05335	-	GBSCOH1_RS05340	-	7	203	34.48	8	0.0036
IR44	GBSCOH1_RS05525	-	GBSCOH1_RS05530	-	6	107	56.07	23	0.0006*
IR45	GBSCOH1_RS05530	-	GBSCOH1_RS05535	+	12	478	25.10	37	0.0024
IR46	GBSCOH1_RS05535	+	GBSCOH1_RS05540	+	7	196	35.71	12	0.0029
IR47	GBSCOH1_RS05600	-	GBSCOH1_RS05605	-	11	411	26.76	18	0.0022
IR48	GBSCOH1_RS05870	-	GBSCOH1_RS05875	-	4	109	36.70	105	0.0206
IR49	GBSCOH1_RS05985	-	GBSCOH1_RS05990	-	8	168	47.62	11	0.0002*
IR50	GBSCOH1_RS06195	-	GBSCOH1_RS06200	-	1	4	250.00	1	0.0370
IR51	GBSCOH1_RS06210	-	GBSCOH1_RS06215	+	8	271	29.52	27	0.0048
IR52	GBSCOH1_RS06250	+	GBSCOH1_RS06255	+	4	91	43.96	3	0.0114
IR53	GBSCOH1_RS06325	-	GBSCOH1_RS06330	+	6	268	22.39	6	0.0436
IR54	GBSCOH1_RS06415	-	GBSCOH1_RS06420	-	6	211	28.44	119	0.0161
IR55	GBSCOH1_RS06600	-	GBSCOH1_RS06605	-	14	740	18.92	20	0.0123
IR56	GBSCOH1_RS06895	-	GBSCOH1_RS06900	-	3	82	36.59	2	0.0435
IR57	GBSCOH1_RS06955	+	GBSCOH1_RS06960	+	7	230	30.43	8	0.0069
IR58	GBSCOH1_RS06985	-	GBSCOH1_RS06990	-	7	242	28.93	13	0.0089
IR59	GBSCOH1_RS07165	-	GBSCOH1_RS07170	-	7	321	21.81	7	0.0346
IR60	GBSCOH1_RS07235	-	GBSCOH1_RS07240	-	15	388	38.66	29	7.03×10^{-6}*

IR61	GBSCOH1_RS07305	-	GBSCOH1_RS07310	+	5	163	30.67	5	0.0202
IR62	GBSCOH1_RS07395	-	GBSCOH1_RS07400	-	11	445	24.72	21	0.0040
IR63	GBSCOH1_RS07510	-	GBSCOH1_RS07515	-	11	154	71.43	42	3.97×10^{-7}*
IR64	GBSCOH1_RS07710	-	GBSCOH1_RS07715	-	8	234	34.19	41	0.0020
IR65	GBSCOH1_RS07955	-	GBSCOH1_RS07960	+	7	305	22.95	17	0.0274
IR66	GBSCOH1_RS08055	+	GBSCOH1_RS08060	-	3	83	36.14	2	0.0448
IR67	GBSCOH1_RS08135	-	GBSCOH1_RS08140	+	19	1093	17.38	34	0.0095
IR68	GBSCOH1_RS08205	+	GBSCOH1_RS08210	-	2	34	58.82	1	0.0415
IR69	GBSCOH1_RS08295	+	GBSCOH1_RS08300	+	7	159	44.03	11	0.0009*
IR70	GBSCOH1_RS08330	+	GBSCOH1_RS08335	-	8	360	22.22	8	0.0227
IR71	GBSCOH1_RS08390	-	GBSCOH1_RS08395	-	6	208	28.85	6	0.0151
IR72	GBSCOH1_RS08545	-	GBSCOH1_RS08550	-	8	376	21.28	11	0.0283
IR73	GBSCOH1_RS08580	-	GBSCOH1_RS08585	-	4	97	41.24	4	0.0141
IR74	GBSCOH1_RS08640	+	GBSCOH1_RS08645	-	2	22	90.91	2	0.0187
IR75	GBSCOH1_RS08805	-	GBSCOH1_RS08810	+	19	277	68.59	51	5.64×10^{-11}*
IR76	GBSCOH1_RS08825	+	GBSCOH1_RS08830	+	22	1197	18.38	69	0.0030
IR77	GBSCOH1_RS08910	-	GBSCOH1_RS08915	-	7	128	54.69	12	0.0003*
IR78	GBSCOH1_RS08955	-	GBSCOH1_RS08960	-	7	248	28.23	18	0.0101
IR79	GBSCOH1_RS08975	-	GBSCOH1_RS08980	+	6	208	28.85	6	0.0151
IR80	GBSCOH1_RS08990	-	GBSCOH1_RS08995	-	6	254	23.62	39	0.0351
IR81	GBSCOH1_RS09060	-	GBSCOH1_RS09065	-	8	246	32.52	11	0.0027
IR82	GBSCOH1_RS09105	-	GBSCOH1_RS09110	-	6	154	38.96	12	0.0038
IR83	GBSCOH1_RS09565	-	GBSCOH1_RS09570	-	7	346	20.23	8	0.0482
IR84	GBSCOH1_RS09645	-	GBSCOH1_RS09650	-	10	313	31.95	18	0.0010*
IR85	GBSCOH1_RS09900	+	GBSCOH1_RS09905	+	5	201	24.88	6	0.0435
IR86	GBSCOH1_RS09960	-	GBSCOH1_RS09965	+	4	102	39.22	9	0.0166

[1]Number of SNPs per kb.

[2]Number of isolates polymorphic in relation to the COH1 reference sequence.

[3]*P* values obtained with the exact Poisson test. Those with an asterisk indicate $P < 0.05$ after multiple testing correction with a false discovery rate (FDR) of 10%.

Supplementary Table 4. Genes that acquired more than five inactivating mutations.

Locus	Product	Total	NS[1]	FS[2]	Isolates[3]
GBSCOH1_RS06985	peptidase	10	5	5	16
GBSCOH1_RS06095	peptidase C5	8	1	7	103
GBSCOH1_RS07380	hypothetical protein	7	3	4	603
GBSCOH1_RS04515	hypothetical protein	7	3	4	8
GBSCOH1_RS07235	nickel ABC transporter	7	2	5	26
GBSCOH1_RS08830	hypothetical protein	7	0	7	24
GBSCOH1_RS05095	cell division protein FtsK	6	5	1	370
GBSCOH1_RS03450	hypothetical protein	6	4	2	68
GBSCOH1_RS07215	peptide ABC transporter ATP-binding protein	6	4	2	45
GBSCOH1_RS01045	4-diphosphocytidyl-2C-methyl-D-erythritol kinase	6	3	3	502
GBSCOH1_RS08960	hypothetical protein	6	3	3	6
GBSCOH1_RS04480	type II CRISPR RNA-guided endonuclease Cas9	6	2	4	7
GBSCOH1_RS09590	hypothetical protein	6	1	5	11
GBSCOH1_RS04855	ABC transporter substrate-binding protein	6	1	5	6

[1]Number of nonsense mutations.

[2]Number of frameshift mutations.

[3]Number of isolates polymorphic in relation to the COH1 reference sequence.

Supplementary Table 5. Accessory genes associated with CC17.

Gene	Locus[1]	Product	CC17[2]	Other[3]
AG1	BM110_01674	hypothetical protein	3	0
AG2	BM110_01670	hypothetical protein	3	0
AG3	BM110_01671	hypothetical protein	3	0
AG4	BM110_01673	unknown	3	0
AG5	BM110_00282	hypothetical protein	3	0
AG6	BM110_01672	hypothetical protein	3	0
AG7	BM110_01422	serine-rich repeat protein Srr2	3	0
AG8	BM110_01071	hypothetical protein	3	0
AG9	BM110_01499	surface-anchored serine protease	3	1
AG10	BM110_01486	hypothetical protein	3	1
AG11	BM110_01498	LPXTG-motif cell wall anchor domain protein	3	1
AG12	BM110_01420	hypothetical protein	3	1
AG13	BM110_01421	preprotein translocase SecY family protein,putative	3	1
AG14	BM110_01418	hypothetical protein	3	1
AG15	BM110_01419	hypothetical protein	3	1
AG16	BM110_00071	Chromosome segregation ATPase	3	1
AG17	BM110_00072	hypothetical protein	3	1
AG18	BM110_01412	glycosyl transferase, family 8	3	1
AG19	BM110_01414	hypothetical protein	3	1
AG20	BM110_01413	hypothetical protein	3	1
AG21	BM110_01121	hypothetical protein	3	1
AG22	BM110_01120	hypothetical protein	3	1
AG23	BM110_01122	protein of unknown function	3	1
AG24	BM110_01416	glycosyl transferase, group 1	3	1
AG25	BM110_01415	hypothetical protein	3	1
AG26	BM110_00857	reverse transcriptase and maturase	3	1
AG27	BM110_00380	acetyltransferase, GNAT family	3	1
AG28	BM110_01119	lipoprotein	3	1
AG29	BM110_01118	lipoprotein	3	1
AG30	BM110_01417	preprotein translocase, SecA chain	3	1
AG31	09mas018883_00475	hypothetical protein	3	1
AG32	BM110_01990	phage transcriptional repressor	3	1
AG33	BM110_02063	hypothetical protein	3	2
AG34	2603V-R_01280	hypothetical protein	3	2
AG35	2603V-R_00487	surface protein Rib	3	2
AG36	2603V-R_01281	hypothetical protein	3	2
AG37	2603V-R_01801	hypothetical protein	3	3
AG38	2603V-R_02072	glycosyl transferase family protein	3	4
AG39	2603V-R_02071	glycosyl transferase family protein	3	4
AG40	BM110_01257	capsular polysaccharide repeating-unit polymerase	3	5
AG41	09mas018883_00499	lipoprotein	3	5
AG42	BM110_00126	hypothetical protein	3	5
AG43	09mas018883_01053	lipoprotein, putative	3	6
AG44	BM110_01258	glycosyl transferase CpsG (V)	3	6
AG45	BM110_00127	lipoprotein, putative	3	6
AG46	09mas018883_01052	hypothetical protein	3	6
AG47	09mas018883_00498	hypothetical protein	3	6
AG48	09mas018883_00497	DNA-damage-inducible protein J, putative	3	6
AG49	BM110_00128	Hypothetical protein	3	6
AG50	BM110_00585	hypothetical protein	3	7
AG51	09mas018883_01950	Unknown	3	7
AG52	A909_01436	surface protein Spb1	3	8
AG53	138P_00417	lipoprotein	3	8
AG54	BM110_02061	radical SAM domain protein protein	2	0
AG55	BM110_02106	hypervirulent GBS adhesin HvgA	2	0
AG56	BM110_02060	SREBP site 2 protease family protein	2	0
AG57	BM110_02059	Exported signaling peptide, YydF/SAG_2028 family protein	2	0
AG58	COH1_00266	type III restriction system methylase	2	0
AG59	COH1_00267	type II restriction endonuclease	2	0

