UNIVERSITÀ
DI SIENA
1240

UNIVERSITY OF SIENA

DEPARTMENT OF MEDICAL BIOTECHNOLOGIES

PhD COURSE IN MEDICAL BIOTECHNOLOGIES

COORDINATOR: PROF. LORENZO LEONCINI

XXXIV CYCLE

Whole genome sequencing and comparative genomics in lactic acid bacteria

Supervisor:
Prof. Gianni Pozzi

Co-supervisors:
Prof. Francesco Iannelli
Dr. Francesco Santoro

PhD candidate:
Lorenzo Colombini

Academic year 2020–2021

Table of Contents

ABSTRACT 3

CHAPTER 1. General introduction 4

1. Lactic acid bacteria 4
  1.1. The genus *Lactobacillus* 5
    1.1.1. *Lactobacillus crispatus* 6
  1.2. The genus *Streptococcus* 7
    1.2.1. *Streptococcus pneumoniae* 8
  1.3. The genus *Enterococcus* 9
    1.3.1. *Enterococcus faecalis* 10

2. Bacterial genomics and whole genome sequencing 12
  2.1. Nanopore sequencing technology 13
  2.2. Genome annotation and comparison tools 15

3. Aim of the thesis 17

4. References 18

CHAPTER 2. The mobilome of probiotic *Lactobacillus crispatus* M247 includes Tn*7088* a novel transposon carrying a biosynthetic gene cluster for a class I bacteriocin 39

CHAPTER 3. The unusual duplication of a 69.9-kb long chromosomal segment produces two copies of an inverted repeat and generates genomic instability in a laboratory strain of *Lactobacillus crispatus* 81

CHAPTER 4. Sequence typing and antimicrobial susceptibility testing of infertility-associated *Enterococcus faecalis* reveals clonality of aminoglycoside resistant strains 126

CHAPTER 5. Nucleotide sequence analysis of the novel composite transposon Tn*7086* carrying aminoglycoside resistance genes in infertility-associated *Enterococcus faecalis* belonging to the clonal complex/sequence type CC16/ST480 158

CHAPTER 6. Complete genome sequence of *Lactobacillus crispatus* type strain ATCC 33820 199

CHAPTER 7. Complete genome sequence of the *Streptococcus pneumoniae* strain Rx1, a Hex mismatch repair-deficient standard transformation recipient 203

CHAPTER 8. General conclusions 208

APPENDIX. Scientific *Curriculum Vitae* 210

ABSTRACT

In the present thesis, the genomes of different microbial species, belonging to the lactic acid bacteria and including *Lactobacillus crispatus*, *Streptococcus pneumoniae* and *Enterococcus faecalis*, were obtained and analyzed using comparative genomics tools. In the first part of the thesis was described the genome of the probiotic *L. crispatus* strain M247, which contains a novel integrative and mobilizable element named Tn*7088*. Tn*7088* carries a biosynthetic gene cluster coding for a class I bacteriocin which is homologous to the listeriolysin S gene cluster of *Listeria monocytogenes* and may confer selective advantages towards related bacterial species. Chromosomal rearrangements mediated by insertion sequences and involving two regions of 69.9-kb and 15.4-kb, were detected in the M247 strain. A *L. crispatus* M247 laboratory strain carried in our laboratory strain collection since 1990 and named M247_Siena, showed an unusual duplication of the 69.9-kb DNA region resulting in the generation of two long inverted repeats (LIRs) and the deletion of the 15.4-kb region. Analysis of ultra-long DNA Nanopore reads showed that the presence of LIRs in strain M247_Siena increased the intrinsic genome instability of strain M247. In the second part, a collection of 41 *E. faecalis* strains isolated from genital tract samples of infertile couples, was subjected to antimicrobial susceptibility testing and whole genome sequencing. Multi locus sequence typing and antimicrobial susceptibility testing results suggested clonality of infertility-associated *E. faecalis* isolates resistant to high-level aminoglycosides. Analysis of the genomic location of aminoglycoside modifying enzyme (AME) genes led to the identification of a family of novel composite transposons, whose reference element was denominated Tn*7086*. Tn*7086* and Tn*7086*-like elements in infertility-associated *E. faecalis* shared the following traits: i) are flanked by two direct repeats of the IS*1216E* element, ii) employ the same chromosomal *panE* gene integration site, iii) excise from the bacterial chromosome leaving an IS*1216E* copy in the chromosome and form circular intermediates in which the ends are joined by the other IS*1216E* copy. Finally, the whole genome sequences of the *L. crispatus* type strain ATCC 33820 and of *S. pneumoniae* laboratory strains Rx1 and R36A, were obtained and analyzed.

CHAPTER 1. General introduction

1. Lactic acid bacteria

The first pure culture of a lactic acid bacterium was isolated in 1873 by Joseph Lister and designated as *Bacterium lactis* (now *Lactococcus lactis*) for its capacity of causing the lactic acid fermentation of milk (Santer, 2010). The metabolic-based term "lactic acid bacteria" (LAB) is now used to define a phylogenetically heterogeneous group of microorganisms which are Gram-positive, usually catalase negative, microaerophilic, acid-tolerant, non-sporulating rods and cocci and characterized by their ability to produce lactic acid as main-end product of their metabolism (Hayek & Ibrahim, 2013; Quinto et al., 2014). LAB have coevolved with plants, invertebrates and vertebrates, establishing mutualism, symbiosis, commensalism, or even parasitism-like behavior with their host and consequentially, are associated with niches of dairy (fermented), meat, and vegetable origin, with the gastrointestinal and urogenital tracts of humans and animals, with soil and water (George et al., 2018; Liu et al., 2014). Lactic acid bacteria can be differentiated upon various criteria as morphology, growth temperature and ability to ferment glucose. Based on their fermentation features LAB are categorized into i) homofermentative LAB which mainly produce lactic acid from sugars and ii) heterofermentative LAB producing as lactic acid side products acetic acid or alcohol and carbon dioxide (Carr et al., 2002). Phylogenetically the LAB group consists of two major branches namely the *Clostridium* branch and the actinomycetes branch, identified essentially by 16S rDNA sequencing (Woese, 1998). The *Clostridium* branch includes LAB of the Firmicutes phylum with genera as *Enterococcus, Lactobacillus, Lactococcus, Leuconostoc, Pediococcus, Streptococcus* and *Weissella*, which all belong to the order *Lactobacillales* and are low-GC content (31-49%) organisms; whereas the actinomycetes branch contains the *Bifidobacterium* genus of the Actinobacteria phylum, which have a high-GC content (58-61%) (H. Zhang & Cai, 2014). Due to their versatile metabolism and properties, strains of Lactic acid bacteria are used as starter cultures in the dairy industry, as probiotics in dietary supplements and as bioconversion agents in the production of interesting compounds (i.e., nutraceuticals) (Ruiz-Rodríguez et al., 2016).

1.1. The genus *Lactobacillus*

Historically the genus *Lactobacillus* constituted the largest and most diverse group among LAB, including more than 250 species, extremely diverse at phenotypic, ecological and genotypic level. Recent work, based on whole genome sequence analysis led to a reclassification into 25 genera including the emended genus *Lactobacillus* which includes host-adapted organisms that have been referred to as the *Lactobacillus delbrueckii* group, the genus *Paralactobacillus* and 23 novel genera for which other names have been proposed (Zheng et al., 2020). *Lactobacillus* species are Gram-positive, homofermentative, thermophilic, non-spore forming rods, which adapted to their vertebrate host, except for the *Lactobacillus melliventris* clade that is adapted to social bees (Martinson et al., 2011). Lactobacilli are normal inhabitants of the oral cavity, the gastrointestinal tract and the urogenital tract (Ahrné et al., 1998; Hillier et al., 1993). *Lactobacillus* spp. has been identified as the most abundant genus throughout the female reproductive system of healthy women and alterations in their quantity and quality have been associated with different gynecological disorders (Kyono et al., 2018; Moreno et al., 2016; Ravel et al., 2011; S. B. Smith & Ravel, 2017). Indeed, *Lactobacillus* spp. contribute to the maintenance of vaginal homeostasis by different direct and indirect anti-pathogenic mechanisms such as i) induction of an acid pH associated to lactic acid production, ii) hydrogen peroxide production, iii) formation of microcolonies that create a physical barrier against pathogen colonization and iv) induction of immune response against pathogen (Aldunate et al., 2015; Tachedjian et al., 2018; Tyssen et al., 2018). Furthermore, strains of genus *Lactobacillus* are able to produce bacteriocins, which are ribosomally synthesized peptides exerting antimicrobial activity toward strains of species related to the producing species (Collins et al., 2017; Zacharof & Lovitt, 2012). In the gastrointestinal tract (GIT) of humans and animals, *Lactobacillus spp*. are found in variable amounts according to the animal species, the age of the host or the location within the gut. However, with the advent of new techniques of identification including Next Generation Sequencing (NGS), it has been estimated that autochthonous lactobacilli of GIT constitute only the 0.01 - 0.6% of human adult fecal microbiota (Heeney et al., 2018; Lebeer et al., 2008).

1.1.1. *Lactobacillus crispatus*

Lactobacillus crispatus was first isolated in 1953 at the Institut of Pasteur by Brygoo and Aladame (Brygoo & Aladame, 1953). Initially considered a new species of the genus *Eubacterium*, *L. crispatus* was later identified as *Lactobacillus* and considered a synonymous with "*L. acidophilus* group A2" (CATO et al., 1983). Recently, Zheng and colleagues (Zheng et al., 2020) reclassified the genus *Lactobacillus* into 25 genera; however, the nomenclature of *L. crispatus* remained unchanged. *L. crispatus* is the most frequently isolated species among the vaginal lactobacilli of the human microbiota of healthy women (Raven et al., 2011). The presence of *L. crispatus* in the vaginal microbiota is associated with reduced risk of preterm delivery, viral sexually transmitted infections, and bacterial vaginosis (Petrova et al., 2015). Furthermore, *L. crispatus* is one of the few *Lactobacillus* species isolated from human gut (El Aila et al., 2009) and it is associated to animal health, particularly to poultry (chicken and turkey) gut health (Dec et al., 2018; Wei et al., 2013). Despite the growing interest in *L. crispatus* strains suitable to be used as probiotic for both women and poultry, only limited information has been elucidated on the genetic bases conferring a pivotal role to *L. crispatus* in the human vaginal and poultry gut niche. To date (December 12, 2021) GenBank hosts 169 deposited genomes of *L. crispatus*, of these only 14 are complete, while 155 are draft genomes assembled either in scaffolds or in contigs (https://www.ncbi.nlm.nih.gov/genome/browse/#!/prokaryotes/1815/). The median total length of the *L. crispatus* complete genome is 2.371 Megabases (Mbs), encoding for 2,223 genes, with an average GC content of 37 %. *L. crispatus* genome is bigger than the genome of other well-known vaginal *Lactobacillus* species such as *L. gasseri* (2.046 Mbs), *L. jensenii* (1.640 Mbs) and *L. iners* (1.404 Mbs). This size difference is consistent with the *L. crispatus* genome coding for several protein families that are involved in organismal interaction such as resistance to phage infections, bacteriocin-type sequences, toxin-antitoxin systems and for a high number of proteins related to mobile genetic elements such as transposases (Mendes-Soares et al., 2014). Comparative genomic analysis of *L. crispatus* genomes were performed to analyze the genetic diversity and the population structure of this species (Abdelmaksoud et al., 2016; Ojala et al., 2014; Pan et al., 2020; van der Veer et al., 2019). A first study based on 10 genome sequences estimated the size of the core genome to level at about 1,116 genes (Ojala et al., 2014), while a more recent work on a larger dataset (105 genomes) estimated that the core genome of *L. crispatus* consists of 465 genes. Phylogenetical analysis indicates a clear separation of *L. crispatus* strains regarding their isolation source (human or poultry) and also that human gut and vaginal isolates cluster separately. Compared to *L. crispatus* gut isolates, strains from vaginal environment present CRISPR loci with a reduced number of spacers (mainly of Type-II-A CRISPR system) (Pan et al., 2020) and this probably reflects the higher prevalence of phages in the gut environment (Stern et al., 2012). In addition, gut isolates tend to display a complete exopolysaccharide (EPS) biosynthesis gene cluster constituted by 16 genes (priming glycosyltransferase (*p-gtf*), glycosyltransferases, flippase, tyrosine kinase (*epsC-D*) tyrosine phosphatase, capsular polysaccharide gene (*cpsA*), rhamnose and membrane transporters, among others), whereas vaginal isolates tend to display a variable *eps* cluster which is very often truncated or incomplete (Ojala et al., 2014; Pan et al., 2020). Comparison of the genetic content of *L. crispatus* isolates from healthy lactobacilli-dominated vaginal microbiomes (LVM) with isolates from dysbiotic vaginal microbiomes (DVM), showed similar content of lactic acid production genes and phages, with similar phage-induced lysis rate, however LVM isolates were more likely to carry glycosyltransferase genes and DVM isolates genes for cellobiose transport (Abdelmaksoud et al., 2016; van der Veer et al., 2019).