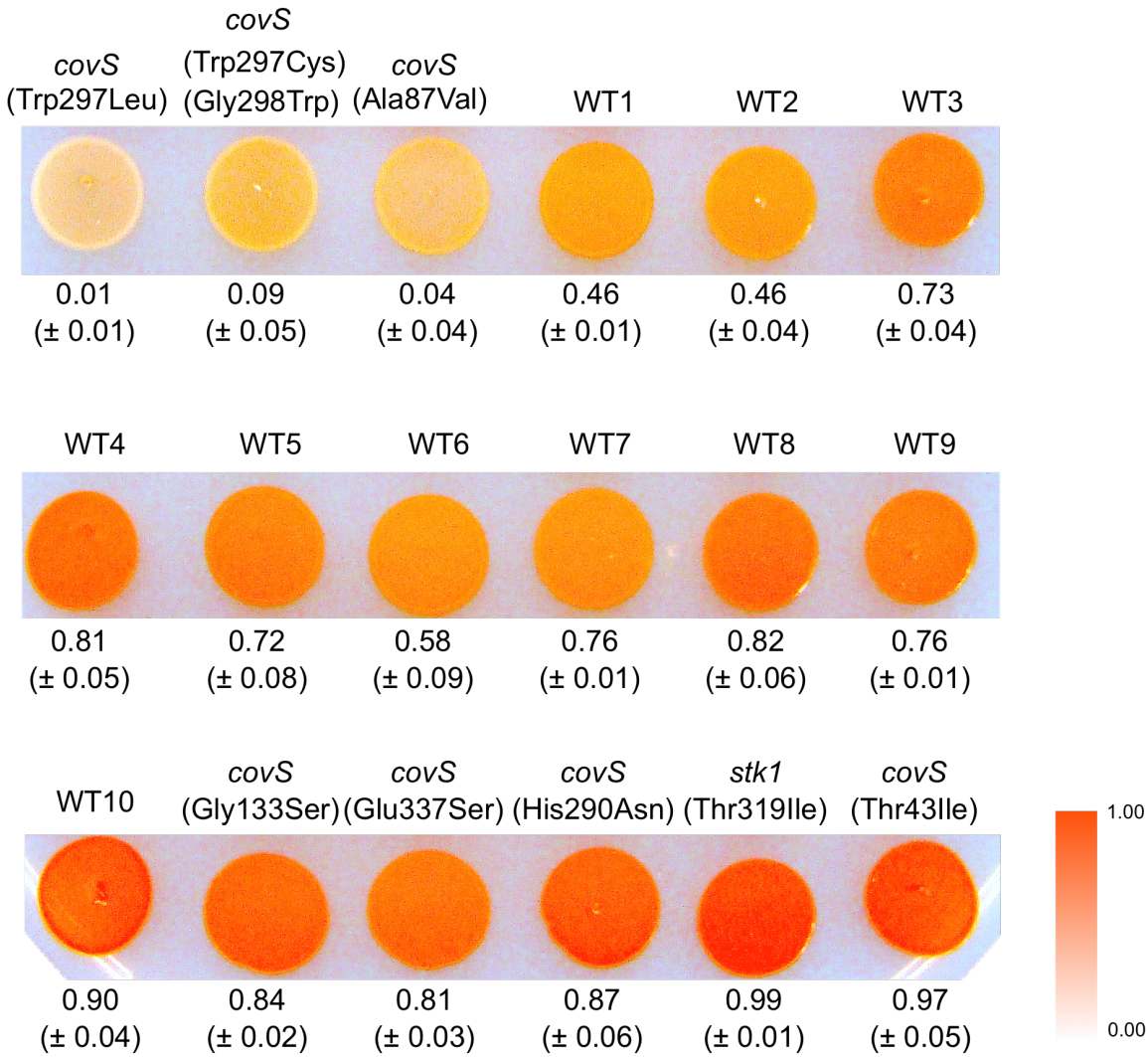
AG60	COH1_00268	hypothetical protein	2	0
AG61	COH1_00927	hypothetical protein	2	0
AG62	COH1_00928	hypothetical protein	2	0
AG63	COH1_00929	hypothetical protein	2	0
AG64	COH1_00941	hypothetical protein	2	0
AG65	COH1_01220	hypothetical protein	2	0
AG66	COH1_01221	FtsK/SpoIIIE family protein	2	0
AG67	COH1_01218	site-specific recombinase, phage integrase family	2	0
AG68	COH1_01219	Unknown-related protein	2	0
AG69	COH1_00946	hypothetical protein	2	0
AG70	COH1_00947	site-specific recombinase, phage integrase family	2	0
AG71	COH1_00944	putative sigma-70 family protein	2	0
AG72	COH1_00945	hypothetical protein	2	0
AG73	COH1_01222	hypothetical protein	2	0
AG74	COH1_01223	hypothetical protein	2	0
AG75	COH1_01225	type II restriction enyme eco47ii	2	0
AG76	COH1_01224	putative serine /threonine protein kinase	2	0
AG77	COH1_01226	C-5 cytosine-specific DNA methylase	2	0
AG78	COH1_00930	hypothetical protein	2	0
AG79	COH1_00931	transcriptional regulator, Cro /CI family	2	0
AG80	COH1_00269	ATP-dependent DNA helicase recG	2	0
AG81	COH1_00943	Tn916, transcriptional regulator, putative	2	0
AG82	COH1_00939	membrane protein, putative	2	0
AG83	COH1_00938	ATP /GTP-binding protein, putative	2	0
AG84	COH1_00937	hypothetical protein	2	0
AG85	COH1_00936	hypothetical protein	2	0
AG86	COH1_00935	hypothetical protein	2	0
AG87	COH1_00934	hypothetical protein	2	0
AG88	COH1_00932	hypothetical protein	2	0
AG89	COH1_00264	Tn5252, Orf 9 protein	2	0
AG90	COH1_00265	hypothetical protein	2	0
AG91	COH1_00259	site-specific recombinase, phage integrase family	2	0
AG92	COH1_00263	Tn5252, Orf 9 protein	2	0
AG93	COH1_00262	Phage replication initiation	2	0
AG94	COH1_00261	hypothetical protein	2	0
AG95	COH1_00260	transcriptional regulator, Cro /CI family,putative	2	0
AG96	A909_02095	hypothetical protein	2	1
AG97	2603V-R_02029	Cro/CI family transcriptional regulator	2	1
AG98	2603V-R_02030	hypothetical protein	2	1
AG99	2603V-R_02027	hypothetical protein	2	1
AG100	2603V-R_02028	hypothetical protein	2	1
AG101	2603V-R_02012	cell wall surface anchor family protein	2	1
AG102	BM110_02047	transcriptional regulator, Cro /CI family	2	2
AG103	BM110_02046	transcriptional regulator, Cro /CI family	2	2
AG104	BM110_02045	hypothetical protein	2	2
AG105	BM110_02052	CAMP factor	2	2
AG106	BM110_02049	lipoprotein, putative	2	2
AG107	BM110_02040	hypothetical protein	2	2
AG108	09mas018883_00281	Hypothetical protein	1	27
AG109	09mas018883_00283	FIG01119511: hypothetical protein	1	27
AG110	09mas018883_00279	Virulence factor esxA	1	27
AG111	09mas018883_00248	lipoprotein	0	21
AG112	09mas018883_01497	cell wall surface anchor family protein	0	21
AG113	09mas018883_01074	ESAT-6/Esx family secreted protein EsxA/YukE	0	26
AG114	09mas018883_01494	Putative CDP-glycerol:glycerophosphate glycerophosphotransferase	0	26
AG115	09mas018883_01491	glycosyl transferase, putative	0	26
AG116	09mas018883_01492	glycosyl transferase family protein	0	26
AG117	09mas018883_01493	Glycosyl transferase, family 8	0	26
AG118	09mas018883_01490	Preprotein translocase SecY2 subunit	0	26
AG119	09mas018883_01489	Accessory secretory protein Asp1	0	27
AG120	09mas018883_01487	Accessory secretory protein Asp3	0	27

AG121	09mas018883_01496	Nucleotide sugar synthetase-like protein	0	27
AG122	09mas018883_01488	Accessory secretory protein Asp2	0	27
AG123	09mas018883_01495	Putative CDP-glycerol:glycerophosphate glycerophosphotransferase	0	27
AG124	138P_00447	phosphorylase	0	28
AG125	09mas018883_01498	Unknown	0	28
AG126	09mas018883_01484	GftB: Glycosyl transferase, family 8	0	28
AG127	09mas018883_01474	hypothetical protein	0	28
AG128	09mas018883_01485	Poly(glycerol-phosphate) alpha-glucosyltransferase GftA	0	28
AG129	09mas018883_01486	Protein export cytoplasm protein SecA2 ATPase RNA helicase	0	28
AG130	09mas018883_01004	hypothetical protein	0	29

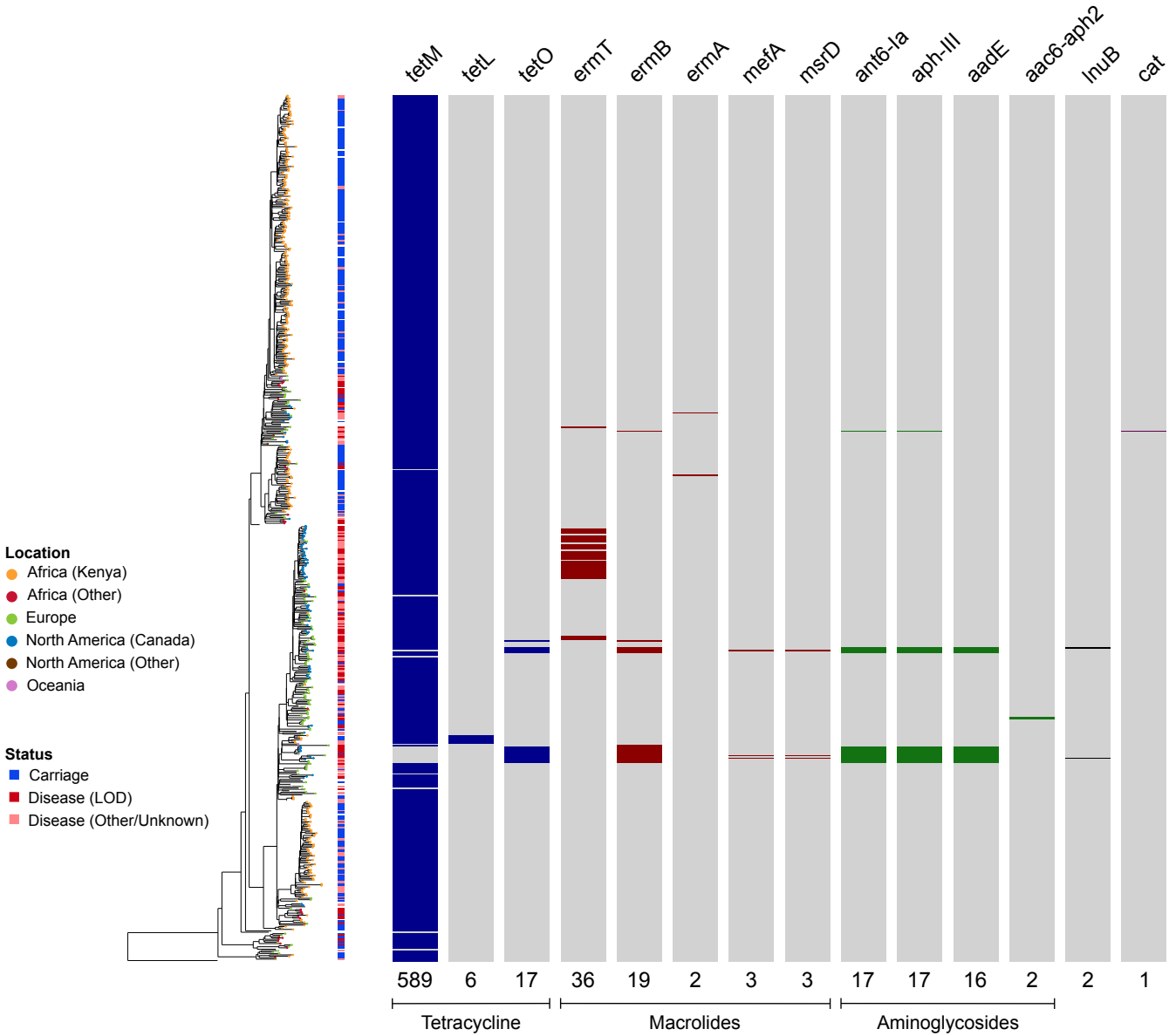
[1]Locus number based on gene annotation with Prokka (Seeman, et al. 2014).

[2]Number of CC17 genomes, out of the three analysed, in which the gene is present.

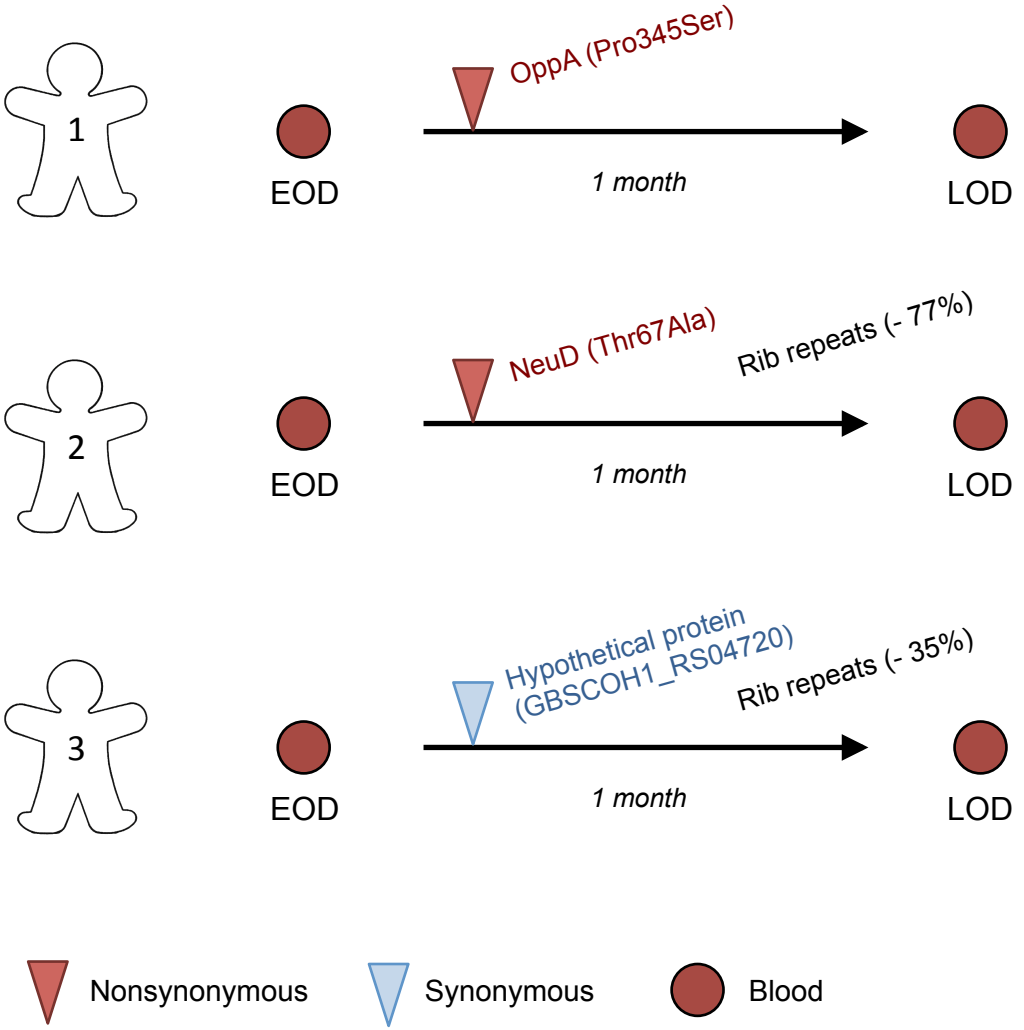
[3]Number of non-CC17 genomes, out of the 29 analysed, in which the gene is present.



Supplementary Figure 1. Pigmentation of strains with *covRS*-related mutations.
Pigment production of 18 different CC17 strains. Strains WT1-10 correspond to isolates with no mutations in genes assumed to affect the activity of the CovRS system (*covR, covS*, *abx1* or *stk1*). For the remaining strains, mutations potentially affecting *covRS* are indicated above each orange spot. Strains WT1/2/4/5/6/7/8/9 were obtained from carriers while all others were collected from infection. Average values and standard deviation from four independent experiments are depicted below each isolate. Ratios ranging from 0 to 1 were the result of normalization against the sample with the highest intensity in each test, calculated with imageJ (https://imagej.nih.gov/ij).



Supplementary Figure 2. Genetic characterization of antibiotic resistance.
Antibiotic resistance determinants present in the ResFinder[39] database, detected in both the sequencing reads and the assembled genomes of the CC17 strains according to their core-genome phylogeny. The number of isolates with each antibiotic resistance gene is depicted below each column. Gene absence is represented in grey, while the remaining colours illustrate the different classes of their corresponding antibiotics.



Supplementary Figure 3. Genomic evolution between EOD and LOD. Mutations differentiating three pairs of isolates collected one-month apart from different blood cultures of infected newborns. The name of the gene affected by the mutation and its effect on the protein sequence is represented according to the figure key. Percentage values alongside Rib represent change in the normalized sequencing coverage.

Discussion and future perspectives

The rise of genomics throughout the last 20 years revolutionized the fields of molecular microbiology, population genetics and microbial ecology. The ability to study the evolution of bacterial pathogens at an unprecedented resolution has paved the way for various methodological and scientific advances. Importantly, it has improved our understanding of many bacterial physiological processes, and of microbial interactions within communities and with their surrounding environment.

Throughout this work, we harnessed the power of whole-genome sequencing to gain new insights into the evolution of GBS and of its differential adaptation to humans and bovines. We observed a stark contrast in the evolutionary pressures present within the host environments that have shaped the distinct genomic landscapes observed for human- and bovine-specific populations. We were able to study these processes at multiple evolutionary scales, from a single strain level to a clonal- and population-wide analysis. Investigating the transition of GBS from carriage to disease in the human host, following maternal transmission, allowed us to focus on the rapid adaptation of this species within a short timeframe, and on the genomic differences between very closely related isolates coming from distinct clinical states. On the other hand, analysis of the hypervirulent CC17 disseminating worldwide for over 50 years provided us with a broader picture of the evolution of a well-adapted and clinically important population. At the same time, studying the bovine-specific clone persisting in Portuguese dairy farms offered a parallel overview of the adaptation of GBS to a substantially different environment.

Colonization and persistence in the bovine host

The prevalence of GBS mastitis in Portugal represents a recurring problem to the dairy industry in the country. The genomic changes accompanying the persistence of one single CC61 clone for over 20 years in Portuguese farms revealed important insights into the selective pressures present in the bovine environment. Indeed, analysing a closely related population with the same genomic background allowed us to more easily detect

parallel evolution in response to similar evolutionary pressures. To our knowledge, this was the first population genomics study carried out to obtain possible clues as to how GBS has been able to successfully colonize the bovine ecosystem. We were able to identify the acquisition of genes/mutations leading to modifications with a fitness benefit, while also predicting the functions being lost during steady colonization of the bovine udder. The most evident indication of positive selection was among mutations within an iron/manganese transporter system. Not only were the mutations preferentially located in a specific transcriptional terminator and affecting all the CC61 strains, but further expression analysis suggested a potential impact of these mutations in the regulation of iron/manganese uptake. In an *in vivo* experiment, *S. uberis* strains devoid of the orthologous manganese transporter MtuA were shown to be unable to infect lactating dairy cows (Smith et al 2003).

Although we discovered important adaptive changes that may confer an evolutionary advantage, the CC61 lineage in general is likely approaching an adaptive peak in the bovine host owing to its long-term colonization. It is likely that one of the critical events in the initial transition of GBS to bovines was the acquisition of a functional lactose operon (Richards et al 2013). Therefore, to complement the evolutionary picture inferred for the bovine-derived CC61 clone, we also studied in parallel the evolution of a human lineage of CC2 that underwent a recent host switch to cattle (Figure S2 of Publication nº1). Since bacterial populations experience a period of rapid adaptation when colonizing a new environment (Didelot et al 2016), this CC2 lineage represented a more straightforward example to study the evolution of GBS in the bovine ecosystem. Fittingly, some evolutionary changes most distinctly observed in the CC61 clone were replicated in the CC2 lineage, such as the loss of the capsule and the *vexp* region. This provided further evidence of the strong and consistent effect of evolutionary pressures in the bovine host. However, selective pressures towards the modification of iron/manganese uptake systems were not detected in the CC2 lineage, meaning that evolution of this transport pathway is more important within the genetic repertoire of CC61. In other species able to infect cattle, different functions have been implicated in the ability to survive in this environment. The biosynthesis of vitamin B5 was reported as a crucial function for bovine-derived *Campylobacter* (Sheppard et al 2013b), while other metabolic determinants were seen as part of the core genetic repertoire of a

prevalent clade of mastitis-causing *E. coli* (Goldstone et al 2016). Understandably, the evolution of metabolic pathways is differentially selected in distinct backgrounds, and depends on the inherent properties of each species or clone.

Transmission and evolution in the human host

The analysis of the human-adapted GBS revealed important insights into the aetiology of GBS neonatal infections and the genetic traits conferring a selective advantage during colonization of the human host. By comparing the genomic makeup of strains from infected newborns and their mothers we showed that in particular instances the transmission from mother to child during or after birth is followed by the selection of specific pathoadaptive mutations. Interestingly, we observed that strains collected from LOD presented proportionally more genomic differences than those from EOD. We speculate that this stems from the fact that some strains from LOD might have colonized the neonate asymptomatically, for instance in the gastrointestinal tract, for a limited period before triggering an infection, which created opportunities for the emergence and selection of new adaptive mutations. However, the sampling strategy used and the relatively small number of pairs tested limits some of the interpretations we are able to draw from this work. Notwithstanding, we deduce that most genomic variants are probably under the effects of positive selection as we see a particular bias in mutation type and in the functions of the genes affected. At the same time, we detected an almost equal number of mother-child GBS pairs with no genotypic differences, meaning that GBS is also inherently able to cause disease under specific circumstances. In fact, GBS is well known as an opportunistic pathogen, so immune deficiencies in the host allow this bacterium to progress to a state of disease. Furthermore, even without immune or pathogen-dependent factors, the sheer amount of bacterial load transferred during transmission of GBS could be enough to result in infection. We conclude that the cause of GBS neonatal infections is likely multifactorial and that adaptive evolution towards increased virulence might just be one of the many possible reasons for disease progression. This intricate relationship between microbe and host in determining disease outcome was reinforced by our detailed evolutionary analysis of the human-specific and hypervirulent CC17. In this work we observed that the difference between

GBS strains from carriage and disease is not well defined, as isolates from infection can emerge from independent genetic backgrounds. Given that disease strains are short-lived (ostensibly not further transmitted from infected infants), the majority of circulating lineages in humans are most likely carried asymptomatically. This supports the notion that genomic mutations that may lead to increased virulence have been acquired or selected during infection, such as after shifting from colonized mothers to infected newborns. Based on the mutation directionality, comparative genomic analysis raised an intriguing hypothesis in that GBS strains may be, to some extent, equally transmitted between mothers and newborns. In the child, as previously discussed, the most likely transmission route is through the birth canal or through aspiration of contaminated amniotic fluid after birth. In the mother, further contamination through breastfeeding can occur if the newborn had been previously colonized in the oral cavity by GBS either from the mother herself or another transient source. Nosocomial transmission of GBS has long been reported (Hastings et al 1981) and can occur through direct contact with the hospital staff, other patients within the ward or contaminated equipment. Along the same line, another question that arises is how GBS is transmitted asymptomatically throughout the human population. Sources for carriage and disease in adults are unknown, but might be related to transient contamination of skin and objects by strains from the gastrointestinal reservoir.

Studying the evolutionary forces acting on a bacterial population allows us to investigate genes or mutations pivotal to the successful colonization of a specific environment. In GBS, we observed that various systems evolve to possibly improve its ability to survive in the human host. Functions related to metabolism, cell adhesion, regulation and immune evasion were among the most recurrently affected in GBS strains from human origin. This means that there is not a single factor that is determinant to the dissemination of GBS in humans, but there are instead various alternative strategies that can be selected within different backgrounds depending on the specific properties of each colonized host. It should also be noted that since the global analysis of CC17 was based mostly on the sequences available from previous studies, some biases might be introduced because of the way the samples were selected and isolated in each study. Nonetheless, mutations among transcriptional regulators, and in particular CovRS, were consistently over-represented in the context of disease. Given the global impact of this

regulator in the expression of virulence-associated genes, little evolution of its protein structure may have dramatic effects on the virulence or colonization potential of GBS. Accordingly, phenotypic analysis suggested that some variants might be linked to a differential expression of genes associated with pathogenic traits of GBS. Evolution of regulatory networks has been reported as an important adaptive strategy in other human pathogens both *in vitro* (Philippe et al 2007) and *in vivo* (Marvig et al 2013, Yang et al 2011).

Besides virulence regulation, genetic variation related to immune evasion was recurrently selected in the course of disease. In this respect, we observed a decreased expression of the immunogenic surface protein Rib in a GBS strain transferred from mother to child. Similarly, the detailed analysis of CC17 showed that disease-associated strains possess a Rib with fewer repeated domains compared to that in carriage strains. Due to the immunogenic properties of this surface protein and their direct correlation with the number of repeats, we deem that phase variation in Rib is highly sensitive to immune selective pressures and represents a frequent adaptive strategy in GBS. This has been previously observed in a mouse model of infection (Madoff et al 1996) and is especially disconcerting given that Rib has been considered a potential vaccine target against GBS (Stalhammarcarlemalm et al 1993). Phase variation has been much discussed as a general mechanism of bacterial adaptation (Bayliss 2009, Moxon et al 2006). Particularly in *Neisseria meningitidis*, a similar trend towards the reduction of surface protein repeats was observed during long-term colonization (Alamro et al 2014).

Evolution of GBS in the course of disease seems to mostly affect genes with known and well-documented functions. However, our genomic predictions also underlined the importance of the evolution of particular genes that have not been extensively studied. In this regard, several transcriptional regulators that evolved in the mother-to-child transition might affect other targets of interest. Furthermore, the protein similar to Zoocin A from *S. zooepidemicus*, together with other peptidases or hypothetical proteins with frequent mutations, may also play critical roles that have been overlooked until now.

Evolutionary history and host-specific adaptation

Analysing the specificities of GBS in humans and bovines provides a better understanding of colonization and disease progression in these environments. Equally important, however, is leveraging this knowledge to make inferences on inter-host transmission and the evolutionary history of the species. Thus, from this work we were able to obtain insightful data complementing what is currently known about the genomics of GBS host adaptation. Globally, our evolutionary analyses revealed that the rate of accumulation of new substitutions in GBS is not constant across the species. Not only is the evolutionary rate higher in bovine-derived strains, but it also varies between lineages from the same host species. This tells us that the population size, environmental conditions and selective pressures between each individual host reservoir may vary, in turn differentially affecting the evolution of each GBS population.

In contrast to the human-specific clones, we observed that antibiotic resistance has not been a major selective force for the expansion of bovine-adapted GBS. Instead, the transfer of bacteriocin (e.g. macedocin) production/resistance genes, among animal pathogens, seems to have had a more significant impact in the dissemination of the bovine-specific clone in Portugal. Yet, we show that it probably had a transient selective advantage for this CC61 lineage, as macedocin resistance was later lost by a large fraction of the population. Tetracycline was introduced around the 1950s and was used as a broad-spectrum antibiotic to treat a variety of human infections. Since the presence of the tetracycline resistance determinant *tetM* is widespread across the human population and absent from the majority of bovine isolates, the main divergence event separating these two populations probably predates the acquisition of tetracycline resistance among human strains. Therefore, the occasional presence of bovine-isolated strains with the *tetM* gene either corresponds to a new acquisition within the bovine environment or to recent host-switches from humans to bovines (Haenni et al 2010).