1.2. The genus *Streptococcus*

Streptococci were first observed in 1874 by Billroth (Jones, 1978) and are characterized by ovoid or spherical, Gram-positive cells arranged in pairs or in chains which can be long up to 50 cells. These cocci are facultatively anaerobic, non-sporing, catalase-negative, homofermentative, and have complex nutritional requirements. Streptococci are normal inhabitants of the mucosal surfaces of humans and other mammalians, with some species found also on the skin and others that may be isolated from milk and dairy products (Wood & Holzapfel, 1995). Streptococci are the dominant species in the oral cavity and upper respiratory tract (Abranches et al., 2018). Based on hemolysis pattern and carbohydrate "group" antigens (Lancefield groups), Streptococci were initially differentiated in two groups namely the "pyogenic" and the "viridans" (Sherman JM., 1937). More recently, based on phylogenetic analysis streptococci were separated into 8 distinct "species groups", namely "mitis", "sanguinis", "anginosus", "salivarius", "downei", "mutans, "pyogenic", and "bovis" which comprise most of the described species in the genus (Richards et al., 2014). The genus *Streptococcus* represents one of the most invasive group of bacteria which includes typical human pathogenic species such as *S. pyogenes*, *S. agalactiae, S. pneumoniae* and many other species capable of acting as opportunistic pathogens under appropriate circumstances (Krzyściak et al., 2013).

1.2.1. *Streptococcus pneumoniae*

S. pneumoniae (the "pneumococcus") is one of the most important human pathogens, causing invasive infections such as meningitis, sepsis, pneumonia, and mild mucosal infections as acute otitis media, sinusitis and conjunctivitis. At the same time pneumococcus is also a common inhabitant of the human nasopharynx, where it can stay as a commensal without causing disease. The transmission, colonization and invasion of *S. pneumoniae* depend on its ability to evade or take advantage of the host inflammatory and immune responses (Weiser et al., 2018). The primary virulence factor of *S. pneumoniae* is an extracellular polysaccharidic capsule which surrounds the bacterium conferring protection from mucus-mediated clearance, environmental stresses and phagocytosis (García et al., 1997; Nelson et al., 2007). A total of 100 different pneumococcal capsular serotypes have been identified to date (Ganaie et al., 2020), based on the biochemical composition and antigenic properties (Paton & Trappetti, 2019). The genetic locus encoding the genes for the synthesis of the capsular polysaccharide is located to the same position on the chromosome between *dexB* and *aliA* genes in all serotypes except type 37 (Llull et al., 1999), it is variable in length (approximately 10–30 kb) and has an essentially conserved block-wise arrangement (Paton & Trappetti, 2019). Other *S. pneumoniae* virulence factors include surface proteins and enzymes, such as the choline-binding proteins (PspA, PspC and LytA), and the toxin pneumolysin (Berry & Paton, 2000; Brown et al., 2015; Kaetzel, 2001; Tomasz et al., 1970; Tu et al., 1999). To date (December 12, 2021) GenBank hosts 8,973 deposited genomes of *S. pneumoniae*, of these 90 are complete, while 8,883 are draft genomes assembled either in scaffolds or in contigs (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/176/). The median total length of the *E. faecalis* genome is 2.085 Mbs, encoding for 1,951 genes, with an average GC content of 39.6 %. One of the keys of the success of pneumococcus as a pathogen is its genome plasticity. Indeed, *S. pneumoniae* is naturally transformable and therefore, readily able to internalize and integrate heterologous DNA into its genome through the competence system (Straume et al., 2015), resulting in rapid variations such as serotype changes (Coffey et al., 1998). Furthermore, *S. pneumoniae* genome is also shaped by the presence of mobile genetic elements such as integrative and conjugative elements responsible for pneumococcal genome evolution and more particularly for virulence and drug resistance acquisition (Croucher et al., 2009, 2011). All pneumococcal plasmids are cryptic as they do not code for genes conferring observable phenotypes. Most of the pneumococcal plasmids isolated over years are nearly identical to pDP1, which was the first plasmid of *S. pneumoniae* to be isolated in 1979 in strain D39 and its derivatives (M. D. Smith & Guild, 1979), with the exception of pSpnP1 (Romero et al., 2007).

1.3. The genus *Enterococcus*

Enterococci were first described in 1899 by Thiercelin (Thiercelin & Jouhaud, 1899), but were phylogenetically grouped within the genus *Streptococcus* (Sherman JM., 1937) and classified as a new genus only in 1984 upon DNA hybridization analysis and 16S rRNA sequencing (Schleifer & Kilpper-Bälz, 1984). Species within the genus Enterococcus have ovoid, Gram-positive cells occurring singly, in pairs or in short chains in which cells are elongated in the direction of the chain. Enterococci are present in the intestinal tracts of humans and animals and in the environments these organisms inhabit (Klein, 2003; Murray, 1990). Enterococci are the predominant Gram-positive cocci found within the gastrointestinal tract and in humans can be isolated at concentrations of 10^5 to 10^7 CFU/gram feces (Jett et al., 1994). Despite being auxotrophic for many amino acids, vitamins and micronutrients (Niven & Sherman, 1944), enterococci present a strong survival ability resulting from their tolerance to UV irradiation (Maraccini et al., 2012), salt concentration, starvation (Hartke et al., 1998) and predation by bacteriophages (Duerkop et al., 2016; Purnell et al., 2011). Enterococci are also able to replicate in the environment, possibly as the results of collaboration within polymicrobial consortium (Byappanahalli et al., 2012; Desmarais et al., 2002; Yamahara et al., 2009). Horizontal gene transfer in enterococci is favored by the fact that they exist in complex microbial ecosystem, in intimate contact with large diversity of potential sources of genetic material; furthermore, due to their high level of intrinsic antibiotic resistance, enterococci occur in environment substantially enriched for antibiotic-resistance elements (Van Tyne & Gilmore, 2014). The ability of enterococci to acquire mobile genetic elements, conveying antimicrobial resistance and virulence traits among both Gram-positive and -negative species, has contributed to their emergence as leading hospital pathogens (Palmer et al., 2010; Pöntinen et al., 2021).

1.3.1. *Enterococcus faecalis*

Enterococcus faecalis is one of the most abundant enterococci in human feces and the species responsible for the majority of enterococcal infections in humans, including urinary tract infections (UTIs), sepsis, endocarditis, peritonitis, abdominal/pelvic and soft tissue infection (Agudelo Higuita & Huycke, 2014; Lebreton et al., 2014). The most frequent clinical manifestation is UTI, of which *E. faecalis* is the second most common agent worldwide after *Escherichia coli* (Flores-Mireles et al., 2015). *E. faecalis* is also the leading pathogen among Gram-positive bacteria of catheter-associated UTIs (CAUTIs) in healthcare settings (Peng et al., 2018). Ascending UTIs and intra-abdominal infections can lead to bacteremia and endocarditis. Both *Enterococcus faecium* and *E. faecalis* have a remarkable tropism for the endocardium and/or the heart valves, but *E. faecalis* alone accounts for about 90% of enterococcal endocarditis cases, especially in risk groups (Fernández-Hidalgo et al., 2020). To date (December 12, 2021) GenBank hosts 2,076 deposited genomes of *E. faecalis*, of these 79 are complete, while 1997 are draft genomes assembled either in scaffolds or in contigs (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/808/). The median total length of the *E. faecalis* genome is 2.973 Mbs, encoding for 2,753 genes, with an average GC content of 37.4 %. Compared to commensal representatives, hospital-adapted *E. faecalis* strains generally contain a larger mobilome which accounts for over a quarter of the genome as observed in the vancomycin resistant *E. faecalis* strain V583 (Bourgogne et al., 2008; Hegstad et al., 2010; Weaver, 2019). Indeed, hospital-adapted, multidrug-resistant lineages of *E. faecalis* include strains of multiple-locus sequence type clonal clusters CC2, CC9, CC28, and CC40 (McBride et al., 2007; Ruiz-Garbajosa et al., 2006) in which a variety of auxiliary traits such as antibiotic resistance genes (Lebreton et al., 2013; McBride et al., 2007), Enterococcal surface protein (Esp)-containing pathogenicity island (Leavis et al., 2004; Tendolkar et al., 2004) and more complex integral cell wall carbohydrate operons (Palmer et al., 2012; Solheim et al., 2011), have converged on mobile genetic elements. Recently, a comparative genomic study involving 2027 *E. faecalis* genomes from isolates spanning a wide range of isolation years and sources, indicated that apparent adaptation to the hospital-associated niche is actually likely to be due to selection for survival in a broader set of niches, consistent with *E. faecalis* having a generalist nature (Pöntinen et al., 2021). Vectors of horizontal transmission of most of the antibiotic resistance genes in *E. faecalis* are pheromone-responsive plasmids such as pCF10 (Dunny, 2007) and pAD1 (Clewell, 2007). Conjugation of these plasmids is induced by pheromones chromosomally encoded within genes for lipoprotein signal peptides and secreted by potential recipient cells (Palmer et al., 2010). Pheromone-responsive plasmids provide also accessory genes encoding bacteriocin, cytolysin production and ultraviolet resistance and probably evolved to shuttle niche specialization traits allowing *E. faecalis* as a species to readily adapt to a particular host (Palmer et al., 2010). However, the majority of the mobilome in enterococci is represented by transposable elements distributed on both chromosomes and plasmids (Lam et al., 2012; Paulsen et al., 2003; Qin et al., 2012) and including (i) composite transposon (class I transposons), (ii) Tn*3*/*21* family transposons (class II transposons) and (iii) integrative and conjugative elements (ICEs) comprehensive of the classical conjugative transposons (Werner et al., 2013). Composite transposons are flanked by copies of insertion sequences of the same family that act together to move the DNA between them, while Tn*3*/*21* family transposons are bounded by short inverted repeats and contain both genes needed for transposon movement and the accessory genes (Harmer et al., 2020). ICEs are conjugative and self-transmissible elements capable of excise from and integrate into the host chromosomes and like conjugative plasmids contribute to the antibiotic resistance genes diffusion among *Enterococcus spp.* ICEs members of the Tn*916/*Tn*1545* family, all related to Tn*916* originally discovered in *E. faecalis* (Franke & Clewell, 1981), are broad host range ICEs responsible for a large proportion of the antibiotic resistance observed in *E. faecalis*, but also in other bacterial genera such as *Staphylococcus* and *Streptococcus* (Hegstad et al., 2010; Roberts & Mullany, 2011).

2. Bacterial genomics and whole genome sequencing

Bacterial genomics involves the study and comparison of whole bacterial genomes using nucleic acid sequencing technologies and computational analysis tools (Casjens, 1998). Whole-genome sequencing (WGS) represents a "top-down" approach to associate genotype with phenotype and holds the potential to enable rapid bacterial profiling and pathogen identification, leading further details about the molecular basis of virulence and antibiotic-resistance acquisition and allowing population studies via comparative analysis. Comparison of whole genomes allows the identification of large genomic variations, including insertions, deletions, inversions, translocations and duplications, which can all contribute to the unique genotypic composition of each isolate. Such structural variations can be identified through both sequence assembly and read mapping. Indeed, disproportionate read coverage of reads mapping to a reference genome can be used to detect deletions (manifested as an absence of reads mapping to that region of the genome) and duplications of the genome (manifested as a doubling of reads mapping to that region of the genome) (Bryant et al., 2012). Bacterial genome sizes can differ over a greater than tenfold range, ranging from 580 kbp for *Mycoplasma genitalium* (Fraser et al., 1995) up to 9,140 kbp for *Myxococcus xanthus* (Goldman et al., 2006) with specialist bacteria having smaller genomes compared to bacteria that are metabolic generalists and/or undergo some form of development such as sporulation, mycelium formation. Genome assembly reconstructs a genome from many shorter reads (Miller et al., 2010; Nagarajan & Pop, 2013; Pop, 2009). The advent of novel sequencing technologies has enabled more rapid, cost-effective and precise microbial sequencing. However, the assembly of complete bacterial genomes remains a challenging process with DNA sequence repeats representing the primary obstacle (Koren & Phillippy, 2015). DNA repeats in prokaryotes are causes and consequences of genome plasticity which may origin through intrachromosomal recombination or horizontal transfer and in turn lead to genetic material amplifications, deletions, and rearrangement via recombination processes (Treangen et al., 2009). Bacterial genomes have been classified based on their repeats content in i) class I genomes having few repeats other than the rDNA operon (7-8kb), ii) class II containing in addition many mid-scale repeats such as insertion sequences, with rDNA operon still being the longest, iii) class III containing repeats significantly larger than the rDNA operon (Koren et al., 2013). Among bacteria harboring a class III genomes it is estimated that a small percentage (~3%) contains long near identical repeats above 30 kb to over 100 kb in length, although this fraction could be under-represented *per se* among completely sequenced genomes (Schmid et al., 2018).