Homologous recombination was described as a major adaptive force among human clones (Brochet et al 2008b, Da Cunha et al 2014). As this is not the case for the human-specific CC17 and the bovine-adapted CC61, they represent prime examples to perform genotype-to-phenotype association studies of bacterial evolution in which vertically

acquired mutations are the main source of diversity. Both in CC17 and CC61, variation within transcriptional regulators such as CovRS represents a recurrent adaptive strategy. This means that CovRS is probably important for adaptation regardless of the host environment. Most notably is that environmental pressures from each specific host promoted a recurrent degradation and loss of a distinct set of genes. Surprisingly, among the most prevalent pseudogenes in the bovine-derived clones were regions that have been implicated in the virulence of GBS (Brochet et al 2008b, Carlin et al 2009, Mistou et al 2009). However, studies describing the important role of these genetic traits have only been carried out in the context of human infection, so we reason that they might be most beneficial in the human ecosystem. This is further confirmed by our similar genomic analysis of the CC17 population, which showed that the genes most degraded in the bovine clones are not the ones predominantly decaying in the human-specific population. In this regard, we did not observe a recurrent inactivation of particularly important traits among the human-adapted CC17 strains, as some of the most inactivated genes remain uncharacterized. Besides revealing different adaptive requirements and selective constraints within humans and cattle, distinct patterns of pseudogenization provides us with clues on the host evolutionary history of the species and on the potential of inter-host transmission in GBS. Considering that some of the regions being lost in bovine-derived strains are maintained in the human population, the common ancestor that initially acquired all these traits possibly inhabited the human intestinal tract or a similar environment. Also, the recurrent decay of genetic traits important for survival in humans suggests that bovine GBS might not be able to as efficiently colonize the human host. Thus, the transmission of GBS from bovines to humans would most likely be unsuccessful. Even still, compared to the fish-adapted population of GBS, this degenerative process in the bovine host is not as pronounced. Following the expansion of the fish-specific ST260/261 lineage, reductive evolution led to an extensive decay of the core-genome of GBS. This is probably reflective of distinctive selective forces in the fish environment and of a longer adaptive period, which suggests that the divergence between human and bovine GBS populations probably occurred more recently (Rosinski-Chupin et al 2013).

Combining both previous and current observations, we propose the following model for the host evolutionary history of GBS (Fig. 14). The divergence of GBS and its adaptation

to fish occurred first, marked by the emergence of the ST260/261 lineage. Subsequently, from a common ancestor inhabiting a human-like environment (e.g. the gastrointestinal tract), the CC61 lineage expanded in cattle and adapted to the bovine udder, mainly through the acquisition of the lactose operon. Then, among the strains that remained in the human host, those that eventually acquired tetracycline resistance outcompeted and replaced the existing population, leading to the emergence of the major human clones seen today. Presently, human CCs can still be stably transmitted to bovines if gained the ability to metabolise lactose. However, they are likely to be outcompeted by CC61 clones, which are better adapted to the bovine environment due to additional adaptive changes that developed within the core-genome. Therefore, the host-restricted ST260/261 and CC61, having streamlined their genomic architecture to their respective environments, are probably incapable of switching hosts. In contrast, the specific distribution of CC7 strains, which include isolates from human, bovine and fish origin, likely reflects a more generalist background and an increased ability to readily switch host environments.

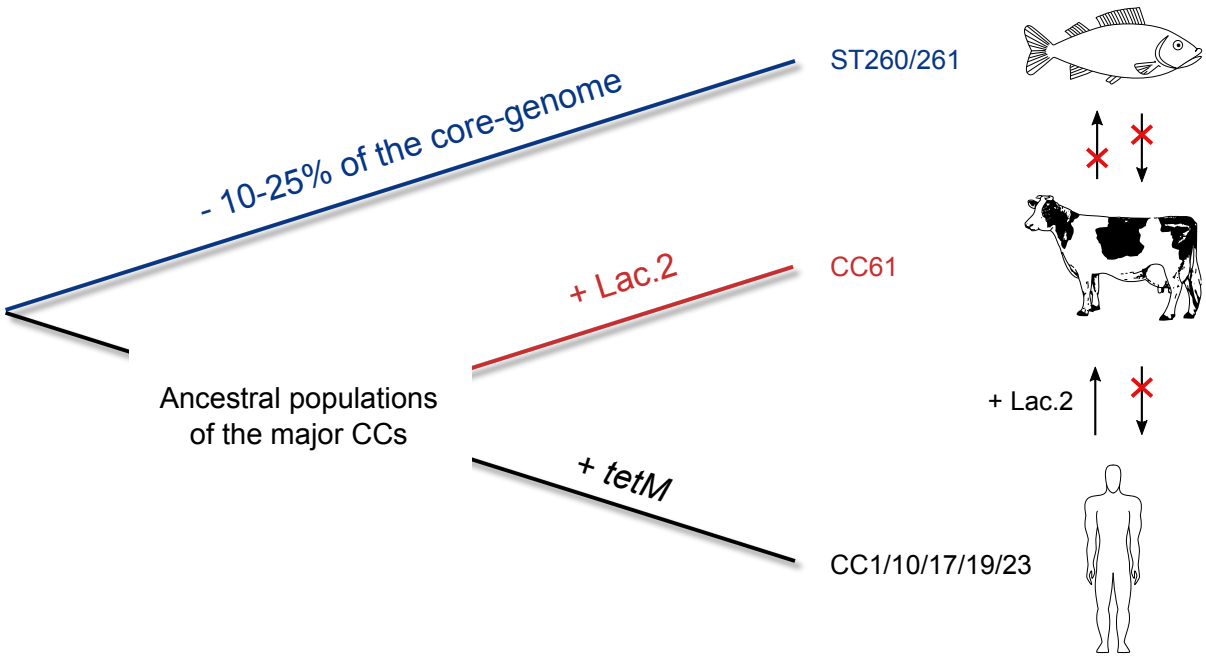


Figure 14. Model of the evolutionary history and inter-host transmission of GBS.

Practical implications and future directions

In the context of neonatal disease, the incidence of CC17 and of LOD continues to be a problematic issue. Even though the capsular polysaccharide and Rib have been discussed as promising vaccine antigens to prevent GBS colonization, their underlying diversity might limit their therapeutic efficiency. The search for other vaccine targets is still underway and will likely need closer investigation (Hughes et al 2002). In this regard, functional experimentation of genes highlighted in this work as important for the dissemination and progression of disease could reveal promising preventive strategies for GBS colonization. This leads into the fact that genomics can only go so far. In spite of being an excellent approach to find patterns of interest from large volumes of data, bringing these results into the clinic requires further experimental validation. To this end, it will not only be important to assess the function of genes under the effects of natural selection, but also evaluate the impact of the most recurrent pathoadaptive mutations we detected. For instance, the biofilm forming and fibrinogen-binding ability of human strains could be tested and related to the recurrent mutations seen in the PI-1 genomic island and the *fbs* genes. Likewise, the effect of modifications in the capsule operon could be measured by directly assessing the function of specific genes, such as the sialic acid production and O-acetyltransferase activity of NeuD. Additionally, the comparison between mother and child GBS samples will need to be extended and validated with a larger number of isolates. This will bring more reliable data on the transmission dynamics of GBS and on the evolutionary pressures acting during its switch from commensal to pathogen.

In Portugal, given that GBS mastitis continues to be a recurring problem with significant financial repercussions, we highlight the need to more closely monitor infected cattle in order to prevent relapse or reinfection. Indeed, both in bovines and neonates we detected several instances of relapsed GBS infections caused by the same clone. The reasons underlying the ability of GBS to persist in both hosts will need to be further investigated to potentially adjust the treatment strategy used in these particular cases. Through collaboration with Portuguese laboratories, we have also provided veterinary clinicians with a PCR method to specifically identify the dominant CC61 clone, which will be of critical importance to control and prevent the spread of this successful lineage.

Alternatively, it will also be useful to investigate additional phenotypes that may be important for adaptation to the bovine host, and strive to establish genotype-to-phenotype associations building on the extensive genomic analyses here presented.

As for the host evolutionary history of GBS, although there is strong genomic evidence supporting our model, additional experimentation might be able to functionally address this intriguing question. One possibility would be to assess in competition experiments the fitness of strains lacking or not the regions decayed in the bovine clones. This could be performed within various genomic backgrounds and tested both in bovine milk and in a mouse model of colonization. In the end, this would reveal whether gene inactivation really does represent an evolutionary trade-off, conferring a fitness benefit in one environment in detriment to the loss of fitness in the other. But, even though the functional impact of the mutations can be tested, assessing their selective advantage in the right environmental conditions is much more challenging, as it is in most cases unclear at which stage of infection or colonization these genetic traits were selected.

Whole-genome sequencing has proven to be an indispensable tool for the study of bacterial evolution and adaptation. However, the success of these analyses depends on the existence of a large number of good quality sequences, as well as on the availability of reliable clinical data. Indeed, several studies have leveraged the volume of bacterial sequence data available to make inferences on the dissemination and evolutionary dynamics of many human pathogens. For instance, population genomics was used to investigate the intercontinental spread and evolution of the USA300 community-acquired methicillin-resistant *S. aureus* clone (Glaser et al 2016), as well as the dissemination of the multidrug resistant ST131 lineage of *E. coli* (Petty et al 2014). In this work, we present another valuable application of genomics in looking at GBS host tropism through complementary evolutionary perspectives. As more sequences become available, the predictive strength of these evolutionary inferences will increase and be able to mitigate potential sampling biases that might be introduced with limited datasets. The methodology here used can be applied in other contexts for the study of various bacterial pathogens, while the results we presented bring useful insights that hold promise to the control and treatment of GBS infections in both humans and bovines.

References

Abachin E, Poyart C, Pellegrini E, Milohanic E, Fiedler F, Berche P, Trieu-Cuot P (2002). Formation of D-alanyl-lipoteichoic acid is required for adhesion and virulence of *Listeria monocytogenes*. *Mol Microbiol* **43:** 1-14.

Addis MF, Tanca A, Uzzau S, Oikonomou G, Bicalho RC, Moroni P (2016). The bovine milk microbiota: insights and perspectives from -omics studies. *Mol Biosyst* **12:** 2359-2372.

Akaike H (1974). New look at statistical model identification. *IEEE T Automat Contr* **AC19:** 716-723.

Alamro M, Bidmos FA, Chan H, Oldfield NJ, Newton E, Bai X, Aidley J, Care R, Mattick C *et al* (2014). Phase variation mediates reductions in expression of surface proteins during persistent meningococcal carriage. *Infect Immun* **82:** 2472-2484.

Allen U, Nimrod C, Macdonald N, Toye B, Stephens D, Marchessault V (1999). Relationship between antenatal group B streptococcal vaginal colonization and premature labour. *Paediatr Child Healt* **4:** 465-469.

Almeida A, Albuquerque P, Araujo R, Ribeiro N, Tavares F (2013). Detection and discrimination of common bovine mastitis-causing streptococci. *Vet Microbiol* **164:** 370-377.

Andersen SB, Marvig RL, Molin S, Johansen HK, Griffin AS (2015). Long-term social dynamics drive loss of function in pathogenic bacteria. *Proc Natl Acad Sci U S A* **112:** 10756-10761.

Bailey SF, Hinz A, Kassen R (2014). Adaptive synonymous mutations in an experimentally evolved *Pseudomonas fluorescens* population. *Nat Commun* **5**.

Baker CJ, Barrett FF, Gordon RC, Yow MD (1973). Suppurative meningitis due to streptococci of lancefield group B - study of 33 infants. *J Pediatr* **82:** 724-729.

Baker CJ, Rench MA, Edwards MS, Carpenter RJ, Hays BM, Kasper DL (1988). Immunization of pregnant women with a polysaccharide vaccine of group B *Streptococcus*. *N Engl J Med* **319:** 1180-1185.

Baron MJ, Filman DJ, Prophete GA, Hogle JM, Madoff LC (2007). Identification of a glycosaminoglycan binding region of the alpha C protein that mediates entry of group B streptococci into host cells. *J Biol Chem* **282:** 10526-10536.

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461:** 1243-U1274.

Barton LL, Feigin RD, Lins R (1973). Group B beta hemolytic streptococcal meningitis in infants. *J Pediatr* **82:** 719-723.

Bayliss CD (2009). Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals. *FEMS Microbiol Rev* **33:** 504-520.

Bayliss CD, Bidmos FA, Anjum A, Manchev VT, Richards RL, Grossier JP, Wooldridge KG, Ketley JM, Barrow PA *et al* (2012). Phase variable genes of *Campylobacter jejuni* exhibit high mutation rates and specific mutational patterns but mutability is not the major determinant of population structure during host colonization. *Nucleic Acids Res* **40:** 5876-5889.

Beckmann C, Waggoner JD, Harris TO, Tamura GS, Rubens CE (2002). Identification of novel adhesins from group B streptococci by use of phage display reveals that C5a peptidase mediates fibronectin binding. *Infect Immun* **70:** 2869-2876.

Bellais S, Six A, Fouet A, Longo M, Dmytruk N, Glaser P, Trieu-Cuot P, Poyart C (2012). Capsular switching in group B *Streptococcus* CC17 hypervirulent clone: a future challenge for polysaccharide vaccine development. *J Infect Dis* **206:** 1745-1752.

Bennett PM (2004). Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methos Mol Biol* **266:** 71-113.

Bhaya D, Davison M, Barrangou R (2011). CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45:** 273-297.

Bidet P, Brahimi N, Chalas C, Aujard Y, Bingen E (2003). Molecular characterization of serotype III group B *Streptococcus* isolates causing neonatal meningitis. *J Infect Dis* **188:** 1132-1137.