2.1. Nanopore sequencing technology

In 2014 the Oxford Nanopore Technologies (ONT) company (United Kingdom) released the Nanopore sequencing technology which directly targets single DNA molecules and is currently the only sequencing technology based on DNA translocation through biological nanopores. The core of the technology is constituted by nanoscale protein pores or “nanopores” embedded in an electrical resistant polymer membrane where an ionic current flows. During sequencing, single strand DNA translocation through the pores induces voltage shifts in the ionic current that are characteristic of each DNA sequence occupying the pore (“squiggles”) and are then computationally interpreted as k-mers of 3-6 nucleotides in length (Jain et al., 2016). Nanopore sequencing technology presents a series of advantages: i) long sequencing reads up to hundreds kbp, with theoretically no-instrument imposed size limitation (Jain et al., 2016), ii) real time data analysis coupled with no fixed run time up to max 72 h, with the possibility of interrupting the sequencing when a certain datum has been seen a certain number of times at a specified confidence level (“read until”), so that in this way the experiment is defined by the user and not by the machine (Loose et al., 2016), iii) reduced costs compared to other sequencing technologies (Q.-F. Zhang et al., 2020), iv) portability and readiness of use with different types of devices and DNA library preparation protocols suitable for different situations and conditions, also for use in non-laboratory settings (Castro-Wallace et al., 2017; Goordial et al., 2017; Pomerantz et al., 2018). Despite the numerous advantages of Nanopore sequencing technology, raw Nanopore reads are still characterized by a relatively high-error rate. Systematic random errors account for 5% up to 15% of the total sequenced bases and consist in insertions/deletions (InDels) of bases at the level of DNA homopolymer tracts or in a minority of cases in nucleotide substitution, due to variations in the DNA translocation speed and to chemical modifications that alter the electric signal, respectively (De Maio et al., 2019). Nanopore sequencing errors are associated to changes in the protein annotation, due to the introduction of premature stop codon or frameshift error in the DNA sequence resulting in incorrect shorter or longer predicted coding sequences. The low Nanopore accuracy may be resolved using either non-hybrid methods or hybrid approaches. Non-hybrid methods include bioinformatic tools that perform self-correction with long reads alone, using the overlap information to generate a consensus sequence and consequently requiring that a sufficient genome coverage has been generated during sequencing (Loman et al., 2015). Hybrid approaches involve bioinformatic tools that implies short high accuracy reads (e.g. Illumina reads) for single base pair correction and long Nanopore reads just for the resolution of genomic architecture (De Maio et al., 2019). Hybrid methods aided by short accurate reads achieve better correction quality, especially when handling low coverage-depth long reads compared to non-hybrid methods and therefore, are currently the preferred approach for accurate genome assembly (De Maio et al., 2019; Madoui et al., 2015; Ruan et al., 2020). Due to its characteristics, Nanopore sequencing technology has been used in microbiology for the *de novo* genome assembly of complete bacterial genomes (Karlsson et al., 2015; Laver et al., 2015), for rapid and accurate pathogen identification (Bialasiewicz et al., 2019; Charalampous et al., 2018; Cusco et al., 2018; Sanderson et al., 2018) and for antimicrobial resistance pathogens profiling (Bainomugisa et al., 2018; Lemon et al., 2017; Pitt et al., 2020; Schmidt et al., 2017; Tamma et al., 2019). Long Nanopore reads have been proven to be particularly useful for *de novo* genome assembly because hold the potential to uniquely span repeated regions of a bacterial genome by anchoring both extremities of each repeat to the flanking sequences (Koren et al., 2013; Loman et al., 2015).

2.2. Genome annotation and comparison tools

Genome annotation consists of a i) gene finding process aimed to predict the section of the genome that contain genes and a ii) function assignment step seeking to predict the function of the coded proteins by sequence similarity across various sequence databases. Bioinformatic annotation pipelines include local annotation pipelines which can be downloaded and run on local computers, such as Prokka (Seemann, 2014) and web based platforms which require users to upload their unannotated genomes to a given server, such as RAST (Overbeek et al., 2014), IMG (Markowitz et al., 2012) and NCBI (Tatusova et al., 2016). A set of annotated genomes allows gene clustering into ortholog families, followed by analysis of the presence/absence of each family in the given genomes. The term core genome refers to the set of distinct families observed in all genomes, the term variable or accessory genome to gene families not included in the core genome, whereas the term pan genome refers to the totality of gene families observed in the genomes being compared (Tettelin et al., 2005; Vernikos et al., 2015). Biologically the core genome identifies the gene families that can be found in the members of a group, whereas the variable genome represents gene families of subgroups which can potentially be associated to different phenotypic tracts. Pan genome can be open if the addition of new genomes increases the pan genome size, or closed if after a certain number of genomes, the addition of a new one will not increase the size of its pan genome. Pan genome analysis programs are Get_Homologues (Contreras-Moreira & Vinuesa, 2013) , which uses 3rd party programs for computing orthologous group and Roary (Page et al., 2015), which instead does not depend on 3rd party. Whole genome alignment can't be obtained with any sequence alignment algorithm due to the length of genomes. Pairwise genome alignment programs rely on a seed-and-extend method which first find short alignments and then connect those close enough such as the MUMmer program (Kurtz et al., 2004) and YOC program (Uricaru et al., 2015). Variation of the seed-and-extend method are also used by program for multiple genome alignment such as MAUVE (Darling et al., 2004). Genome alignment is useful to evaluate the sinteny, which is the conservation of the genes order that may suggest functional evolutionary constraints. However, despite all the technological advancements the main obstacle to genome comparison analysis remains the presence of deposited prokaryotic genomes in the form of draft genomes which may contain incorrectly assembled contigs.

3. Aim of the thesis

The aim of this thesis was to study strains of three bacterial species part of the group of lactic acid bacteria by means of whole genome sequencing and comparative genomic analysis. In particular, the thesis focused on *L. crispatus* probiotic strain M247 and type strain ATCC 33820, *S. pneumoniae* laboratory strains Rx1 and R36A, and *E. faecalis* clinical isolates retrieved from infertile couples. To obtain a complete genome for each strain analyzed, sequencing approach was based on Nanopore sequencing technology, accompanied by Illumina sequencing for better sequence quality. *L. crispatus* probiotic strain M247 genome was investigated using bioinformatic tools, complemented with PCR analysis (Chapters 2 and 3). A collection of infertility-associated *E. faecalis* clinical isolates was characterized by using antimicrobial susceptibility testing, whole genome sequencing and multilocus sequence typing to investigate the presence and features of antimicrobial resistance determinants (Chapters 4 and 5). Finally, the genome sequences of *L. crispatus* type strain ATCC 33820 and *S. pneumoniae* strains Rx1 and R36A were obtained and analyzed (Chapters 6 and 7).

4. References

Abdelmaksoud, A. A., Koparde, V. N., Sheth, N. U., Serrano, M. G., Glascock, A. L., Fettweis, J. M., Strauss, J. F., Buck, G. A., & Jefferson, K. K. (2016). Comparison of *Lactobacillus crispatus* isolates from *Lactobacillus*-dominated vaginal microbiomes with isolates from microbiomes containing bacterial vaginosis-associated bacteria. *Microbiology*, *162*(3), 466–475. https://doi.org/10.1099/mic.0.000238.

Abranches, J., Zeng, L., Kajfasz, J. K., Palmer, S. R., Chakraborty, B., Wen, Z. T., Richards, V. P., Brady, L. J., & Lemos, J. A. (2018). Biology of oral streptococci. *Microbiology Spectrum*, *6*(5). https://doi.org/10.1128/microbiolspec.GPP3-0042-201.

Agudelo Higuita, N. I., & Huycke, M. M. (2014). Enterococcal disease, epidemiology, and implications for treatment. In M. S. Gilmore, D. B. Clewell, Y. Ike, & N. Shankar (Eds.), *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection*. Massachusetts Eye and Ear Infirmary. http://www.ncbi.nlm.nih.gov/books/NBK190429/.

Ahrné, Nobaek, Jeppsson, Adlerberth, Wold, & Molin. (1998). The normal *Lactobacillus* flora of healthy human rectal and oral mucosa. *Journal of Applied Microbiology*, *85*(1), 88–94. https://doi.org/10.1046/j.1365-2672.1998.00480.

Aldunate, M., Srbinovski, D., Hearps, A. C., Latham, C. F., Ramsland, P. A., Gugasyan, R., Cone, R. A., & Tachedjian, G. (2015). Antimicrobial and immune modulatory effects of lactic acid and short chain fatty acids produced by vaginal microbiota associated with eubiosis and bacterial vaginosis. *Frontiers in Physiology*, *6*. https://doi.org/10.3389/fphys.2015.00164.

Bainomugisa, A., Duarte, T., Lavu, E., Pandey, S., Coulter, C., Marais, B. J., & Coin, L. M. (2018). A complete high-quality MinION Nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. *Microbial Genomics*, *4*(7). https://doi.org/10.1099/mgen.0.000188.

Berry, A. M., & Paton, J. C. (2000). Additive attenuation of virulence of *Streptococcus pneumoniae* by mutation of the genes encoding pneumolysin and other putative pneumococcal virulence proteins. *Infection and Immunity*, *68*(1), 133–140. https://doi.org/10.1128/IAI.68.1.133-140.2000.

Bialasiewicz, S., Duarte, T. P. S., Nguyen, S. H., Sukumaran, V., Stewart, A., Appleton, S., Pitt, M. E., Bainomugisa, A., Jennison, A. V., Graham, R., Coin, L. J. M., & Hajkowicz, K. (2019). Rapid diagnosis of *Capnocytophaga canimorsus* septic shock in an immunocompetent individual using real-time Nanopore sequencing: A case report. *BMC Infectious Diseases*, *19*(1), 660. https://doi.org/10.1186/s12879-019-4173-2.

Bourgogne, A., Garsin, D. A., Qin, X., Singh, K. V., Sillanpaa, J., Yerrapragada, S., Ding, Y., Dugan-Rocha, S., Buhay, C., Shen, H., Chen, G., Williams, G., Muzny, D., Maadani, A., Fox, K. A., Gioia, J., Chen, L., Shang, Y., Arias, C. A., Weinstock, G. M. (2008). Large scale variation in *Enterococcus faecalis* illustrated by the genome analysis of strain OG1RF. *Genome Biology*, *9*(7), R110. https://doi.org/10.1186/gb-2008-9-7-r110.

Brown, A. O., Millett, E. R. C., Quint, J. K., & Orihuela, C. J. (2015). Cardiotoxicity during invasive pneumococcal disease. *American Journal of Respiratory and Critical Care Medicine*, *191*(7), 739–745. https://doi.org/10.1164/rccm.201411-1951PP.

Bryant, J., Chewapreecha, C., & Bentley, S. D. (2012). Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiology*, *7*(11), 1283–1296. https://doi.org/10.2217/fmb.12.108.

Brygoo, E. R., & Aladame, N. (1953). Study of a new strictly anaerobic species of the genus *Eubacterium: Eubacterium crispatum* n. Sp.. *Annales de l'Institut Pasteur*, *84*(3), 640–641.

Byappanahalli, M. N., Nevers, M. B., Korajkic, A., Staley, Z. R., & Harwood, V. J. (2012). Enterococci in the environment. *Microbiology and Molecular Biology Reviews*, *76*(4), 685–706. https://doi.org/10.1128/MMBR.00023-12.

Carr, F. J., Chill, D., & Maida, N. (2002). The lactic acid bacteria: a literature survey. *Critical Reviews in Microbiology*, *28*(4), 281–370. https://doi.org/10.1080/1040-840291046759.

Casjens, S. (1998). The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics*, *32*(1), 339–377. https://doi.org/10.1146/annurev.genet.32.1.339.

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., Dworkin, J. P., Lupisella, M. L., Smith, D. J., Botkin, D. J., Stephenson, T. A., Juul, S., Turner, D. J., Izquierdo, F., Federman, S., Stryke, D., Somasekar, S., Alexander, N., Yu, G., … Burton, A. S. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Scientific Reports*, *7*(1), 18022. https://doi.org/10.1038/s41598-017-18364-0.

Cato, E. P., Moore, W. E. C., & Johnson, J. L. (1983). Synonymy of strains of "*Lactobacillus acidophilus*" group A2 (Johnson et al. 1980) with the type strain of *Lactobacillus crispatus* (Brygoo and Aladame 1953) Moore and Holdeman 1970. In *International Journal of Systematic and Evolutionary Microbiology,* (Vol. 33, Issue 2, pp. 426–428). Microbiology Society.

Charalampous, T., Richardson, H., Kay, G. L., Baldan, R., Jeanes, C., Rae, D., Grundy, S., Turner, D. J., Wain, J., Leggett, R. M., Livermore, D. M., & O'Grady, J. (2018). Rapid diagnosis of lower respiratory infection using Nanopore-based clinical metagenomics [Preprint]. *Microbiology*. https://doi.org/10.1101/387548.

Clewell, D. B. (2007). Properties of *Enterococcus faecalis* plasmid pAD1, a member of a widely disseminated family of pheromone-responding, conjugative, virulence elements encoding cytolysin. *Plasmid*, *58*(3), 205–227. https://doi.org/10.1016/j.plasmid.2007.05.001.

Coffey, T. J., Enright, M. C., Daniels, M., Morona, J. K., Morona, R., Hryniewicz, W., Paton, J. C., & Spratt, B. G. (1998). Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of

Streptococcus pneumoniae. *Molecular Microbiology*, *27*(1), 73–83. https://doi.org/10.1046/j.1365-2958.1998.00658.

Collins, F. W. J., O'Connor, P. M., O'Sullivan, O., Gómez-Sala, B., Rea, M. C., Hill, C., & Ross, R. P. (2017). Bacteriocin Gene-Trait matching across the complete *Lactobacillus* Pan-genome. *Scientific Reports*, *7*(1), 3481. https://doi.org/10.1038/s41598-017-03339-y.