Bisharat N, Crook DW, Leigh J, Harding RM, Ward PN, Coffey TJ, Maiden MC, Peto T, Jones N (2004). Hyperinvasive neonatal group B *Streptococcus* has arisen from a bovine ancestor. *J Clin Microbiol* **42:** 2161-2167.

Bisharat N, Jones N, Marchaim D, Block C, Harding RM, Yagupsky P, Peto T, Crook DW (2005). Population structure of group B streptococcus from a low-incidence region for invasive neonatal disease. *Microbiol-Sgm* **151:** 1875-1881.

Bohnsack JF, Takahashi S, Hammitt L, Miller DV, Aly AA, Adderson EE (2000). Genetic polymorphisms of group B streptococcus *scpB* alter functional activity of a cell-associated peptidase that inactivates C5a. *Infect Immun* **68:** 5018-5025.

Bolduc GR, Madoff LC (2007). The group B streptococcal alpha C protein binds alpha(1)beta(1)-integrin through a novel KTD motif that promotes internalization of GBS within human epithelial cells. *Microbiol-Sgm* **153:** 4039-4049.

Boyd DA, Cvitkovitch DG, Bleiweis AS, Kiriukhin MY, Debabov DV, Neuhaus FC, Hamilton IR (2000). Defects in D-alanyl-lipoteichoic acid synthesis in *Streptococcus mutans* results in acid sensitivity. *J Bacteriol* **182:** 6055-6065.

Brimil N, Barthell E, Heindrichs U, Kuhn M, Lutticken R, Spellerberg B (2006). Epidemiology of *Streptococcus agalactiae* colonization in Germany. *Zentralbl Bakteriol* **296:** 39-44.

Brochet M, Couve E, Zouine M, Vallaeys T, Rusniok C, Lamy MC, Buchrieser C, Trieu-Cuot P, Kunst F *et al* (2006). Genomic diversity and evolution within the species *Streptococcus agalactiae*. *Microbes Infect* **8:** 1227-1243.

Brochet M (2008). Génomique des populations et flux géniques au sein de l'espèce *Streptococcus agalactiae*. PhD thesis, Université Paris VI.

Brochet M, Couve E, Glaser P, Guedon G, Payot S (2008a). Integrative conjugative elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae*. *J Bacteriol* **190:** 6913-6917.

Brochet M, Rusniok C, Couve E, Dramsi S, Poyart C, Trieu-Cuot P, Kunst F, Glaser P (2008b). Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **105:** 15961-15966.

Brochet M, Da Cunha V, Couve E, Rusniok C, Trieu-Cuot P, Glaser P (2009). Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. *Mol Microbiol* **71:** 948-959.

Broker G, Spellerberg B (2004). Surface proteins of *Streptococcus agalactiae* and horizontal gene transfer. *Zentralbl Bakteriol* **294:** 169-175.

Bromham L, Penny D (2003). The modern molecular clock. *Nat Rev Genet* **4:** 216-224.

Burrus V, Waldor MK (2004). Shaping bacterial genomes with integrative and conjugative elements. *Res Microbiol* **155:** 376-386.

Buscetta M, Papasergi S, Firon A, Pietrocola G, Biondo C, Mancuso G, Midiri A, Romeo L, Teti G *et al* (2014). FbsC, a novel fibrinogen-binding protein, promotes *Streptococcus agalactiae*-host cell interactions. *J Biol Chem* **289:** 21003-21015.

Cai Y, Kong F, Gilbert GL (2007). Three new macrolide efflux (*mef*) gene variants in *Streptococcus agalactiae*. *J Clin Microbiol* **45:** 2754-2755.

Caliot E, Dramsi S, Chapot-Chartier M-P, Courtin P, Kulakauskas S, Pechoux C, Trieu-Cuot P, Mistou M-Y (2012). Role of the group B antigen of *Streptococcus agalactiae*: a peptidoglycan-anchored polysaccharide involved in cell wall biogenesis. *PLOS Pathog* **8**.

Carlin AF, Uchiyama S, Chang YC, Lewis AL, Nizet V, Varki A (2009). Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. *Blood* **113:** 3333-3336.

Charlesworth B (2009). Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10:** 195-205.

Chattopadhyay D, Carey AJ, Caliot E, Webb RI, Layton JR, Wang Y, Bohnsack JF, Adderson EE, Ulett GC (2011). Phylogenetic lineage and pilus protein Spb1/SAN1518 affect opsonin-independent phagocytosis and intracellular survival of Group B Streptococcus. *Microbes Infect* **13:** 369-382.

Cheng Q, Stafslien D, Purushothaman SS, Cleary P (2002). The group B streptococcal C5a peptidase is both a specific protease and an invasin. *Infect Immun* **70:** 2408-2413.

Chmouryguina I, Suvorov A, Ferrieri P, Cleary PP (1996). Conservation of the C5a peptidase genes in group A and B streptococci. *Infect Immun* **64:** 2387-2390.

Christie K, Atkins NE, Munch-Petersen E (1944). A note on a lytic phenomenon shown by group B streptococci. *Aust J Exp Biol Med* **22:** 197-200.

Cieslewicz MJ, Chaffin D, Glusman G, Kasper D, Madan A, Rodrigues S, Fahey J, Wessels MR, Rubens CE (2005). Structural and genetic diversity of group B *Streptococcus* capsular polysaccharides. *Infect Immun* **73:** 3096-3103.

Cingolani P, Platts A, Wang LL, Coon M, Tung N, Wang L, Land SJ, Lu X, Ruden DM (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. *Fly* **6:** 80-92.

Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S (2011). Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* **44:** 106-110.

Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T *et al* (2014). Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *mBio* **5**.

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH *et al* (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science* **331:** 430-434.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR (2015). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**.

Da Cunha V, Davies MR, Douarre PE, Rosinski-Chupin I, Margarit I, Spinali S, Perkins T, Lechat P, Dmytruk N *et al* (2014). *Streptococcus agalactiae* clones infecting humans were selected and fixed through the extensive use of tetracycline. *Nat Commun* **5:** 11.

Daley AJ, Garland SM (2004). Prevention of neonatal group B streptococcal disease: progress, challenges and dilemmas. *J Paediatr Child H* **40:** 664-668.

David S, Rusniok C, Mentasti M, Gomez-Valero L, Harris SR, Lechat P, Lees J, Ginevra C, Glaser P *et al* (2016). Multiple major disease-associated clones of *Legionella pneumophila* have emerged recently and independently. *Genome Res*.

Davidson AL, Dassa E, Orelle C, Chen J (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol Mol Biol Rev* **72:** 317-364.

de Groot A, Roche D, Fernandez B, Ludanyi M, Cruveiller S, Pignol D, Vallenet D, Armengaud J, Blanchard L (2014). RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium *Deinococcus deserti*. *Genome Biol and Evol* **6:** 932-948.

Delannoy CM, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, Turnbull JF, Zadoks RN (2013). Human *Streptococcus agalactiae* strains in aquatic mammals and fish. *BMC Microbiol* **13:** 1-9.

Deveau H, Garneau JE, Moineau S (2010). CRISPR/Cas System and its role in phage-bacteria interactions. *Annu Rev Microbiol* **64:** 475-493.

Di Palo B, Rippa V, Santi I, Brettoni C, Muzzi A, Metruccio MME, Grifantini R, Telford JL, Paccani SR *et al* (2013). Adaptive response of group B *Streptococcus* to high glucose conditions: new insights on the covrs regulation network. *PLOS One* **8:** 11.

Didelot X, Eyre DW, Cule M, Ip CLC, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T *et al* (2012a). Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* **13**.

Didelot X, Meric G, Falush D, Darling AE (2012b). Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**.

Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ (2016). Within-host evolution of bacterial pathogens. *Nat Rev Microbiol* **14:** 150-162.

Dogan B, Schukken YH, Santisteban C, Boor KJ (2005). Distribution of serotypes and antimicrobial resistance genes among *Streptococcus agalactiae* isolates from bovine and human hosts. *J Clin Microbiol* **43:** 5899-5906.

Dong Y, Speer CP (2015). Late-onset neonatal sepsis: recent developments. *Arch Dis Child Fetal Neonatal Ed* **100:** F257-F263.

Doran KS, Liu GY, Nizet V (2003). Group B streptococcal beta-hemolysin/cytolysin activates neutrophil signaling pathways in brain endothelium and contributes to development of meningitis. *J Clin Invest* **112:** 736-744.

Doran KS, Nizet V (2004). Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy. *Mol Microbiol* **54:** 23-31.

Douarre PE, Sauvage E, Poyart C, Glaser P (2015). Host specificity in the diversity and transfer of *lsa* resistance genes in group B *Streptococcus*. *J Antimicrob Chemother* **70:** 3205-3213.

Dramsi S, Caliot E, Bonne I, Guadagnini S, Prevost MC, Kojadinovic M, Lalioui L, Poyart C, Trieu-Cuot P (2006). Assembly and role of pili in group B streptococci. *Mol Microbiol* **60:** 1401-1413.

Edmond KM, Kortsalioudaki C, Scott S, Schrag SJ, Zaidi AK, Cousens S, Heath PT (2012). Group B streptococcal disease in infants aged younger than 3 months: systematic review and meta-analysis. *Lancet* **379:** 547-556.

Edwards MS, Baker CJ (2005). Group B streptococcal infections. In: R. J.S aKJO (ed). *Infectious Diseases of the Fetus and Newborn Infant*. Saunders. pp 1091-1156.

Eickhoff TC, Klein JO, Daly AK, Ingall D, Finland M (1964). Neonatal sepsis and other infections due to group B beta-hemolytic streptococci. *N Engl J Med* **271:** 1221-1228.

Eldar A, Bejerano Y, Bercovier H (1994). *Streptococcus shiloi* and *Streptococcus difficile* - 2 new streptococcal species causing a meningoencephalitis in fish. *Curr Microbiol* **28:** 139-143.

Elliott JA, Thompson TA, Facklam RR, Slotved HC (2004). Increased sensitivity of a latex agglutination method for serotyping group B streptococcus. *J Clin Microbiol* **42:** 3907-3907.

Evans JJ, Klesius PH, Gilbert PM, Shoemaker CA, Al Sarawi MA, Landsberg J, Duremdez R, Al Marzouk A, Al Zenki S (2002). Characterization of β-haemolytic group B *Streptococcus agalactiae* in cultured seabream, *Sparus auratus* L., and wild mullet, *Liza klunzingeri* (Day), in Kuwait. *J Fish Dis* **25:** 505-513.

Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC, Shoemaker CA, Takahashi S (2008). Phylogenetic relationships among *Streptococcus agalactiae* isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan. *J Med Microbiol* **57:** 1369-1376.

Eyre DW, Babakhani F, Griffiths D, Seddon J, Elias CD, Gorbach SL, Peto TEA, Crook D, Walker AS (2014). Whole-genome sequencing demonstrates that fidaxomicin is superior to vancomycin for preventing reinfection and relapse of infection with *Clostridium difficile*. *J Infect Dis* **209:** 1446-1451.

Fabretti F, Theilacker C, Baldassarri L, Kaczynski Z, Kropec A, Holst O, Huebner J (2006). Alanine esters of enterococcal lipoteichoic acid play a role in biofilm formation and resistance to antimicrobial peptides. *Infect Immun* **74:** 4164-4171.

Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, Suerbaum S (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: Estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A* **98:** 15056-15061.

Faralla C, Metruccio MM, De Chiara M, Mu R, Patras KA, Muzzi A, Grandi G, Margarit I, Doran KS *et al* (2014). Analysis of two-component systems in group B Streptococcus shows that RgfAC and the novel FspSR modulate virulence and bacterial fitness. *mBio* **5**.

Fasola E, Livdahl C, Ferrieri P (1993). Molecular analysis of multiple isolates of the major serotypes of group B streptococci. *J Clin Microbiol* **31:** 2616-2620.

Ferrieri P (1988). Surface-localized protein antigens of group B-streptococci. *Rev Infect Dis* **10:** S363-S366.

Firon A, Tazi A, Da Cunha V, Brinster S, Sauvage E, Dramsi S, Golenbock DT, Glaser P, Poyart C *et al* (2013). The Abi-domain protein Abx1 interacts with the CovS histidine kinase to control virulence gene expression in group B *Streptococcus*. *PLOS Pathog* **9**.

Fischer A, Liljander A, Kaspar H, Muriuki C, Fuxelius HH, Bongcam-Rudloff E, de Villiers EP, Huber CA, Frey J *et al* (2013). Camel *Streptococcus agalactiae* populations are associated with specific disease complexes and acquired the tetracycline resistance gene *tetM* via a Tn916-like element. *Vet Res* **44**.

Flores AR, Galloway-Pena J, Sahasrabhojane P, Saldana M, Yao H, Su X, Ajami NJ, Holder ME, Petrosino JF *et al* (2015). Sequence type 1 group B *Streptococcus*, an emerging cause of invasive disease in adults, evolves by small genetic changes. *Proc Natl Acad Sci U S A* **112:** 6431-6436.

Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC *et al* (2011). Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* **43:** 482-+.

Foxman B, Gillespie BW, Manning SD, Marrs CF (2007). Risk factors for group B streptococcal colonization: potential for different transmission systems by capsular type. *Ann Epidemiol* **17:** 854-862.

Franciosi RA, Knostman JD, Zimmerma RA (1973). Group B streptococcal neonatal and infant infections. *J Pediatr* **82:** 707-718.

Franke AE, Clewell DB (1981). Evidence for a chromosome-borne resistance transposon (Tn916) in *Streptococcus faecalis* that is capable of conjugal transfer in the absence of a conjugative plasmid. *J Bacteriol* **145:** 494-502.

Franken C, Haase G, Brandt C, Weber-Heynemann J, Martin S, Lammler C, Podbielski A, Lutticken R, Spellerberg B (2001). Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing *scpB* and *lmb*. *Mol Microbiol* **41:** 925-935.

Frost LS, Leplae R, Summers AO, Toussaint A (2005). Mobile genetic elements: The agents of open source evolution. *Nat Rev Microbiol* **3:** 722-732.

Furuya EY, Lowy FD (2006). Antimicrobial-resistant bacteria in the community setting. *Nat Rev Microbiol* **4:** 36-45.

Gao W, Chua K, Davies JK, Newton HJ, Seemann T, Harrison PF, Holmes NE, Rhee H-W, Hong J-I *et al* (2010). Two novel point mutations in clinical *Staphylococcus aureus* reduce linezolid susceptibility and switch on the stringent response to promote persistent infection. *PLOS Pathog* **6:** e1000944.

Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, Zouine M, Couve E, Lalioui L *et al* (2002). Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. *Mol Microbiol* **45:** 1499-1513.

Glaser P, Martins-Simões P, Villain A, Barbier M, Tristan A, Bouchier C, Ma L, Bes M, Laurent F *et al* (2016). Demography and intercontinental spread of the USA300 community-acquired methicillin-resistant *Staphylococcus aureus* lineage. *mBio* **7**.

Gleich-Theurer U, Aymanns S, Haas G, Mauerer S, Vogt J, Spellerberg B (2009). Human serum induces streptococcal C5a peptidase expression. *Infect Immun* **77:** 3817-3825.

Goldstone RJ, Harris S, Smith DGE (2016). Genomic content typifying a prevalent clade of bovine mastitis-associated *Escherichia coli*. *Sci Rep* **6**.

Gottschalk B, Broeker G, Kuhn M, Aymanns S, Gleich-Theurer U, Spellerberg B (2006). Transport of multidrug resistance substrates by the *Streptococcus agalactiae* hemolysin transporter. *J Bacteriol* **188:** 5984-5992.

Gravekamp C, Kasper DL, Michel JL, Kling DE, Carey V, Madoff LC (1997). Immunogenicity and protective efficacy of the alpha C protein of group B streptococci are inversely related to the number of repeats. *Infect Immun* **65:** 5216-5221.

Gravey F, Galopin S, Grall N, Auzou M, Andremont A, Leclercq R, Cattoir V (2013). Lincosamide resistance mediated by *lnu*(C) (L phenotype) in a *Streptococcus anginosus* clinical isolate. *J Antimicrob Chemother* **68:** 2464-2467.

Guerillot R, Da Cunha V, Sauvage E, Bouchier C, Glaser P (2013). Modular evolution of TnGBSs, a new family of integrative and conjugative elements associating insertion sequence transposition, plasmid replication, and conjugation for their spreading. *J Bacteriol* **195:** 1979-1990.

Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder BV, Cartwright RA, Smyth DS, Smyth CJ, Lindsay JA *et al* (2010). Evolutionary genomics of staphylococcus

aureus reveals insights into the origin and molecular basis of ruminant host adaptation. *Genome Biol and Evol* **2:** 454-466.

Gutekunst H, Eikmanns BJ, Reinscheid DJ (2004). The novel fibrinogen-binding protein FbsB promotes *Streptococcus agalactiae* invasion into epithelial cells. *Infect Immun* **72:** 3495-3504.

Haenni M, Saras E, Bertin S, Leblond P, Madec JY, Payot S (2010). Diversity and mobility of integrative and conjugative elements in bovine isolates of *Streptococcus agalactiae*, *S. dysgalactiae subsp. dysgalactiae*, and *S. uberis*. *Appl Environ Microbiol* **76:** 7957-7965.

Harmon RJ (1994). Physiology of mastitis and factors affecting somatic cell counts. *J Dairy Sci* **77:** 2103-2112.

Hastings MJ, Easmon CSF, Bloxham B, Clare AJ (1981). Nosocomial transmission of group B streptococci. *J Med Microbiol* **14:** R9-R9.

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57:** 97-109.

Heath A, DiRita VJ, Barg NL, Engleberg NC (1999). A two-component regulatory system, CsrR-CsrS, represses expression of three *Streptococcus pyogenes* virulence factors, hyaluronic acid capsule, streptolysin S, and pyrogenic exotoxin B. *Infect Immun* **67:** 5298-5305.

Hedge J, Wilson DJ (2016). Practical approaches for detecting selection in microbial genomes. *PLOS Comput Biol* **12**.

Heelan JS, Hasenbein ME, McAdam AJ (2004). Resistance of group B *Streptococcus* to selected antibiotics, including erythromycin and clindamycin. *J Clin Microbiol* **42:** 1263-1264.

Henderson IR, Owen P, Nataro JP (1999). Molecular switches - the ON and OFF of bacterial phase variation. *Mol Microbiol* **33:** 919-932.

Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond AJ (2007). Evidence for time dependency of molecular rate estimates. *Syst Biol* **56:** 515-522.

Hogeveen H, Huijps K, Lam T (2011). Economic aspects of mastitis: new developments. *N Z Vet J* **59:** 16-23.

Holden MTG, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, Jourdan T, Bason NC, Holroyd NE *et al* (2009). Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLOS Pathog* **5**.

Horvath P, Barrangou R (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327:** 167-170.

Hughes MJG, Moore JC, Lane JD, Wilson R, Pribul PK, Younes ZN, Dobson RJ, Everest P, Reason AJ *et al* (2002). Identification of major outer surface proteins of *Streptococcus agalactiae*. *Infect Immun* **70:** 1254-1259.

Imperi M, Pataracchia M, Alfarone G, Baldassarri L, Orefici G, Creti R (2010). A multiplex PCR assay for the direct identification of the capsular type (Ia to IX) of *Streptococcus agalactiae*. *J Microbiol Meth* **80:** 212-214.

Ippolito DL, James WA, Tinnemore D, Huang RR, Dehart MJ, Williams J, Wingerd MA, Demons ST (2010). Group B *Streptococcus* serotype prevalence in reproductive-age women at a tertiary care military medical center relative to global serotype distribution. *BMC Infect Dis* **10**.

Jacob F, Monod J (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3:** 318-356.

Jansen R, van Embden JDA, Gaastra W, Schouls LM (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43:** 1565-1575.

Janulczyk R, Ricci S, Bjorck L (2003). MtsABC is important for manganese and iron transport, oxidative stress resistance, and virulence of *Streptococcus pyogenes*. *Infect Immun* **71:** 2656-2664.

Jiang SM, Cieslewicz MJ, Kasper DL, Wessels MR (2005). Regulation of virulence by a two-component system in group B *Streptococcus*. *J Bacteriol* **187:** 1105-1113.

Jiang SM, Ishmael N, Hotopp JD, Puliti M, Tissi L, Kumar N, Cieslewicz MJ, Tettelin H, Wesselsl MR (2008). Variation in the group B *Streptococcus* CsrRS regulon and effects on pathogenicity. *J Bacteriol* **190:** 1956-1965.

Johnson CM, Grossman AD (2015). Integrative and Conjugative Elements (ICEs): what they do and how they work. *Annu Rev Genet* **49:** 577-601.

Johri AK, Paoletti LC, Glaser P, Dua M, Sharma PK, Grandi G, Rappuoli R (2006). Group B *Streptococcus*: global incidence and vaccine development. *Nat Rev Microbiol* **4:** 932-942.

Jones AL, Knoll KM, Rubens CE (2000). Identification of *Streptococcus agalactiae* virulence genes in the neonatal rat sepsis model using signature-tagged mutagenesis. *Mol Microbiol* **37:** 1444-1455.

Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, Glaser P, Rusniok C, Crook DWM *et al* (2003). Multilocus sequence typing system for group B *Streptococcus*. *J Clin Microbiol* **41:** 2530-2536.

Jorgensen HJ, Nordstoga AB, Sviland S, Zadoks RN, Solverod L, Kvitle B, Mork T (2016). *Streptococcus agalactiae* in the environment of bovine dairy herds - rewriting the textbooks? *Vet Microbiol* **184:** 64-72.

Jukes TH, Cantor CR (1969). Evolution of protein molecules. *Mammalian protein metabolism* **3:** 132.

Jürgens D, Sterzik B, Fehrenbach FJ (1987). Unspecific binding of group B streptococcal cocytolysin (CAMP factor) to immunoglobulins and its possible role in pathogenicity. *J Exp Med* **165:** 720.

Kawamura Y, Fujiwara H, Mishima N, Tanaka Y, Tanimoto A, Ilawa S, Itoh Y, Ezaki T (2003). First *Streptococcus agalactiae* isolates highly resistant to quinolones, with point mutations in *gyrA* and *parC*. *Antimicrob Agents Chemother* **47:** 3605-3609.

Kexel G, Schoenbohm S (1965). *Streptococcus agalactiae* as the causative agent in infantile meningitis. *Dtsch Med Wochenschr* **90:** 258-261.

Kim JH (1996). General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Syst Biol* **45:** 363-374.

Kling DE, Gravekamp C, Madoff LC, Michel JL (1997). Characterization of two distinct opsonic and protective epitopes within the alpha C protein of the group B *Streptococcus*. *Infect Immun* **65:** 1462-1467.

Kobayashi M, Vekemans J, Baker CJ, Ratner AJ, Le Doare K, Schrag SJ (2016). Group B *Streptococcus* vaccine development: present status and future considerations, with emphasis on perspectives for low and middle income countries. *F1000Research* **5:** 2355-2355.

Koehler W (2007). The present state of species within the genera *Streptococcus* and *Enterococcus*. *Zentralbl Bakteriol* **297:** 133-150.

Koeser CU, Ellington MJ, Peacock SJ (2014). Whole-genome sequencing to control antimicrobial resistance. *Trends Genet* **30:** 401-407.

Kolaczkowski B, Thornton JW (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431:** 980-984.

Kong F, Lambertsen LM, Slotved HC, Ko D, Wang H, Gilbert GL (2008). Use of phenotypic and molecular serotype identification methods to characterize previously nonserotypeable group B streptococci. *J Clin Microbiol* **46:** 2745-2750.

Kong FR, Gowan S, Martin D, James G, Gilbert GL (2002). Molecular profiles of group B streptococcal surface protein antigen genes: relationship to molecular serotypes. *J Clin Microbiol* **40:** 620-626.

Kuang Y, Tani K, Synnott AJ, Ohshima K, Higuchi H, Nagahata H, Tanji Y (2009). Characterization of bacterial population of raw milk from bovine mastitis by culture-independent PCR-DGGE method. *Biochem Eng J* **45:** 76-81.

Kupczok A, Bollback JP (2013). Probabilistic models for CRISPR spacer content evolution. *BMC Evol Biol* **13**.

Lachenauer CS, Creti R, Michel JL, Madoff LC (2000). Mosaicism in the alpha-like protein genes of group B streptococci. *Proc Natl Acad Sci U S A* **97:** 9630-9635.

Lammler C, Abdulmawjood A, Weiss R (1998). Properties of serological group B streptococci of dog, cat and monkey origin. *J Vet Med B* **45:** 561-566.

Lamy MC, Zouine M, Fert J, Vergassola M, Couve E, Pellegrini E, Glaser P, Kunst F, Msadek T *et al* (2004). CovS/CovR of group B *Streptococcus*: a two-component global regulatory system involved in virulence. *Mol Microbiol* **54:** 1250-1268.

Lancefield RC (1933). A serological differentiation of human and other groups of hemolytic streptococci. *J Exp Med* **57:** 571-595.

Lancefield RC (1934). A serological differentiation of specific types of bovine hemolytic streptococci (group B). *J Exp Med* **59:** 441-458.

Lancefield RC, Hare R (1935). The serological differentiation of pathogenic and non-pathogenic strains of hemolytic streptococci from parturient women. *J Exp Med* **61:** 335-349.

Lauer P, Rinaudo CD, Soriani M, Margarit I, Maione D, Rosini R, Taddei AR, Mora M, Rappuoli R *et al* (2005). Genome analysis reveals pili in group B *Streptococcus*. *Science* **309:** 105-105.

Lemey P, Salemi M, Vandamme AM (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Levin BR, Perrot V, Walker N (2000). Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* **154:** 985-997.

Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Jr., Skurnik D, Leiby N, LiPuma JJ *et al* (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* **43:** 1275-U1148.

Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R (2014). Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nat Genet* **46:** 82-+.

Ligozzi M, Bernini C, Bonora MG, de Fatima M, Zuliani J, Fontana R (2002). Evaluation of the VITEK 2 system for identification and antimicrobial susceptibility testing of medically relevant gram-positive cocci. *J Clin Microbiol* **40:** 1681-1686.

Lin W-J, Walthers D, Connelly JE, Burnside K, Jewell KA, Kenney LJ, Rajagopal L (2009). Threonine phosphorylation prevents promoter DNA binding of the group B *Streptococcus* response regulator CovR. *Mol Microbiol* **71:** 1477-1495.