Contreras-Moreira, B., & Vinuesa, P. (2013). GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology*, *79*(24), 7696–7701. https://doi.org/10.1128/AEM.02411-13.

Croucher, N. J., Harris, S. R., Fraser, C., Quail, M. A., Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J. H., Ko, K. S., Pichon, B., Baker, S., Parry, C. M., Lambertsen, L. M., Shahinas, D., Pillai, D. R., Mitchell, T. J., Dougan, G., Tomasz, A., Bentley, S. D. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science*, *331*(6016), 430–434. https://doi.org/10.1126/science.1198545.

Croucher, N. J., Walker, D., Romero, P., Lennard, N., Paterson, G. K., Bason, N. C., Mitchell, A. M., Quail, M. A., Andrew, P. W., Parkhill, J., Bentley, S. D., & Mitchell, T. J. (2009). Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* $^{\text{Spain23F}}$ ST81. *Journal of Bacteriology*, *191*(5), 1480–1489. https://doi.org/10.1128/JB.01343-08.

Cusco, A., Catozzi, C., Vines, J., Sanchez, A., & Francino, O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon [Preprint]. *Microbiology*. https://doi.org/10.1101/450734.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, *14*(7), 1394–1403. https://doi.org/10.1101/gr.2289704.

De Maio, N., Shaw, L. P., Hubbard, A., George, S., Sanderson, N. D., Swann, J., Wick, R., AbuOun, M., Stubberfield, E., Hoosdally, S. J., Crook, D. W., Peto, T. E. A., Sheppard, A.

E., Bailey, M. J., Read, D. S., Anjum, M. F., Walker, A. S., Stoesser, N., & on behalf of the REHAB consortium. (2019). Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*, *5*(9). https://doi.org/10.1099/mgen.0.000294.

Dec, M., Nowaczek, A., Stępień-Pyśniak, D., Wawrzykowski, J., & Urban-Chmiel, R. (2018). Identification and antibiotic susceptibility of lactobacilli isolated from turkeys. *BMC Microbiology*, *18*(1), 168. https://doi.org/10.1186/s12866-018-1269-6.

Desmarais, T. R., Solo-Gabriele, H. M., & Palmer, C. J. (2002). Influence of Soil on Fecal Indicator Organisms in a tidally influenced subtropical environment. *Applied and Environmental Microbiology*, *68*(3), 1165–1172. https://doi.org/10.1128/AEM.68.3.1165-1172.2002.

Duerkop, B. A., Huo, W., Bhardwaj, P., Palmer, K. L., & Hooper, L. V. (2016). Molecular basis for lytic bacteriophage resistance in Enterococci. *MBio*, *7*(4). https://doi.org/10.1128/mBio.01304-16.

Dunny, G. M. (2007). The peptide pheromone-inducible conjugation system of *Enterococcus faecalis* plasmid pCF10: Cell–cell signalling, gene transfer, complexity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1483), 1185–1193. https://doi.org/10.1098/rstb.2007.2043.

El Aila, N. A., Tency, I., Claeys, G., Verstraelen, H., Saerens, B., Lopes dos Santos Santiago, G., De Backer, E., Cools, P., Temmerman, M., Verhelst, R., & Vaneechoutte, M. (2009). Identification and genotyping of bacteria from paired vaginal and rectal samples from pregnant women indicates similarity between vaginal and rectal microflora. *BMC Infectious Diseases*, *9*(1), 167. https://doi.org/10.1186/1471-2334-9-167

Fernández-Hidalgo, N., Escolà-Vergé, L., & Pericàs, J. M. (2020). *Enterococcus faecalis* endocarditis: What’s next? *Future Microbiology*, *15*, 349–364. https://doi.org/10.2217/fmb-2019-0247

Flores-Mireles, A. L., Walker, J. N., Caparon, M., & Hultgren, S. J. (2015). Urinary tract infections: Epidemiology, mechanisms of infection and treatment options. *Nature Reviews. Microbiology*, *13*(5), 269–284. https://doi.org/10.1038/nrmicro3432.

Franke, A. E., & Clewell, D. B. (1981). Evidence for a chromosome-borne resistance transposon (Tn*916*) in *Streptococcus faecalis* that is capable of 'conjugal' transfer in the absence of a conjugative plasmid. *Journal of Bacteriology*, *145*(1), 494–502. https://doi.org/10.1128/jb.145.1.494-502.1981.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., … Venter, J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, *270*(5235), 397–404. https://doi.org/10.1126/science.270.5235.397.

Ganaie, F., Saad, J. S., McGee, L., van Tonder, A. J., Bentley, S. D., Lo, S. W., Gladstone, R. A., Turner, P., Keenan, J. D., Breiman, R. F., & Nahm, M. H. (2020). A new pneumococcal capsule type, 10d, is the 100th serotype and has a large *cps* fragment from an oral *Streptococcus*. *MBio*, *11*(3). https://doi.org/10.1128/mBio.00937-20.

García, E., Arrecubieta, C., Muñoz, R., Mollerach, M., & López, R. (1997). A functional analysis of the *Streptococcus pneumoniae* genes involved in the synthesis of type 1 and type 3 capsular polysaccharides. *Microbial Drug Resistance*, *3*(1), 73–88. https://doi.org/10.1089/mdr.1997.3.73.

George, F., Daniel, C., Thomas, M., Singer, E., Guilbaud, A., Tessier, F. J., Revol-Junelles, A.-M., Borges, F., & Foligné, B. (2018). Occurrence and dynamism of lactic acid bacteria in distinct ecological niches: a multifaceted functional health perspective. *Frontiers in Microbiology*, *9*, 2899. https://doi.org/10.3389/fmicb.2018.02899.

Goldman, B. S., Nierman, W. C., Kaiser, D., Slater, S. C., Durkin, A. S., Eisen, J. A., Ronning, C. M., Barbazuk, W. B., Blanchard, M., Field, C., Halling, C., Hinkle, G., Iartchuk, O., Kim,

H. S., Mackenzie, C., Madupu, R., Miller, N., Shvartsbeyn, A., Sullivan, S. A., … Kaplan, H. B. (2006). Evolution of sensory complexity recorded in a myxobacterial genome. *Proceedings of the National Academy of Sciences*, *103*(41), 15200–15205. https://doi.org/10.1073/pnas.0607335103.

Goordial, J., Altshuler, I., Hindson, K., Chan-Yam, K., Marcolefas, E., & Whyte, L. G. (2017). In Situ Field Sequencing and Life Detection in Remote (79°26′N) Canadian High Arctic Permafrost Ice Wedge Microbial Communities. *Frontiers in Microbiology*, *8*, 2594. https://doi.org/10.3389/fmicb.2017.02594.

Harmer, C. J., Pong, C. H., & Hall, R. M. (2020). Structures bounded by directly-oriented members of the IS*26* family are pseudo-compound transposons. *Plasmid*, *111*, 102530. https://doi.org/10.1016/j.plasmid.2020.102530.

Hartke, A., Giard, J.-C., Laplace, J.-M., & Auffray, Y. (1998). Survival of *Enterococcus faecalis* in an oligotrophic microcosm: changes in morphology, development of general stress resistance, and analysis of protein synthesis. *Appl. Environ. Microbiol.*, *64*, 8.

Hayek, S. A., & Ibrahim, S. A. (2013). Current limitations and challenges with lactic acid bacteria: a review. *Food and Nutrition Sciences*, *04*(11), 73–87. https://doi.org/10.4236/fns.2013.411A010.

Heeney, D. D., Gareau, M. G., & Marco, M. L. (2018). Intestinal *Lactobacillus* in health and disease, a driver or just along for the ride? *Current Opinion in Biotechnology*, *49*, 140–147. https://doi.org/10.1016/j.copbio.2017.08.004.

Hegstad, K., Mikalsen, T., Coque, T. M., Werner, G., & Sundsfjord, A. (2010). Mobile genetic elements and their contribution to the emergence of antimicrobial resistant *Enterococcus faecalis* and *Enterococcus faecium*. *Clinical Microbiology and Infection*, *16*(6), 541–554. https://doi.org/10.1111/j.1469-0691.2010.03226.

Hillier, S. L., Krohn, M. A., Rabe, L. K., Klebanoff, S. J., & Eschenbach, D. A. (1993). The Normal Vaginal Flora, H202-Producing Lactobacilli, and Bacterial Vaginosis in Pregnant Women. *Clin Infect Dis*, 16 Suppl 4:S273-81. doi: 10.1093/clinids/16.supplement_4.s273.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of Nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0.

Jett, B. D., Huycke, M. M., & Gilmore, M. S. (1994). Virulence of enterococci. *Clinical Microbiology Reviews*, *7*(4), 462–478. https://doi.org/10.1128/CMR.7.4.462.

Jones, D. (1978). Composition and differentiation of the genus *Streptococcus*. *Society for Applied Bacteriology Symposium Series*, *7*, 1–49.

Kaetzel, C. S. (2001). Polymeric Ig receptor: defender of the fort or Trojan horse? *Current Biology : CB*, *11*(1), R35-38. https://doi.org/10.1016/s0960-9822(00)00041-5.

Karlsson, E., Lärkeryd, A., Sjödin, A., Forsman, M., & Stenberg, P. (2015). Scaffolding of a bacterial genome using MinION Nanopore sequencing. *Scientific Reports*, *5*(1), 11996. https://doi.org/10.1038/srep11996.

Klein, G. (2003). Taxonomy, ecology and antibiotic resistance of enterococci from food and the gastro-intestinal tract. *International Journal of Food Microbiology*, *88*(2–3), 123–131. https://doi.org/10.1016/S0168-1605(03)00175-2.

Koren, S., Harhay, G. P., Smith, T. P., Bono, J. L., Harhay, D. M., Mcvey, S. D., Radune, D., Bergman, N. H., & Phillippy, A. M. (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology*, *14*(9), R101. https://doi.org/10.1186/gb-2013-14-9-r101.

Koren, S., & Phillippy, A. M. (2015). One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, *23*, 110–120. https://doi.org/10.1016/j.mib.2014.11.014.

Krzyściak, W., Pluskwa, K. K., Jurczak, A., & Kościelniak, D. (2013). The pathogenicity of the *Streptococcus* genus. *European Journal of Clinical Microbiology & Infectious Diseases*, *32*(11), 1361–1376. https://doi.org/10.1007/s10096-013-1914-9.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology* 5, R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Kyono, K., Hashimoto, T., Nagai, Y., & Sakuraba, Y. (2018). Analysis of endometrial microbiota by 16S ribosomal RNA gene sequencing among infertile patients: a single-center pilot study. *Reproductive Medicine and Biology*, *17*(3), 297–306. https://doi.org/10.1002/rmb2.12105.

Lam, M. M. C., Seemann, T., Bulach, D. M., Gladman, S. L., Chen, H., Haring, V., Moore, R. J., Ballard, S., Grayson, M. L., Johnson, P. D. R., Howden, B. P., & Stinear, T. P. (2012). Comparative analysis of the first complete *Enterococcus faecium* genome. *Journal of Bacteriology*, *194*(9), 2334–2341. https://doi.org/10.1128/JB.00259-12.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001.

Leavis, H., Top, J., Shankar, N., Borgen, K., Bonten, M., van Embden, J., & Willems, R. J. L. (2004). A novel putative enterococcal pathogenicity island linked to the *esp* virulence gene of *Enterococcus faecium* and associated with epidemicity. *Journal of Bacteriology*, *186*(3), 672–682. https://doi.org/10.1128/JB.186.3.672-682.2004.

Lebeer, S., Vanderleyden, J., & De Keersmaecker, S. C. J. (2008). Genes and molecules of Lactobacilli supporting probiotic action. *Microbiology and Molecular Biology Reviews*, *72*(4), 728–764. https://doi.org/10.1128/MMBR.00017-08.

Lebreton, F., van Schaik, W., Manson McGuire, A., Godfrey, P., Griggs, A., Mazumdar, V., Corander, J., Cheng, L., Saif, S., Young, S., Zeng, Q., Wortman, J., Birren, B., Willems, R. J. L., Earl, A. M., & Gilmore, M. S. (2013). Emergence of epidemic multidrug-resistant *Enterococcus faecium* from animal and commensal strains. *MBio*, *4*(4). https://doi.org/10.1128/mBio.00534-13.

Lebreton, F., Willems, R. J. L., & Gilmore, M. S. (2014). Enterococcus Diversity, Origins in Nature, and Gut Colonization. In M. S. Gilmore, D. B. Clewell, Y. Ike, & N. Shankar (Eds.), Enterococci: from commensals to leading causes of drug resistant infection. *Massachusetts Eye and Ear Infirmary*. http://www.ncbi.nlm.nih.gov/books/NBK190427/.

Lemon, J. K., Khil, P. P., Frank, K. M., & Dekker, J. P. (2017). Rapid Nanopore sequencing of plasmids and resistance gene detection in clinical isolates. *Journal of Clinical Microbiology*, *55*(12), 3530–3543. https://doi.org/10.1128/JCM.01069-17.

Liu, W., Pang, H., Zhang, H., & Cai, Y. (2014). Biodiversity of lactic acid bacteria. In H. Zhang & Y. Cai (Eds.), *Lactic Acid Bacteria* (pp. 103–203). Springer Netherlands. https://doi.org/10.1007/978-94-017-8841-0_2.

Llull, D., Muñoz, R., López, R., & García, E. (1999). a single gene (*tts*) located outside the *cap* locus directs the formation of *Streptococcus pneumoniae* type 37 capsular polysaccharide: type 37 pneumococci are natural, genetically binary strains. *J Exp Med, 190*(2): 241-251. doi:10.108/jem.190.2.241.

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled *de novo* using only Nanopore sequencing data. *Nature Methods*, *12*(8), 733–735. https://doi.org/10.1038/nmeth.3444.

Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using Nanopore technology. *Nature Methods*, *13*(9), 751–754. https://doi.org/10.1038/nmeth.3930.

Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., & Aury, J.-M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC Genomics*, *16*(1), 327. https://doi.org/10.1186/s12864-015-1519-z.

Maraccini, P. A., Ferguson, D. M., & Boehm, A. B. (2012). Diurnal variation in *Enterococcus* species composition in polluted ocean water and a potential role for the enterococcal carotenoid in protection against photoinactivation. *Applied and Environmental Microbiology*, *78*(2), 6.

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., & Kyrpides, N. C. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(D1), D115–D122. https://doi.org/10.1093/nar/gkr1044.

Martinson, V. G., Danforth, B. N., Minckley, R. L., Rueppell, O., Tingek, S., & Moran, N. A. (2011). A simple and distinctive microbiota associated with honey bees and bumble bees: the microbiota of honey bees and bumble bees. *Molecular Ecology*, *20*(3), 619–628. https://doi.org/10.1111/j.1365-294X.2010.04959.

McBride, S. M., Fischetti, V. A., LeBlanc, D. J., Moellering, R. C., & Gilmore, M. S. (2007). Genetic diversity among *Enterococcus faecalis*. *PLoS ONE*, *2*(7), e582. https://doi.org/10.1371/journal.pone.0000582.

Mendes-Soares, H., Suzuki, H., Hickey, R. J., & Forney, L. J. (2014). Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment. *Journal of Bacteriology*, *196*(7), 1458–1470. https://doi.org/10.1128/JB.01439-13.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001.

Moreno, I., Codoñer, F. M., Vilella, F., Valbuena, D., Martinez-Blanch, J. F., Jimenez-Almazán, J., Alonso, R., Alamá, P., Remohí, J., Pellicer, A., Ramon, D., & Simon, C. (2016).

Evidence that the endometrial microbiota has an effect on implantation success or failure. *American Journal of Obstetrics and Gynecology*, *215*(6), 684–703. https://doi.org/10.1016/j.ajog.2016.09.075.

Murray, B. E. (1990). The life and times of the *Enterococcus*. *Clinical Microbiology Reviews*, *3*(1), 46–65. https://doi.org/10.1128/cmr.3.1.46.

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*(3), 157–167. https://doi.org/10.1038/nrg3367.

Nelson, A. L., Roche, A. M., Gould, J. M., Chim, K., Ratner, A. J., & Weiser, J. N. (2007). Capsule enhances pneumococcal colonization by limiting mucus-mediated clearance. *Infection and Immunity*, *75*(1), 83–90. https://doi.org/10.1128/IAI.01475-06.

Niven, C. F., & Sherman, J. M. (1944). Nutrition of the Enterococci. *Journal of Bacteriology*, *47*(4), 335–342. https://doi.org/10.1128/jb.47.4.335-342.1944.

Ojala, T., Kankainen, M., Castro, J., Cerca, N., Edelman, S., Westerlund-Wikström, B., Paulin, L., Holm, L., & Auvinen, P. (2014). Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. *BMC Genomics*, *15*(1), 1070. https://doi.org/10.1186/1471-2164-15-1070.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(D1), D206–D214. https://doi.org/10.1093/nar/gkt1226.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

Palmer, K. L., Godfrey, P., Griggs, A., Kos, V. N., Zucker, J., Desjardins, C., Cerqueira, G., Gevers, D., Walker, S., Wortman, J., Feldgarden, M., Haas, B., Birren, B., & Gilmore, M. S. (2012). Comparative genomics of Enterococci: variation in *Enterococcus faecalis*, clade structure in *E. faecium*, and defining characteristics of *E . gallinarum* and *E . casseliflavus*. *MBio*, *3*(1). https://doi.org/10.1128/mBio.00318-11.

Palmer, K. L., Kos, V. N., & Gilmore, M. S. (2010). Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Current Opinion in Microbiology*, *13*(5), 632–639. https://doi.org/10.1016/j.mib.2010.08.004.

Pan, M., Hidalgo-Cantabrana, C., & Barrangou, R. (2020). Host and body site-specific adaptation of *Lactobacillus crispatus* genomes. *NAR Genomics and Bioinformatics*, *2*(1), lqaa001. https://doi.org/10.1093/nargab/lqaa001.

Paton, J. C., & Trappetti, C. (2019). *Streptococcus pneumoniae* capsular polysaccharide. *Microbiology Spectrum*, *7*(2). https://doi.org/10.1128/microbiolspec.GPP3-0019-2018.

Paulsen, I. T., Banerjei, L., Myers, G. S. A., Nelson, K. E., Seshadri, R., Read, T. D., Fouts, D. E., Eisen, J. A., Gill, S. R., Heidelberg, J. F., Tettelin, H., Dodson, R. J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R. T., Durkin, S., Kolonay, J., … Fraser, C. M. (2003). Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*, *299*(5615), 2071–2074. https://doi.org/10.1126/science.1080613.

Peng, D., Li, X., Liu, P., Luo, M., Chen, S., Su, K., Zhang, Z., He, Q., Qiu, J., & Li, Y. (2018). Epidemiology of pathogens and antimicrobial resistanceof catheter-associated urinary tract infections in intensive care units: A systematic review and meta-analysis. *American Journal of Infection Control*, *46*(12), e81–e90. https://doi.org/10.1016/j.ajic.2018.07.012.

Petrova, M. I., Lievens, E., Malik, S., Imholz, N., & Lebeer, S. (2015). *Lactobacillus* species as biomarkers and agents that can promote various aspects of vaginal health. *Frontiers in Physiology*, *6*. https://doi.org/10.3389/fphys.2015.00081.

Pitt, M. E., Nguyen, S. H., Duarte, T. P. S., Teng, H., Blaskovich, M. A. T., Cooper, M. A., & Coin, L. J. M. (2020). Evaluating the genome and resistome of extensively drug-resistant *Klebsiella pneumoniae* using native DNA and RNA Nanopore sequencing. *GigaScience*, *9*(2), giaa002. https://doi.org/10.1093/gigascience/giaa002.

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., Barrio-Amorós, C. L., Salazar-Valenzuela, D., & Prost, S. (2018). Real-time DNA barcoding in a rainforest using Nanopore sequencing: Opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, *7*(4). https://doi.org/10.1093/gigascience/giy033.

Pöntinen, A. K., Top, J., Arredondo-Alonso, S., Tonkin-Hill, G., Freitas, A. R., Novais, C., Gladstone, R. A., Pesonen, M., Meneses, R., Pesonen, H., Lees, J. A., Jamrozy, D., Bentley, S. D., Lanza, V. F., Torres, C., Peixe, L., Coque, T. M., Parkhill, J., Schürch, A. C., … Corander, J. (2021). Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. *Nature Communications*, *12*(1), 1523. https://doi.org/10.1038/s41467-021-21749-5.

Pop, M. (2009). Genome assembly reborn: Recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354–366. https://doi.org/10.1093/bib/bbp026.

Purnell, S. E., Ebdon, J. E., & Taylor, H. D. (2011). Bacteriophage Lysis of *Enterococcus* Host Strains: A Tool for Microbial Source Tracking? *Environmental Science & Technology*, *45*(24), 10699–10705. https://doi.org/10.1021/es202141.

Qin, X., Galloway-Peña, J. R., Sillanpaa, J., Roh, J. H., Nallapareddy, S. R., Chowdhury, S., Bourgogne, A., Choudhury, T., Muzny, D. M., Buhay, C. J., Ding, Y., Dugan-Rocha, S., Liu, W., Kovar, C., Sodergren, E., Highlander, S., Petrosino, J. F., Worley, K. C., Gibbs, R. A., … Murray, B. E. (2012). Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes. *BMC Microbiology*, *12*(1), 135. https://doi.org/10.1186/1471-2180-12-135.

Quinto, E. J., Jiménez, P., Caro, I., Tejero, J., Mateo, J., & Girbés, T. (2014). Probiotic lactic acid bacteria: a review. *Food and Nutrition Sciences*, *05*(18), 1765–1775. https://doi.org/10.4236/fns.2014.518190.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S. K., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L., & Forney, L. J. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, *108*(Supplement_1), 4680–4687. https://doi.org/10.1073/pnas.1002611107.

Richards, V. P., Palmer, S. R., Pavinski Bitar, P. D., Qin, X., Weinstock, G. M., Highlander, S. K., Town, C. D., Burne, R. A., & Stanhope, M. J. (2014). Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biology and Evolution*, *6*(4), 741–753. https://doi.org/10.1093/gbe/evu048.

Roberts, A. P., & Mullany, P. (2011). Tn *916* -like genetic elements: A diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiology Reviews*, *35*(5), 856–871. https://doi.org/10.1111/j.1574-6976.2011.00283.

Romero, P., Llull, D., García, E., Mitchell, T. J., López, R., & Moscoso, M. (2007). Isolation and characterization of a new plasmid pSpnP1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. *Plasmid*, *58*(1), 51–60. https://doi.org/10.1016/j.plasmid.2006.12.006.

Ruan, Z., Wu, J., Chen, H., Draz, M. S., Xu, J., & He, F. (2020). Hybrid genome assembly and annotation of a pandrug-resistant *Klebsiella pneumonia*e strain using Nanopore and Illumina sequencing. *Infection and Drug Resistance*, *Volume 13*, 199–206. https://doi.org/10.2147/IDR.S240404.

Ruiz-Garbajosa, P., Bonten, M. J. M., Robinson, D. A., Top, J., Nallapareddy, S. R., Torres, C., Coque, T. M., Cantón, R., Baquero, F., Murray, B. E., del Campo, R., & Willems, R. J. L. (2006). Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. *Journal of Clinical Microbiology*, *44*(6), 2220–2228. https://doi.org/10.1128/JCM.02596-05.

Ruiz-Rodríguez, L., Bleckwedel, J., Eugenia Ortiz, M., Pescuma, M., & Mozzi, F. (2016). Lactic acid bacteria. In C. Wittmann & J. C. Liao (Eds.), *Industrial Biotechnology* (pp. 395–451). Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527807796.ch11.

Sanderson, N. D., Street, T. L., Foster, D., Swann, J., Atkins, B. L., Brent, A. J., McNally, M. A., Oakley, S., Taylor, A., Peto, T. E. A., Crook, D. W., & Eyre, D. W. (2018). Real-time analysis of Nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics*, *19*(1), 714. https://doi.org/10.1186/s12864-018-5094-y.

Santer, M. (2010). Joseph Lister: first use of a bacterium as a ‘model organism’ to illustrate the cause of infectious disease of humans. *Notes and Records of the Royal Society*, *64*(1), 59–65. https://doi.org/10.1098/rsnr.2009.0029.

Schleifer, K. H., & Kilpper-Bälz, R. (1984). Transfer of *Streptococcus faecalis* and *Streptococcus faecium* to the genus *Enterococcus* nom. Rev. As *Enterococcus faecalis* comb. Nov. And *Enterococcus faecium* comb. Nov. *International Journal of Systematic and Evolutionary Microbiology, 34*(1), 31–34. https://doi.org/10.1099/00207713-34-1-31.

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J. E., Remus-Emsermann, M. N. P., & Ahrens, C. H. (2018). Pushing the limits of *de novo* genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Research*, *46*(17), 8953–8965. https://doi.org/10.1093/nar/gky726.

Schmidt, K., Mwaigwisya, S., Crossman, L. C., Doumith, M., Munroe, D., Pires, C., Khan, A. M., Woodford, N., Saunders, N. J., Wain, J., O’Grady, J., & Livermore, D. M. (2017). Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by Nanopore-based metagenomic sequencing. *Journal of Antimicrobial Chemotherapy*, *72*(1), 104–114. https://doi.org/10.1093/jac/dkw397.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Sherman JM. (1937). The Streptococci. *Bacteriological reviews*, 1(1), 3–97. https://doi.org/10.1128/br.1.1.3-97.1937.

Smith, M. D., & Guild, W. R. (1979). A plasmid in *Streptococcus pneumoniae*. *Journal of Bacteriology*, *137*(2), 735–739. https://doi.org/10.1128/jb.137.2.735-739.1979-

Smith, S. B., & Ravel, J. (2017). The vaginal microbiota, host defence and reproductive physiology: Vaginal microbiota in defence and physiology. *The Journal of Physiology*, *595*(2), 451–463. https://doi.org/10.1113/JP271694.

Solheim, M., Brekke, M. C., Snipen, L. G., Willems, R. J., Nes, I. F., & Brede, D. A. (2011). Comparative genomic analysis reveals significant enrichment of mobile genetic elements and genes encoding surface structure-proteins in hospital-associated clonal complex 2 *Enterococcus faecalis*. *BMC Microbiology*, *11*(1), 3. https://doi.org/10.1186/1471-2180-11-3.