Lindahl G, Stalhammar-Carlemalm M, Areschoug T (2005). Surface proteins of *Streptococcus agalactiae* and related proteins in other bacterial pathogens. *Clin Microbiol Rev* **18:** 102-127.

Lopez-Sanchez MJ, Sauvage E, Da Cunha V, Clermont D, Hariniaina ER, Gonzalez-Zorn B, Poyart C, Rosinski-Chupin I, Glaser P (2012). The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Mol Microbiol* **85:** 1057-1071.

Luan SL, Granlund M, Sellin M, Lagergard T, Spratt BG, Norgren M (2005). Multilocus sequence typing of Swedish invasive group B streptococcus isolates indicates a neonatally associated genetic lineage and capsule switching. *J Clin Microbiol* **43:** 3727-3733.

Madoff LC, Michel JL, Gong EW, Kling DE, Kasper DL (1996). Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein. *Proc Natl Acad Sci U S A* **93:** 4131-4136.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou JJ, Zurth K *et al* (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95:** 3140-3145.

Maisey HC, Hensler M, Nizet V, Doran KS (2007). Group B streptococcal pilus proteins contribute to adherence to and invasion of brain microvascular endothelial cells. *J Bacteriol* **189:** 1464-1467.

Maisey HC, Quach D, Hensler ME, Liu GY, Gallo RL, Nizet V, Doran KS (2008). A group B streptococcal pilus protein promotes phagocyte resistance and systemic virulence. *Faseb Journal* **22:** 1715-1724.

Malbruny B, Werno AM, Murdoch DR, Leclercq R, Cattoir V (2011). Cross-resistance to lincosamides, streptogramins A, and pleuromutilins due to the *lsa*(c) gene in *Streptococcus agalactiae* UCN70. *Antimicrob Agents Chemother* **55:** 1470-1474.

Manning SD, Springman AC, Lehotzky E, Lewis MA, Whittam TS, Davies HD (2009). Multilocus sequence types associated with neonatal group B streptococcal sepsis and meningitis in Canada. *J Clin Microbiol* **47:** 1143-1148.

Manning SD, Springman AC, Million AD, Milton NR, McNamara SE, Somsel PA, Bartlett P, Davies HD (2010). Association of group B *Streptococcus* colonization and bovine exposure: a prospective cross-sectional cohort study. *PLOS One* **5:** 6.

Margarit I, Rinaudo CD, Galeotti CL, Maione D, Ghezzo C, Buttazzoni E, Rosini R, Runci Y, Mora M *et al* (2009). Preventing bacterial infections with pilus-based vaccines: the group B *Streptococcus* paradigm. *J Infect Dis* **199:** 108-115.

Marques MB, Kasper DL, Pangburn MK, Wessels MR (1992). Prevention of C3 deposition by capsular polysaccharide is a virulence mechanism of type II group B streptococci. *Infect Immun* **60:** 3986-3993.

Martins ER, Pessanha MA, Ramirez M, Melo-Cristino J, Portuguese Grp Study S (2007). Analysis of group B streptococcal isolates from infants and pregnant women in Portugal revealing two lineages with enhanced invasiveness. *J Clin Microbiol* **45:** 3224-3229.

Marvig RL, Johansen HK, 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**.

Marvig RL, Sommer LM, Molin S, Johansen HK (2014). Convergent evolution and adaptation of *Pseudomonas aeruginosa* within patients with cystic fibrosis. *Nat Genet* **47:** 57–64.

Mathers AJ, Stoesser N, Sheppard AE, Pankhurst L, Giess A, Yeh AJ, Didelot X, Turner SD, Sebra R *et al* (2015). *Klebsiella pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae* at a single institution: insights into endemicity from whole-genome sequencing. *Antimicrob Agents Chemother* **59:** 1661-1668.

McDonald JS (1979). Bovine mastitis - introductory remarks. *J Dairy Sci* **62:** 117-118.

McNab R, Forbes H, Handley PS, Loach DM, Tannock GW, Jenkinson HF (1999). Cell wall-anchored CshA polypeptide (259 kilodaltons) in *Streptococcus gordonii* forms surface fibrils that confer hydrophobic and adhesive properties. *J Bacteriol* **181:** 3087-3095.

Melnyk AH, Wong A, Kassen R (2015). The fitness costs of antibiotic resistance mutations. *Evol Appl* **8:** 273-283.

Merritt K, Jacobs NJ (1976). Improved medium for detecting pigment production by group B streptococci. *J Clin Microbiol* **4:** 379-380.

Michon F, Katzenellenbogen E, Kasper DL, Jennings HJ (1987). Structure of the complex group-specific polysaccharide of group B *Streptococcus*. *Biochemistry* **26:** 476-486.

Mistou M-Y, Dramsi S, Brega S, Poyart C, Trieu-Cuot P (2009). Molecular dissection of the *secA2* locus of group B *Streptococcus* reveals that glycosylation of the Srr1 LPXTG protein is required for full virulence. *J Bacteriol* **191:** 4195-4206.

Mitchell TJ (2003). The pathogenesis of streptococcal infections: from tooth decay to meningitis. *Nat Rev Microbiol* **1:** 219-230.

Mora M, Bensi G, Capo S, Falugi F, Zingaretti C, Manetti AGO, Maggi T, Taddei AR, Grandi G *et al* (2005). Group A *Streptococcus* produce pilus-like structures containing protective antigens and Lancefield T antigens. *Proc Natl Acad Sci U S A* **102:** 15641-15646.

Moxon R, Bayliss C, Hood D (2006). Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics*. pp 307-333.

Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E *et al* (2007). Tracking the *in vivo* evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci U S A* **104:** 9451-9456.

Mweu MM, Nielsen SS, Halasa T, Toft N (2012). Annual incidence, prevalence and transmission characteristics of *Streptococcus agalactiae* in Danish dairy herds. *Prev Vet Med* **106:** 244-250.

Neave FK, Dodd FH, Kingwill RG, Westgarth DR (1969). Control of mastitis in the dairy herd by hygiene and management. *J Dairy Sci* **52:** 696-707.

Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3:** 418-426.

Nocard M, Mollereau H (1887). Sur unemammite contagieuse des vaches laitieres. *Ann Inst Pasteur* **1:** 109-126.

Ochman H, Wilson AC (1987). Evolution in bacteria - evidence for a universal substitution rate in cellular genomes. *J Mol Evol* **26:** 74-86.

Okoro CK, Kingsley RA, Quail MA, Kankwatira AM, Feasey NA, Parkhill J, Dougan G, Gordon MA (2012). High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal *Salmonella typhimurium* disease. *Clin Infect Dis* **54:** 955-963.

Oliveira ICM, De Mattos MC, Areal MFT, Ferreira-Carvalho BT, Figueiredo AMS, Benchetrit LC (2005). Pulsed-field gel electrophoresis of human group B streptococci isolated in Brazil. *J Chemotherapy* **17:** 258-263.

Ounissi H, Derlot E, Carlier C, Courvalin P (1990). Gene homogeneity for aminoglycoside-modifying enzymes in gram-positive cocci. *Antimicrob Agents Chemother* **34:** 2164-2168.

Papadelli M, Karsioti A, Anastasiou R, Georgalaki M, Tsakalidou E (2007). Characterization of the gene cluster involved in the biosynthesis of macedocin, the lantibiotic produced by *Streptococcus macedonicus*. *FEMS Microbiol Lett* **272:** 75-82.

Park SE, Jiang SM, Wessels MR (2012). CsrRS and environmental pH regulate group B *Streptococcus* adherence to human epithelial cells and extracellular matrix. *Infect Immun* **80:** 3975-3984.

Paulsen IT, Banerjei L, Myers GSA, Nelson KE, Seshadri R, Read TD, Fouts DE, Eisen JA, Gill SR *et al* (2003). Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* **299:** 2071-2074.

Peschel A, Otto M, Jack RW, Kalbacher H, Jung G, Gotz F (1999). Inactivation of the dlt operon in *Staphylococcus aureus* confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *J Biol Chem* **274:** 8405-8410.

Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, Forde BM, Phan M-D, Gomes Moriel D, Peters KM *et al* (2014). Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc Natl Acad Sci U S A* **111:** 5694-5699.

Pezzicoli A, Santi I, Lauer P, Rosini R, Rinaudo D, Grandi G, Telford JL, Soriani M (2008). Pilus backbone contributes to group B *Streptococcus* paracellular translocation through epithelial cells. *J Infect Dis* **198:** 890-898.

Philippe N, Crozat E, Lenski RE, Schneider D (2007). Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. *Bioessays* **29:** 846-860.

Pillai P, Srinivasan U, Zhang L, Borchardt SM, Debusscher J, Marrs CF, Foxman B (2009). *Streptococcus agalactiae* pulsed-field gel electrophoresis patterns cross capsular types. *Epidemiol Infect* **137:** 1420-1425.

Poutrel B, Ryniewicz HZ (1984). Evaluation of the API 20 Strep system for species identification of streptococci isolated from bovine mastitis. *J Clin Microbiol* **19:** 213-214.

Poyart C, Lamy MC, Boumaila C, Fiedler F, Trieu-Cuot P (2001). Regulation of D-alanyl-lipoteichoic acid biosynthesis in Streptococcus agalactiae involves a novel two-component regulatory system. *J Bacteriol* **183:** 6324-6334.

Poyart C, Jardy L, Quesne G, Berche P, Trieu-Cuot P (2003). Genetic basis of antibiotic resistance in *Streptococcus agalactiae* strains isolated in a French hospital. *Antimicrob Agents Chemother* **47:** 794-797.

Poyart C, Tazi A, Reglier-Poupet H, Billoet A, Tavares N, Raymond J, Trieu-Cuot P (2007). Multiplex PCR assay for rapid and accurate capsular typing of group B streptococci. *J Clin Microbiol* **45:** 1985-1988.

Poyart C, Reglier-Poupet H, Tazi A, Billoet A, Dmytruk N, Bidet P, Bingen E, Raymond J, Trieu-Cuot P (2008). Invasive group B streptococcal infections in infants, France. *Emerg Infect Dis* **14:** 1647-1649.

Price EP, Sarovich DS, Mayo M, Tuanyok A, Drees KP, Kaestli M, Beckstrom-Sternberg SM, Babic-Sternberg JS, Kidd TJ *et al* (2013). Within-host evolution of *Burkholderia pseudomallei* over a twelve-year chronic carriage infection. *mBio* **4**.

Pritzlaff CA, Chang JCW, Kuo SP, Tamura GS, Rubens CE, Nizet V (2001). Genetic basis for the beta-haemolytic/cytolytic activity of group B *Streptococcus*. *Mol Microbiol* **39:** 236-247.

Rato MG, Bexiga R, Florindo C, Cavaco LM, Vilela CL, Santos-Sanches I (2013). Antimicrobial resistance and molecular epidemiology of streptococci from bovine mastitis. *Vet Microbiol* **161:** 286-294.

Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR *et al* (2011). Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLOS One* **6**.

Reichmann NT, Cassona CP, Gruendling A (2013). Revised mechanism of D-alanine incorporation into cell wall polymers in Gram-positive bacteria. *Microbiol-Sgm* **159:** 1868-1877.

Richards VP, Lang P, Bitar PDP, Lefebure T, Schukken YH, Zadoks RN, Stanhope MJ (2011). Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. *Infect Genet Evol* **11:** 1263-1275.

Richards VP, Choi SC, Bitar PDP, Gurjar AA, Stanhope MJ (2013). Transcriptomic and genomic evidence for *Streptococcus agalactiae* adaptation to the bovine environment. *BMC Genomics* **14:** 15.

Rinaudo CD, Rosini R, Galeotti CL, Berti F, Necchi F, Reguzzi V, Ghezzo C, Telford JL, Grandi G *et al* (2010). Specific involvement of pilus type 2a in biofilm formation in group B *Streptococcus*. *PLOS One* **5**.

Ring A, Braun JS, Pohl J, Nizet V, Stremmel W, Shenep JL (2002). Group B streptococcal beta-hemolysin induces mortality and liver injury in experimental sepsis. *J Infect Dis* **185:** 1745-1753.

Roberts AP, Mullany P (2009). A modular master on the move: the *Tn916* family of mobile genetic elements. *Trends Microbiol* **17:** 251-258.

Roberts MC (2005). Update on acquired tetracycline resistance genes. *FEMS Microbiol Lett* **245:** 195-203.

Rosa-Fraile M, Rodriguez-Granger J, Cueto-Lopez M, Sampedro A, Gaye EB, Haro JM, Andreu A (1999). Use of Granada medium to detect group B streptococcal colonization in pregnant women. *J Clin Microbiol* **37:** 2674-2677.

Rosa-Fraile M, Dramsi S, Spellerberg B (2014). Group B streptococcal haemolysin and pigment, a tale of twins. *FEMS Microbiol Rev* **38:** 932-946.

Rosini R, Rinaudo CD, Soriani M, Lauer P, Mora M, Maione D, Taddei A, Santi I, Ghezzo C *et al* (2006). Identification of novel genomic islands coding for antigenic pilus-like structures in *Streptococcus agalactiae*. *Mol Microbiol* **61:** 126-141.