Stern, A., Mick, E., Tirosh, I., Sagy, O., & Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Research*, *22*(10), 1985–1994. https://doi.org/10.1101/gr.138297.112.

Straume, D., Stamsås, G. A., & Håvarstein, L. S. (2015). Natural transformation and genome evolution in *Streptococcus pneumoniae*. *Infection, Genetics and Evolution*, *33*, 371–380. https://doi.org/10.1016/j.meegid.2014.10.020.

Tachedjian, G., O'Hanlon, D. E., & Ravel, J. (2018). The implausible "in vivo" role of hydrogen peroxide as an antimicrobial factor produced by vaginal microbiota. *Microbiome*, *6*(1), 29. https://doi.org/10.1186/s40168-018-0418-3.

Tamma, P. D., Fan, Y., Bergman, Y., Pertea, G., Kazmi, A. Q., Lewis, S., Carroll, K. C., Schatz, M. C., Timp, W., & Simner, P. J. (2019). Applying rapid whole-genome sequencing to predict phenotypic antimicrobial susceptibility testing results among carbapenem-resistant

Klebsiella pneumoniae clinical isolates. *Antimicrobial Agents and Chemotherapy*, *63*(1). https://doi.org/10.1128/AAC.01923-18.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569.

Tendolkar, P. M., Baghdayan, A. S., Gilmore, M. S., & Shankar, N. (2004). Enterococcal surface protein, esp, enhances biofilm formation by *Enterococcus faecalis*. *Infection and Immunity*, *72*(10), 6032–6039. https://doi.org/10.1128/IAI.72.10.6032-6039.2004.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102.

Thiercelin, M. E., & Jouhaud, L. (1899). Sur un diplococque saprophyte de l'intestin susceptible de devenir pathogene. *Sur Un Diplococque Saprophyte de l'intestin Susceptible de Devenir Pathogene.*, 269–27.

Tomasz, A., Albino, A., & Zanati, E. (1970). Multiple antibiotic resistance in a bacterium with suppressed autolytic system. *Nature*, *227*(5254), 138–140. https://doi.org/10.1038/227138a0.

Treangen, T. J., Abraham, A.-L., Touchon, M., & Rocha, E. P. C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews*, *33*(3), 539–571. https://doi.org/10.1111/j.1574-6976.2009.00169.

Tu, A.-H. T., Fulgham, R. L., McCrory, M. A., Briles, D. E., & Szalai, A. J. (1999). Pneumococcal surface protein A inhibits complement activation by *Streptococcus pneumoniae*. *Infection and Immunity*, *67*(9), 4720–4724. https://doi.org/10.1128/IAI.67.9.4720-4724.

Tyssen, D., Wang, Y.-Y., Hayward, J. A., Agius, P. A., DeLong, K., Aldunate, M., Ravel, J., Moench, T. R., Cone, R. A., & Tachedjian, G. (2018). Anti-HIV-1 activity of lactic acid in human cervicovaginal fluid. *MSphere*, *3*(4). https://doi.org/10.1128/mSphere.00055-18.

Uricaru, R., Michotey, C., Chiapello, H., & Rivals, E. (2015). YOC, A new strategy for pairwise alignment of collinear genomes. *BMC Bioinformatics*, *16*(1), 111. https://doi.org/10.1186/s12859-015-0530-3.

van der Veer, C., Hertzberger, R. Y., Bruisten, S. M., Tytgat, H. L. P., Swanenburg, J., de Kat Angelino-Bart, A., Schuren, F., Molenaar, D., Reid, G., de Vries, H., & Kort, R. (2019). Comparative genomics of human Lactobacillus crispatus isolates reveals genes for glycosylation and glycogen degradation: Implications for in vivo dominance of the vaginal microbiota. *Microbiome*, *7*(1), 49. https://doi.org/10.1186/s40168-019-0667-9.

Van Tyne, D., & Gilmore, M. S. (2014). Friend turned foe: evolution of enterococcal virulence and antibiotic resistance. *Annual Review of Microbiology*, *68*(1), 337–356. https://doi.org/10.1146/annurev-micro-091213-113003.

Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, *23*, 148–154. https://doi.org/10.1016/j.mib.2014.11.016.

Weaver, K. E. (2019). *Enterococcal Genetics*. *Microbiol Spectr*, 7(2). doi: 10.1128/microbiolspec.GPP3-0055-2018.

Wei, S., Morrison, M., & Yu, Z. (2013). Bacterial census of poultry intestinal microbiome. *Poultry Science*, *92*(3), 671–683. https://doi.org/10.3382/ps.2012-02822.

Weiser, J. N., Ferreira, D. M., & Paton, J. C. (2018). *Streptococcus pneumoniae*: Transmission, colonization and invasion. *Nature Reviews Microbiology*, *16*(6), 355–367. https://doi.org/10.1038/s41579-018-0001-8.

Werner, G., Coque, T. M., Franz, C. M. A. P., Grohmann, E., Hegstad, K., Jensen, L., van Schaik, W., & Weaver, K. (2013). Antibiotic resistant enterococci—tales of a drug resistance gene trafficker. *International Journal of Medical Microbiology*, *303*(6–7), 360–379. https://doi.org/10.1016/j.ijmm.2013.03.001.

Woese, C. (1998). The universal ancestor. *Proceedings of the National Academy of Sciences*, *95*(12), 6854–6859. https://doi.org/10.1073/pnas.95.12.6854.

Wood, B. J. B., & Holzapfel, W. H. (Eds.). (1995). The genera of lactic acid bacteria. *Springer US*. https://doi.org/10.1007/978-1-4615-5817-0.

Yamahara, K. M., Walters, S. P., & Boehm, A. B. (2009). Growth of enterococci in unaltered, unseeded beach sands subjected to tidal wetting. *Applied and Environmental Microbiology*, *75*(6), 1517–1524. https://doi.org/10.1128/AEM.02278-08.

Zacharof, M. P., & Lovitt, R. W. (2012). Bacteriocins produced by lactic acid bacteria a review article. *APCBEE Procedia*, *2*, 50–56. https://doi.org/10.1016/j.apcbee.2012.06.010.

Zhang, H., & Cai, Y. (Eds.). (2014). Lactic acid bacteria. *Springer Netherlands.* https://doi.org/10.1007/978-94-017-8841-0.

Zhang, Q.-F., Zhang, Y.-J., Wang, S., Wei, Y., Li, F., & Feng, K.-J. (2020). The effect of screening and treatment of *Ureaplasma urealyticum* infection on semen parameters in asymptomatic leukocytospermia: A case-control study. *BMC Urology*, *20*(1), 165. https://doi.org/10.1186/s12894-020-00742-y.

Zheng, J., Wittouck, S., Salvetti, E., Franz, C. M. A. P., Harris, H. M. B., Mattarelli, P., O'Toole, P. W., Pot, B., Vandamme, P., Walter, J., Watanabe, K., Wuyts, S., Felis, G. E., Gänzle, M. G., & Lebeer, S. (2020). A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus Beijerinck* 1901, and union of Lactobacillaceae and Leuconostocaceae. *International Journal of Systematic and Evolutionary Microbiology*, *70*(4), 2782–2858. https://doi.org/10.1099/ijsem.0.004107.

CHAPTER 2

The mobilome of probiotic *Lactobacillus crispatus* M247 includes Tn*7088* a novel transposon carrying a biosynthetic gene cluster for a class I bacteriocin

Lorenzo Colombini[1], Francesco Santoro[1], Lorenzo Morelli[2], Francesco Iannelli[1]
and Gianni Pozzi[1]

Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy

Università Cattolica del Sacro Cuore, Department of Food Science and Technologies for a Sustainable Agri-food Supply Chain (DiSTAS), University of Piacenza, 53100 Piacenza, Italy

Manuscript in preparation

1. ABSTRACT

Background: The probiotic *Lactobacillus crispatus* strain M247 is known to exhibit beneficial effects on intestinal inflammatory disorders, strong aggregation phenotype and adherence to intestinal mucus as well as counteracting effects on vaginal dysbiosis and on papilloma virus infections. In this study, the *L. crispatus* M247 complete genome sequence was obtained and analyzed, resulting in the identification and characterization of a novel mobile genetic element carrying a biosynthetic gene cluster for a class I bacteriocin.

Methods: The complete genome sequence of *L. crispatus* M247 was obtained combining Nanopore and Illumina sequencing technologies. M247 genomic features and its mobilome were evaluated with bioinformatic tools. The DNA sequence of a novel mobile genetic element was analyzed. PCR mapping was performed to evaluate the excision mechanism, and quantitative PCR was used to quantify the number of circular intermediates and reconstituted chromosomal integration sites.

Results: The M247 genome consists of a 2.33 Mb circular chromosome, with 2,305 open reading frames (ORFs) and a GC content of 37.04%. The M247 mobilome accounts for 13.6% of the whole genome, including a 42.6-kb long prophage, a 14.1-kb long novel integrative and mobilizable element named Tn*7088*, and various insertion sequences (ISs). Tn*7088* integrates at a 79-bp long *att*B site on the M247 chromosome containing the last 12 nucleotides at the 3' end of a threonine tRNA encoding gene, and upon integration is flanked by *att*L and *att*R. It was shown that Tn*7088* it is able to excise from the M247 chromosome, with consequent reconstitution of the integration site *att*B identical to *att*R, and to form circular intermediates where the left and right ends are joined by *att*Tn identical to *att*L. *att*L-*att*Tn contain 12 nucleotide changes and 11 nucleotides insertion compared to *att*R-*att*B. Tn*7088* contains 18 ORFs, of which 15 ORFs code for hypothetical proteins with a homology-based predicted functions including i) genes coding for proteins involved in the integration/excision, ii) genes coding for proteins contributing to the putative horizontal transfer of the element and iii) a gene cluster homologous to the listeriolysin S locus of *Listeria monocytogenes*, coding for a class I bacteriocin and enzymes involved in its production. The *att*B site of Tn*7088* was found also in other *Lactobacillus* species sharing a core sequence of 12 nucleotides. Tn*7088*-like elements were found integrated in 7 out of 14 *L. crispatus* complete genomes, with certain variabilities within the bacteriocin gene cluster.

Conclusion: Our work reports the characterization of the novel mobile genetic element Tn*7088*, identified in the genome of the probiotic *L. crispatus* strain M247, which integrates at an *att*B site present in the chromosomes of *L. crispatus* strains and other *Lactobacillus* species. Tn*7088* contains a class I bacteriocin biosynthetic gene cluster homologous of the listeriolysin S gene cluster of *Listeria monocytogenes* suggesting that this element may contribute to the niche-adaptive traits and to the probiotic potential of its host bacterial strain.

2. INTRODUCTION

Lactobacillus crispatus is the most frequently isolated species among the vaginal lactobacilli of the human microbiota of healthy women and it is also one of the commensal bacteria of the human gastrointestinal tract (Petrova et al., 2015; Walter, 2008). *L. crispatus* genomic content varies and correlates with the isolation source. Differences among strains involve mainly CRISPR-cas systems, metabolism genes, exopolysaccharides (EPS)-production and prophages (Ojala et al., 2014; Pan et al., 2020). Fundamental differences in the genetic content can translate to a better performance of specific strains in a particular ecological niche, increasing survival, colonization and functionalities. The production of antimicrobial compounds and EPS are *L. crispatus* traits of interest in the design and formulation of probiotics for host and body site-specific applications (Donnarumma et al., 2014; Nardini et al., 2016). Among antimicrobial compounds, bacteriocins represent a large family of ribosomally produced peptide antibiotics that increase the fitness of individual bacterial strains in competition with other microorganisms or with host defense mechanisms, playing an important role in shaping the microbiome (Heilbronner et al., 2021). Bacteriocins may represent alternatives to antibiotics due to i) their potency both in vitro and in vivo paired with low toxicity, ii) the specific spectrum of activity, iii) the possibility to be bioengineered and to be produced *in situ* by probiotics (Cotter et al., 2013). Based on the occurrence of post-translational modification, bacteriocins are distinguished in class II (peptides which remain unaltered after synthesis) or class I (peptide which are modified by enzymatic tailoring). M247 strain is a *L. crispatus* newborn fecal sample isolate (Cesena et al., 2001) largely studied for its probiotic potential. It was demonstrated that strain M247 exhibits beneficial effects on intestinal inflammatory disorders (Castagliuolo et al., 2005; Voltan et al., 2007, 2008), shows strong aggregation phenotype and adherence to intestinal mucus (Cesena et al., 2001; Hynönen & Palva, 2013; Kirjavainen et al., 1998; Marcotte et al., 2004; Siciliano et al., 2008), helps counteracting vaginal dysbiosis (Pierro et al., 2018) and also seems to contribute to papilloma virus clearance (Pierro et al., 2021). In this work, we report the complete genome sequence of the

L. crispatus strain M247 and we characterize the novel transposon Tn*7088*, containing a biosynthetic gene cluster for a class I bacteriocin, which may confer niche adaptive advantages to its bacterial host.

3. MATERIALS AND METHODS

3.1. Bacterial strains and growth conditions

L. crispatus strain M247 isolated from feces of human newborns (Cesena et al., 2001) was used in this study. The *L. crispatus* type strain ATCC 33820 (Teodori et al., 2021) purchased from the American Type Culture Collection was also used as reference. Frozen starter cultures were grown in DeMan-Rogosa-Sharpe medium (MRS) broth (Oxoid LTD, Basingstoke, Hampshire, England) in anaerobic condition at 37°C.

3.2. DNA purification and quantification

Bacterial cells were harvested by centrifugation (5,000 x *g* for 30 minutes at 4°C) in exponential phase growth (OD_{590}=1.9). *Lactobacillus* cells pellet was dry vortex-mixed for 2-3 min and incubated for 1 hour at 37°C in Protoplasting Buffer (20% Raffinose, 50 mM Tris-HCl [pH 8.0], 5 mM EDTA) containing 4 mg/ml lysozyme. Protoplasts were centrifuged (5,000 x *g* for 5 minutes), resuspended in 15 ml of deionized H_2O with 100 µg/ml proteinase K (Merck KGaA, Darmstadt, Germany) and incubated for 30 minutes at 37°C to obtain osmotic lysis, adding 0.5% SDS after 15 minutes. Then, 0.55 M NaCl was added and the mixture was incubated for 10 minutes at room temperature. High-molecular-weight DNA was purified three times with 1 volume of chloroform-isoamyl alcohol (24:1 [v:v]), precipitated in 0.6 volumes of ice-cold isopropanol, and spooled on a glass rod. DNA was resuspended in 10-fold diluted saline-sodium citrate (SSC) 1x buffer, then adjusted to 1x SSC and maintained at 4°C. The DNA solution was homogenized using a rotator mixer. DNA was quantified with Qubit 2.0 Fluorometer (Invitrogen, Life Technologies, Carlsbad, CA, United States) by using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific) and results were confirmed by spectrophotometer measurement (Implen, Munich, Germany).

DNA integrity and size were assayed by horizontal gel electrophoresis using 0.6% Seakem LE (Lonza, Rockland, ME USA) agarose in 0.5X Tris Borate EDTA running buffer.

3.3. Illumina Whole Genome Sequencing

Illumina sequencing was performed at MicrobesNG (University of Birmingham, United Kingdom) using Nextera library preparation kit (Illumina Inc., San Diego, USA) followed by sequencing on a NovaSeq 6000 device (Illumina Inc.) (2x250 bp paired-end sequencing). Illumina reads were analyzed with NanoPlot v1.18.2 (De Coster et al., 2018). Illumina reads properties and accession numbers were reported in Supplementary Table S1.

3.4. Nanopore Whole Genome Sequencing

Sequencing reactions were carried out in 1.5 ml LoBind tubes (Sarstedt, Nümbrecht, Germany) using wide bore (∅1.2 mm) tips for DNA manipulation in order to reduce physical shearing. DNA size selection of the genomic DNA was obtained with 0.5 volume of AMPure XP beads (Beckman Coulter, Milano, Italy) according to manufacturer’s instructions. 2 µg of size-selected DNA were employed for library construction by using the SQK-LSK 108 kit (Oxford Nanopore Technologies, Oxford, United Kingdom). Pooling of multiple samples was obtained with the Nanopore “Native Barcoding Expansion 1-12 kit” (Oxford Nanopore Technologies). Library preparation was obtained following the manufacturer’s protocol with the following modifications: (i) incubation on rotator mix for 15 min; (ii) the Library Loading Beads (LLB) were not added. Finally, 1 µg of DNA library was loaded onto a R9.4 flow cell (FLO- MIN106) (Oxford Nanopore Technologies). A 21-h sequencing run was performed on a GridION device (Oxford Nanopore Technologies). Real time base calling was performed with Guppy v3.2.6 (Oxford Nanopore Technologies), filtering out reads with a quality cut off >Q7. Base called reads were analyzed with NanoPlot v1.18.2 (De Coster et al., 2018). Nanopore reads properties and accession numbers are reported in Supplementary Table S1.

3.5. Genome assembly and annotation

The overall M247 Nanopore reads were filtered to obtain a 95x coverage taking 2.3Mbp as genome size estimate by using Filtlong0.2.0 software (https://github.com/rrwick/Filtlong) with parameter --*target_bases* and assembled using Flye v2.7.1 (Kolmogorov et al., 2019). The resulting circular contig was polished with Medaka v0.7.1 software (https://github.com/Nanoporetech/medaka) using the overall Nanopore reads, followed by two polishing rounds with the Pilon v1.22 tool (Walker et al., 2014) using the Illumina reads. Assembly completeness was assessed with the Bandage v.0.8.1 tool (Wick et al., 2015), whereas assembly quality was evaluated with both Ideel (https://github.com/mw55309/ideel) and CheckM v1.1.3 tools (Parks et al., 2015). Bwa v0.7.17 (Li, 2013) and minimap2 v2.13 (Li, 2018) programs were used to align the Illumina reads and the Nanopore reads to the assembled genome, respectively. Reads genome mapping was visualized with the Tablet tool v1.17.08.17 (Milne et al., 2013) and used to further verify the assembled structure. M247 genome was automatically annotated with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v5.1 (Tatusova et al., 2016). Default parameters were used for all software unless otherwise specified.

3.6. Genome Analysis

The presence of bacterial integrative and conjugative elements (ICEs) and integrative and mobilizable elements (IMEs) in the M247 genome was investigated with ICEberg 2.0 (M. Liu et al., 2019), while insertion sequences (ISs) were detected with ISsaga (Varani et al., 2011). Integrated prophages were investigated using PHASTER (Arndt et al., 2016) and analyzed with Virfam (Lopes et al., 2014). The presence of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) was evaluated with CRISPRCasFinder (Couvin et al., 2018). Bacteriocins were tested using the *in silico* prediction tool Bagel4 (van Heel et al., 2018). Those bacteriocin predicted were further visualized with Artemis visualization tool and tested against the Bactibase bacteriocin database (Hammami et al., 2007). Antibiotic resistance genes (ARG) analysis was performed using RGI (v3.2.1) (Jia et al., 2017) with parameter “-loose_criteria=no”. Genomic sequence analysis was performed using the Basic Local Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and Artemis/ACT v17.0.1 (Carver et al., 2008).

3.7. DNA sequence analysis

Manual gene annotation of each open reading frame (ORF) was carried out by BLAST homology searching of the databases available at the National Center for Biotechnology Information (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). Protein domains were identified using the protein family database Pfam (https://pfam.xfam.org). Transposon name was assigned by the Tn Registry website curators (https://transposon.lstmed.ac.uk/tn-registry).

3.8. Nucleotide sequence accession numbers

The complete genome sequence of *L. crispatus* M247 is available under GenBank accession no. CP088015, whereas Nanopore and Illumina sequencing reads are available under Sequence Read Archive (SRA) accession no. SRR17479173 and SRR17479172, respectively.

3.9. PCR

PCR and direct PCR sequencing were carried out following an already described protocol (Iannelli et al., 1998; Santoro et al., 2010) Convergent primers designed on the chromosomal regions flanking the integration site were used to investigate the excision of each putative mobile genetic element from the chromosome, whereas divergent primers designed on the ends of each element served to evaluate its ability to form circular intermediate (CI). Oligonucleotide primers are listed in Table 1.

3.10. Quantitative Real-Time PCR

Real-time PCR experiments were carried out with the KAPA SYBR FAST qPCR kit Master Mix Universal (2X) (Merck KGaA, Darmstadt, Germany) on a LightCycler 1.5 apparatus (Roche Diagnostics GmbH, Mannheim, Germany). Real-time PCR mixture contained, in a final volume of 20 μl, 1× KAPA SYBR FAST qPCR reaction mix, 5 pmol of each primer and 2 μl (20 ng) of bacterial gDNA as starting template. Thermal profile was an initial 3 min denaturation step at 95°C followed by 40 cycles of repeated denaturation (0 s at 95°C), annealing (20 s at 62°C), and polymerization (30 s at 72°C). The temperature transition rate was 20°C/s in the denaturation and annealing step and 5°C/s in the polymerization step. A standard curve for the *gyrB* gene of *L. crispatus* M247 was built plotting the threshold cycle against the number of chromosome copies using serial dilutions of chromosomal DNA with known concentration. This external standard curve was used to quantify in each sample the number of (i) chromosome copies, (ii) CIs, and (iii) reconstituted site of integration. Phage CIs and relative reconstituted integration sites were quantified with the primer pairs IF1513/IF1514 and IF1511/IF1512, respectively (Table 1). Transposon CIs were quantified using the primer pair IF1487/IF1488, whereas quantification of the relative reconstituted integration sites was obtained with the primer pairs IF1349/IF1350 and IF1349/IF1444 for strains M247 and ATCC 33820, respectively (Table 1). Melting curve analysis was performed to differentiate the amplified products from primer dimers. Electrophoresis gel run was performed to further verified the amplification products.

4. RESULTS AND DISCUSSION

4.1. The M247 genome

The complete genome sequence of *L. crispatus* M247 was obtained combining Illumina and Nanopore sequencing technologies. Sequence analysis showed that the M247 genome is organized in one circular chromosome of 2,336,126 base pairs (bp) in length, with an average GC content of 37.04%. A schematic representation of the chromosome containing the main sequence characteristics is reported in Figure 1. The putative origin of replication corresponds to the base pair 1 of the chromosome. The genome contains 2,305 open reading frames which are distributed unequally between sense (1,140 ORFs) and antisense (1,165 ORFs) strands with a high percentage (75.84%) of ORFs predicted to be transcribed in the same direction as DNA replication. An annotation with prediction of a biological function was possible for 2,123 CDSs of which 248 CDSs code for hypothetical proteins, while 102 are pseudo genes (48 frameshifted, 28 incompletes, 11 with an internal stop, 15 with multiple problems). 12 ribosomal RNA (rRNA) genes are grouped in 3 rRNA operons, 28 out of 65 tRNA genes are not adjacent to rRNA operons, 3 structural RNAs are also present: (i) tRNA-like/mRNA-like RNA (Felden et al., 1996), (ii) signal recognition particle RNA (Lutcke, 1995) and (iii) ribonuclease P RNA (Pace & Brown, 1995). M247 genome contains three CRISPRs loci; one locus, containing 6 direct repeats (DRs) of 36 bp and 5 spacers, is associated to a type II-A CRISPR-associated (Cas) system displaying genes encoding Cas9, Cas1, Cas2 and Csn2 proteins (Makarova et al., 2011). Type II-A CRISPR have been previously described as uniquely present in *L. crispatus* human vaginal isolates with the exception of the chicken isolate C25 (Ojala et al., 2014; Pan et al., 2020). The other two loci contain 5 DRs of 23 bp with 4 spacers and 2 DRs of 25 bp, respectively, without *cas* genes. *In silico* analysis of the M247 genome sequence indicated that it contains one copy of the *apf* gene as previously described (Marcotte et al., 2004), spanning nucleotides (nt) 1,837,167 to 1,837,838 and one S-layer locus (AY941197) which spans nt 197,947 to 209,532, and includes the paralogous genes *slpA* and *slpB/csbA* (501/557 nt identity) arranged in opposite direction and located 4,624-bp apart, but no *slpC* (Fagan & Fairweather, 2014; Sun et al., 2013).

4.2. The Mobilome of M247

Bioinformatic analysis of M247 genome showed the presence of: i) a prophage, ii) a novel putative IME, that here we denominated Tn*7088* and iii) various ISs. Altogether, these mobile genetic elements account for 13.6% of the whole M247 genome (318,133 out of 2,336,126 bp).

4.2.1. The prophage

The prophage is 42,649-bp in length with an overall GC content (35.2%) lower than the average of the whole genome (37.04%). The element spans nt 1,001,016 to 1,043,664 and integrates at a 139-bp target site (*att*B) including 97 bp at the 5' end of the peptide-methionine (S)-S-oxide reductase encoding gene *msrA* (NCBI locus tag LQF73_05270). Upon integration into the M247 chromosome, the prophage is flanked by *att*L and *att*R. Using divergent PCR primers directed to the ends of the prophage we showed that the element is able to excise from the bacterial chromosome producing a circular form where the left and right ends are joined by a 138-bp sequence (*att*P) identical to *att*L. Furthermore, the reconstitution of the prophage 139-bp insertion site *att*B identical to *att*R, was demonstrated by using convergent PCR primers directed to the chromosomal flanking regions. *att*L-*att*P contain 15 nucleotide changes and 1 nucleotide deletion compared to *att*R-*att*B. In the M247 strain, circular intermediates of the phage were present at a concentration of 3.40 x 10^{-5} (± 5.26 x 10^{-6}) copies per chromosome, whereas reconstituted *att*B sites were at 2.52 x 10^{-5} (± 1.83 x 10^{-7}) copies per chromosome (Table 2). BLAST homology search identified the prophage as homologous of the phage DNA sequence named Isolate ct06w1 (GenBank accession no. BK036340) of the *Siphoviridae* family, obtained by sequencing of a human posterior fornix sample isolate during a metagenomic study of the human virome (Tisza & Buck, 2021). Manual homology-based annotation with functional prediction of the hypothetical gene product was possible for 33 out of 57 prophage ORFs, including genes coding for phage structural proteins, the terminase, the portal protein, a Clp protease, the integrase and the lytic cycle related proteins (Table 3). Compared to the deposited phage DNA sequence (BK036340), the prophage of M247 contains two additional ISs of family IS*256* corresponding to *orf*31 and *orf*51, of which the latter disrupts a tail protein encoding gene which is split in *orf*50 and *orf*52. Genomic sequence analysis revealed the presence of the prophage in 4 out of the 14 *L. crispatus* complete genomes available in NCBI sequences database (accessed in December 2021), namely strains Lc1226, PRL2021, Lc1700 and CO3MRSI1. The prophage DNA sequence was not found in any other deposited sequence, thus suggesting the specificity of the phage host-interaction and identifying the *L. crispatus* as the host bacterial species of this phage.