Rosini R, Campisi E, De Chiara M, Tettelin H, Rinaudo D, Toniolo C, Metruccio M, Guidotti S, Sorensen UBS *et al* (2015). Genomic analysis reveals the molecular basis for capsule loss in the group B *Streptococcus* population. *PLOS One* **10**.

Rosinski-Chupin I, Sauvage E, Mairey B, Mangenot S, Ma L, Da Cunha V, Rusniok C, Bouchier C, Barbe V *et al* (2013). Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage. *BMC Genomics* **14:** 252.

Rosinski-Chupin I, Sauvage E, Sismeiro O, Villain A, Da Cunha V, Caliot M-E, Dillies M-A, Trieu-Cuot P, Bouloc P *et al* (2015). Single nucleotide resolution RNA-seq uncovers new regulatory mechanisms in the opportunistic pathogen *Streptococcus agalactiae*. *BMC Genomics* **16**.

Ross KF, Ronson CW, Tagg JR (1993). Isolation and characterization of the lantibiotic salivaricin A and its structural gene salA from *Streptococcus salivarius* 20P3. *Appl Environ Microbiol* **59:** 2014-2021.

Rubens CE, Raff HV, Jackson JC, Chi EY, Bielitzki JT, Hillier SL (1991). Pathophysiology and histopathology of group-B streptococcal sepsis in macaca-nemestrina primates induced after intraamniotic inoculation - evidence for bacterial cellular invasion. *J Infect Dis* **164:** 320-330.

Ruegg PL, Reinemann DJ (2002). Milk quality and mastitis tests. In: Smith RA (ed). *Bovine Practitioner, Vol 36, No 1*. pp 41-54.

Santi I, Scarselli M, Mariani M, Pezzicoli A, Masignani V, Taddei A, Grandi G, Telford JL, Soriani M (2007). BibA: a novel immunogenic bacterial adhesin contributing to group B *Streptococcus* survival in human blood. *Mol Microbiol* **63:** 754-767.

Schneewind O, Friedrich K, Lutticken R (1988). Cloning and expression of the CAMP factor of group B streptococci in *Escherichia coli*. *Infect Immun* **56:** 2174-2179.

Schubert A, Zakikhany K, Pietrocola G, Meinke A, Speziale P, Eikmanns BJ, Reinscheid DJ (2004). The fibrinogen receptor FbsA promotes adherence of *Streptococcus agalactiae* to human epithelial cells. *Infect Immun* **72:** 6197-6205.

Schwartz DC, Cantor CR (1984). Separation of yeast chromosome-sized dnas by pulsed field gradient gel-electrophoresis. *Cell* **37:** 67-75.

Schwarz G (1978). *Estimating the dimension of a model*. The Institute of Mathematical Statistics.

Seale AC, Koech AC, Sheppard AE, Barsosio HC, Langat J, Anyango E, Mwakio S, Mwarumba S, Morpeth SC *et al* (2016). Maternal colonization with *Streptococcus agalactiae* and associated stillbirth and neonatal disease in coastal Kenya. *Nat Microbiol* **1:** 16067-16067.

Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarrraga AM *et al* (2006). The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet* **38:** 779-786.

Seifert KN, Adderson EE, Whiting AA, Bohnsack JF, Crowley PJ, Brady LJ (2006). A unique serine-rich repeat protein (Srr-2) and novel surface antigen (epsilon) associated with a virulent lineage of serotype III *Streptococcus agalactiae*. *Microbiol-Sgm* **152:** 1029-1040.

Seo HS, Mu R, Kim BJ, Doran KS, Sullam PM (2012). Binding of glycoprotein Srr1 of *Streptococcus agalactiae* to fibrinogen promotes attachment to brain endothelium and the development of meningitis. *PLOS Pathog* **8**.

Sharkey LKR, Edwards TA, O'Neill AJ (2016). ABC-F proteins mediate antibiotic resistance through ribosomal protection. *mBio* **7**.

Sheen TR, Jimenez A, Wang N-Y, Banerjee A, van Sorge NM, Doran KS (2011). Serine-rich repeat proteins and pili promote *Streptococcus agalactiae* colonization of the vaginal tract. *J Bacteriol* **193:** 6834-6842.

Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, Kelly DJ, Cody A, Colles FM *et al* (2013a). Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol Ecol* **22:** 1051-1064.

Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, Bentley SD, Maiden MCJ, Parkhill J *et al* (2013b). Genome-wide association study identifies vitamin B-5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc Natl Acad Sci U S A* **110:** 11923-11927.

Shome BR, Das Mitra S, Bhuvana M, Krithiga N, Velu D, Shome R, Isloor S, Barbuddhe SB, Rahman H (2011). Multiplex PCR assay for species identification of bovine mastitis pathogens. *J Appl Microbiol* **111:** 1349-1356.

Simonsen KA, Anderson-Berry AL, Delair SF, Davies HD (2014). Early-onset neonatal sepsis. *Clin Microbiol Rev* **27:** 21-47.

Smith AJ, Ward PN, Field TR, Jones CL, Lincoln RA, Leigh JA (2003). MtuA, a lipoprotein receptor antigen from *Streptococcus uberis*, is responsible for acquisition of manganese during growth in milk and is essential for infection of the lactating bovine mammary gland. *Infect Immun* **71:** 4842-4849.

Sorek R, Kunin V, Hugenholtz P (2008). CRISPR - a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6:** 181-186.

Sorensen UB, Poulsen K, Ghezzo C, Margarit I, Kilian M (2010). Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones. *mBio* **1**.

Spellerberg B, Pohl B, Haase G, Martin S, Weber-Heynemann J, Lutticken R (1999). Identification of genetic determinants for the hemolytic activity of *Streptococcus agalactiae* by ISS1 transposition. *J Bacteriol* **181:** 3212-3219.

Springman AC, Lacher DW, Waymire EA, Wengert SL, Singh P, Zadoks RN, Davies HD, Manning SD (2014). Pilus distribution among lineages of group B *Sreptococcus*: an evolutionary and clinical perspective. *BMC Microbiol* **14**.

Stalhammarcarlemalm M, Stenberg L, Lindahl G (1993). Protein Rib - a novel group-B streptococcal cell-surface protein that confers protective immunity and is expressed by most strains causing invasive infections. *J Exp Med* **177:** 1593-1603.

Taddei F, Radman M, MaynardSmith J, Toupance B, Gouyon PH, Godelle B (1997). Role of mutator alleles in adaptive evolution. *Nature* **387:** 700-702.

Tapsall JW (1987). Relationship between pigment production and hemolysin formation by lancefield group-B streptococci. *J Med Microbiol* **24:** 83-87.

Tavaré S (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*. Amer Mathematical Society. pp 57-86.

Tazi A, Disson O, Bellais S, Bouaboud A, Dmytruk N, Dramsi S, Mistou MY, Khun H, Mechler C *et al* (2010). The surface protein HvgA mediates group B *Streptococcus* hypervirulence and meningeal tropism in neonates. *J Exp Med* **207:** 2313-2322.

Teatero S, Ramoutar E, McGeer A, Li A, Melano RG, Wasserscheid J, Dewar K, Fittipaldi N (2016). Clonal Complex 17 group B *Streptococcus* strains causing invasive disease in neonates and adults originate from the same genetic pool. *Sci Rep* **6**.

Tenhagen BA, Koster G, Wallmann J, Heuwieser W (2006). Prevalence of mastitis pathogens and their resistance against antimicrobial agents in dairy cows in Brandenburg, Germany. *J Dairy Sci* **89:** 2542-2551.

Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I *et al* (2002). Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc Natl Acad Sci U S A* **99:** 12391-12396.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL *et al* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102:** 13950-13955.

Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, Quail MA, Stevens M, Jones MA *et al* (2008). Comparative genome analysis of *Salmonella Enteritidis* PT4 and *Salmonella Gallinarum* 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Res* **18:** 1624-1637.

Tibary A, Fite C, Anouassi A, Sghiri A (2006). Infectious causes of reproductive loss in camelids. *Theriogenology* **66:** 633-647.

Todd EW (1934). A comparative serological study of streptolysins derived from human and from animal infections, with notes on pneumococcal hæmolysin, tetanolysin and staphylococcus toxin. *J Pathol Bacteriol* **39:** 299-321.

Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T *et al* (2009). The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459:** 950-956.

Tong SY, Holden MT, Nickerson EK, Cooper BS, Koser CU, Cori A, Jombart T, Cauchemez S, Fraser C *et al* (2015). Genome sequencing defines phylogeny and spread of methicillin-resistant *Staphylococcus aureus* in a high transmission setting. *Genome Res* **25:** 111-118.

Toprak E, Veres A, Michel J-B, Chait R, Hartl DL, Kishony R (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet* **44:** 101-U140.

Toussaint A, Merlin C (2002). Mobile elements as a combination of functional modules. *Plasmid* **47:** 26-35.

Udo EE, Aly NYA, Sarkhoo E, Al-Sawan R, Al-Asar A-SM (2011). Detection and characterization of an ST97-SCCmec-V community-associated meticillin-resistant *Staphylococcus aureus* clone in a neonatal intensive care unit and special care baby unit. *J Med Microbiol* **60:** 600-604.

van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, Fussing V, Green J, Feil E *et al* (2007). Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin Microbiol Infec* **13:** 1-46.

van der Mee-Marquet N, Fourny L, Arnault L, Domelier A-S, Salloum M, Lartigue M-F, Quentin R (2008). Molecular characterization of human-colonizing Streptococcus agalactiae strains isolated from throat, skin, anal margin, and genital body sites. *J Clin Microbiol* **46:** 2906-2911.

van Sorge NM, Quach D, Gurney MA, Sullam PM, Nizet V, Doran KS (2009). The group B streptococcal serine-rich repeat 1 glycoprotein mediates penetration of the blood-brain barrier. *J Infect Dis* **199:** 1479-1487.

Viana D, Comos M, McAdam PR, Ward MJ, Selva L, Guinane CM, Gonzalez-Munoz BM, Tristan A, Foster SJ *et al* (2015). A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat Genet* **47:** 361-U195.

Vockenhuber MP, Sharma CM, Statt MG, Schmidt D, Xu ZJ, Dietrich S, Liesegang H, Mathews DH, Suess B (2011). Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. *RNA Biol* **8:** 468-477.

Wastfelt M, StalhammarCarlemalm M, Delisse AM, Cabezon T, Lindahl G (1996). Identification of a family of streptococcal surface proteins with extremely repetitive structure. *J Biol Chem* **271:** 18892-18897.

Watts JL (1988). Etiological agents of bovine mastitis. *Vet Microbiol* **16:** 41-66.

Wexler DE, Chenoweth DE, Cleary PP (1985). Mechanism of action of the group-A streptococcal C5a inactivator. *Proc Natl Acad Sci U S A* **82:** 8144-8148.

Whidbey C, Harrell MI, Burnside K, Ngo L, Becraft AK, Iyer LM, Aravind L, Hitti J, Waldorf KMA *et al* (2013). A hemolytic pigment of group B *Streptococcus* allows bacterial penetration of human placenta. *J Exp Med* **210:** 1265-1281.

Whidbey C, Vornhagen J, Gendrin C, Boldenow E, Samson JM, Doering K, Ngo L, Ezekwe EAD, Jr., Gundlach JH *et al* (2015). A streptococcal lipid toxin induces membrane permeabilization and pyroptosis leading to fetal injury. *EMBO Mol Med* **7:** 488-505.

Wilkinso HW, Thacker LG, Facklam RR (1973). Nonhemolytic group B streptococci of human, bovine and ichthyic origin. *Infect Immun* **7:** 496-498.

Wilson AC, Ochman H, Prager EM (1987). Molecular time scale for evolution. *Trends Genet* **3:** 241-247.

Wirawan RE, Kleese NA, Jack RW, Tagg JR (2006). Molecular and genetic characterization of a novel nisin variant produced by *Streptococcus uberis*. *Appl Environ Microbiol* **72:** 1148-1156.

Worby CJ, Lipsitch M, Hanage WP (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLOS Comput Biol* **10:** e1003549.

Wyder AB, Boss R, Naskova J, Kaufmann T, Steiner A, Graber HU (2011). *Streptococcus* spp. and related bacteria: Their identification and their pathogenic potential for chronic mastitis - A molecular approach. *Res Vet Sci* **91:** 349-357.

Xavier JB (2011). Social interaction in synthetic and natural microbial communities. *Mol Syst Biol* **7**.

Yang L, Jelsbak L, Marvig RL, Damkiaer S, Workman CT, Rau MH, Hansen SK, Folkesson A, Johansen HK *et al* (2011). Evolutionary dynamics of bacteria in a human host environment. *Proc Natl Acad Sci U S A* **108:** 7481-7486.

Yildirim A, Lammler C, Weiss R (2002a). Identification and characterization of *Streptococcus agalactiae* isolated from horses. *Vet Microbiol* **85:** 31-35.

Yildirim AO, Fink K, Lammler C (2002b). Distribution of the hyaluronate lyase encoding gene hylB and the insertion element is 1548 in streptococci of serological group B isolated from animals and humans. *Res Vet Sci* **73:** 131-135.

Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K *et al* (2012). Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A* **109:** 4550-4555.

Zharkikh A, Li WH (1992). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide-sequences. *Mol Biol Evol* **9:** 1119-1147.