4.2.2. The integrative and mobilizable element Tn*7088*

The 14,184-bp long IME, namely Tn*7088*, spans nucleotides 21,914 to 36,097 and has a GC content of 30.96%. PCR analysis showed that the element excises from chromosome and produces a circular form where the left and right ends are joined by *att*Tn restoring the *att*B insertion site. *att*Tn is 90-bp long and is identical to *att*L while *att*B is 79-bp long and is identical to *att*R. *att*L-*att*Tn contain 12 nucleotide changes and 11 nucleotides insertion compared to *att*R-*att*B. *att*R-*att*B contain the last 12 nucleotides at the 3’ end of a threonine tRNA encoding gene (LQF73_00105), which are part of a 15-bp direct repeat included in the *att* sites (Supplementary Figure S1). To obtain a quantitative estimate of Tn*7088* excision from the *L. crispatus* chromosome, real-time PCR was used to quantify concentration of circular forms and reconstituted *att*B sites in extracted and purified gDNA. In the M247 strain, circular forms of Tn*7088* were present at a concentration of 3.92×10^{-5} ($\pm 2.17 \times 10^{-7}$) copies per chromosome, whereas reconstituted *att*B sites were at 1.03×10^{-4} ($\pm 3.27 \times 10^{-5}$) copies per chromosome. These values were comparable to those obtained in the *L. crispatus* type strain ATCC 33820 (Table 4).

4.2.3. Insertions sequences

The M247 genome contains 226 ISs from 15 known families, distributed as follows: IS*256* (83 copies), IS*982* (24 copies), IS*3* sub-group IS*150* (24 copies), IS*110* (21 copies), IS*30* (21 copies), IS*4* sub-group IS*Pepr1* (12 copies), IS*Lre2* (11 copies), IS*4* subgroup IS*4* (10 copies), IS*66* (8 copies), IS*200*/IS*605* sub-group IS*1341* (3 copies), IS*1182* (3 copies), IS*NCY* (2 copies), IS*200*/IS*605* (2), IS*L3* (2 copies) (Table 5). The transposase gene in 26 out of 226 ISs contains insertions or deletions producing frameshift.

4.3. The bacteriocin-encoding biosynthetic gene cluster of Tn*7088*

Nucleotide sequence analysis of Tn*7088* DNA sequence indicate that it contains 18 ORFs (Figure 2) of which o*rfs*3, 15 start with the GTG codon. *orfs*2 to 9, *orfs*10 to 15 and *orfs*16 to 17 are organized as operons each with a non-canonical promoter upstream and a rho-independent terminator downstream. Manual homology-based with functional prediction of the hypothetical gene product was possible for 15 out of 18 Tn*7088* predicted ORFs, whereas 3 ORFs encoded hypothetical proteins that showed no homology to other characterized sequences (Table 6). Predicted gene products were blasted against public protein databases and the Pfam protein family database, taking into account significant homologies with functionally characterized proteins or good matches with Pfam domains. *orf*15 and *orf*18 are ISs elements, of which the first integrated into the 5’ end of *orf*14 resulting in a loss of part of the DNA sequence (truncated *orf*14). Tn*7088* showed a typical IME modular organization (Bellanger et al., 2014) with a integration/excision module (*orfs*7, 8), a mobilization module (*orfs*3, 4, 5) and an adaptation module (*orf*9 to 17). *orf*8 codes for a site-specific integrase belonging to the family of tyrosine recombinases as found in most of IME described (Guédon et al., 2017). *orf7* contains a helix-turn-helix domain for DNA binding, thus can be speculated that codes for a recombination directionality factor, also known as excisionase, that generally helps to reverse the direction of the recombination toward the excision as described in most of IMEs encoding tyrosine integrases (Guédon et al., 2017). *orfs*3, 4 and 5 code for FtsK homologous proteins and a relaxase, respectively, which all take part to the protein-protein complex required for the hypothetical horizontal transfer of the element (Shoemaker et al., 2000). Tn*7088* contains a biosynthetic gene cluster (*orf*9 to 14 and *orfs*16, 17) for a class I bacteriocin belonging to the family of the thiazole/oxazole-modified microcin (TOMM) (Heilbronner et al., 2021). Biosynthetic gene cluster for the synthesis of thiazole and oxazole heterocycles on ribosomally produced peptide are conserved and widely distributed among prokaryotes, being found in both Gram-negative and -positive bacteria as well as in distantly related prokaryotes as cyanobacteria and archaea (S. W. Lee et al., 2008). The TOMM biosynthetic gene cluster of Tn*7088* contains 8 genes which are all homologous of genes contained in the previously described listeriolysin S (*lls*) locus of *Listeria monocytogenes* (Cotter et al., 2008; S. Lee, 2020) (Table 3). *orf*9 is homologous to the *llsA* gene and codes for the structural 44-aa length pro-bacteriocin peptide containing a 13-aa long C/S/T sequence which likely act as the target site for post-translational modifications (McAuliffe et al., 2001). *orfs*10, 11 are homologous to *llsGH* encoding ATP-binding cassette transporter that could potentially export the modified bacteriocin, whereas *orf*12 is homologous to *llsX* of unknown function which is specific of genus *Listeria* (Cotter et al., 2008; S. Lee, 2020). *orfs*13, 14, 16 are homologous to *llsBYD*, encoding enzymes (dehydrogenase and cyclodehydratase) involved in post-translational modifications of the *orf*9-encoded peptide (Melby et al., 2014). *orf*17 is homologous to *llsP*, whose product is a metalloprotease putatively responsible for bacteriocin leader region cleavage. Recently, it has been demonstrated that listeriolysin S doesn't contribute to *L. monocytogenes* tissue injury and virulence in inner host organs, but it is an SLS-like virulence factor targeting exclusively prokaryotic cells thus suggesting a role in the modulation of the host microbiota (Quereda et al., 2016, 2017). A similar function may therefore be hypothesized also for the bacteriocin encoded by Tn*7088* of *L. crispatus* M247.

4.4. Genomic sequence analysis of the Tn*7088*

The NCBI database of 26,353 complete microbial genomes (accessed in December 2021) was interrogated by using as a query the 79-bp *att*B. Homology search identified the Tn*7088* *att*B site in other 23 *Lactobacillus* strains (Figure 3). Sequence homology analysis identified eleven allelic variants of *att*B. The most represented *att*B allelic variant, namely *att*B1, found in 12 (52%) *L. crispatus* genomes, is identical to the *att*B of M247. The other variants were found also in different *Lactobacillus* species other than *crispatus*, namely *amyloliticus*, *kefiranofaciens, helveticus, amylovorus, kullabergensis* (Figure 3). Interestingly, in the genome of the *L. crispatus* type strain ATCC 33820, the *att*B site was 12 nucleotides in length. These 12 nucleotides, namely the last nucleotides at the 3' of the threonine tRNA encoding gene are conserved among the allelic variants of *att*B and can be considered the core of the integration site. In seven *L. crispatus* genomes, namely strains Lc116, Lc1700, Lc1226, 2029, PMC209, CO3MRSI1, ATCC 33820, of which the latter carrying the *att*B variant 11, Tn*7088*-like elements were integrated into the bacterial chromosome. DNA sequence comparison of Tn*7088* with six of those Tn*7088*-like elements indicate that length ranges from 11,757-bp for strain 2029 up to 16,080-bp for strain Lc1700. The 6,026-bp DNA sequence spanning *orf*1 to *orf*9 of Tn*7088* is conserved among the Tn*7088*-like elements, whereas the remaining DNA sequence containing the bacteriocin biosynthetic gene cluster (*orfs*10 to 18) is subject to insertions and deletions. Indeed, ISs (mainly IS*1201* and IS*Lhe5*) are found integrated in the bacteriocin biosynthetic gene cluster of Tn*7088* and Tn*7088*-like elements and cause disruption of different *orfs* in different strains, except for the Tn*7088*-like element of strain 2029 which harbors an undisrupted biosynthetic gene cluster devoid of additional inserted genetic material (Figure 4).

5. CONCLUSIONS

In the present study we reported the complete genome of the *L. crispatus* probiotic strain M247 and we characterized a novel transposon named Tn*7088*. We found that Tn*7088* i) has the structure of an integrative and mobilizable element and integrates at a 79-bp *att*B site involving the last 12 nt at the 3' end of a threonine tRNA encoding gene, ii) excises from the M247 and ATCC chromosomes producing circular intermediate and reconstitution of the *att*B site, at similar frequencies, iii) contains a bacteriocin biosynthetic gene cluster homologous of the listeriolysin S gene cluster of *L. monocytogenes* and iv) is present in the complete genomes of other seven *L. crispatus* strains showing variability within the biosynthetic gene cluster consisting of insertion and deletions caused by integration of ISs elements. Tn*7088* is the first example of an integrative and mobilizable element in *L. crispatus* containing a bacteriocin-encoding biosynthetic gene cluster, which may contribute to the niche-adaptive traits and to the probiotic potential of its host bacterial strain.

TABLES

Table 1. **Oligonucleotide primers.**

Name	Sequence (5' to 3')	GenBank ID: nucleotides
IF1191	TTTAGGATAAGTCCTGGTCAA	CP088015: 1551029 – 1551050
IF1192	ATGTAAGAAGCTGCCTTAGAT	CP088015: 1561768 – 1561748
IF1193	TAGTTCAAGCAGAGCACCAA	CP088015: 1571687 – 1571710
IF1194	CTTGTCTGTAAAATACGATCA	CP088015: 1571687 – 1571710
IF1349	CGGGTAAGACAACGAAGAGT	CP088015: 1526733 – 1526752
IF1350	TCCAACGCCTGTTAAATCACTA	CP088015: 1562987 – 1562969
IF1444	GCAAATTACTGTTACGAGTCTT	CP072197: 1560295 – 1560315
IF1487	TGTGCCACACGGTTTCTAGA	CP088015: 1553452 – 1553431
IF1488	TTTATAGTACCTTTGCCACACAA	CP088015: 1528105 – 1528084
IF1511	GAAATAAAATGGGATACATCAGGT	CP088015: 1000943 – 1000966
IF1512	CCACCAGTATAACCAGAAACTA	CP088015: 1043722 – 1043701
IF1513	CCGTAAGGAGGAGATGCTAA	CP088015: 1001184 – 1001165
IF1514	CGCTCTAGGGGTAAAACTCTA	CP088015: 1043408 – 1043428

Table 2. **Real-time PCR quantification of the M247 phage circular forms and reconstituted *att*B integration site[a].**

Strain	Circular Forms	Reconstituted *att*B site
M247	3.40×10^{-5} (± 5.26×10^{-6})	2.52×10^{-5} (± 1.83×10^{-7})

[a] Concentration was expressed as copies per chromosome.

***Table 3.* Annotated ORFs of the prophage of *L. crispatus* M247.**

		Homologous protein			
ORF (aa)[a]	**Annotation and comments (reference)**	**Protein ID Origin [E value]**[b]	**aa identity**	**aa similarity**	**Pfam domain [E value]**[c]
orf1 (36)	Peptide-methionine sulfoxide reductase MsrA, truncated (Wizemann et al., 1996)	WP_005726248.1 *L. crispatus* [8e-14]	32/32 (100%)	32/32 (100%)	
*orf*2 (407)	DNA integrase (Kwon et al., 1997)	DAW29718.1 *Siphoviridae sp. isolate ct06w1* [0.0]	407/407 (100%)	407/407 (100%)	Phage_integrase (180-373) [1.8e-25] Phage_int_SAM_5 (32-164) [1.8e-05]
orf3 (334)	Abortive infection protein (Garvey et al., 1995)	DAW29696.1 *Siphoviridae sp. isolate ct06w1* [0.0]	334/334 (100%)	334/334 (100%)	Abi_2 (31-239) [1.5e-36]
orf6 (126)	Winged helix-like DNA-binding domain superfamily YjcQ (Brennan & Matthews, 1989)	DAW29760.1 *Siphoviridae sp. isolate ct06w* [2e-96]	125/126 (99%)	126/126 (100%)	
orf7 (208)	Repressor protein CI (Paetzel et al., 1998)	DAW29717.1 *Siphoviridae sp. isolate ct06w* [9e-27]	208/208 (100%)	208/208 (100%)	Peptidase_S24 (86-202) [3.1e-26]
orf8 (74)	Helix-turn-helix XRE-family like protein (Brennan & Matthews, 1989)	DAW29695.1 *Siphoviridae sp. isolate ct06w* [9e-56]	74/74 (100%)	74/74 (100%)	
orf12 (285)	Dna polymerase B	DAW29710.1 *Siphoviridae sp. isolate ct06w1* [1e-158]	285/285 (100%)	285/285 (100%)	HTH_36 (24-74) [4.4e-05]
orf16 (145)	HNH endonuclease (Krishna, 2003)	DAW29694.1 *Siphoviridae sp. isolate ct06w1* [3e-111]	145/145 (100%)	145/145 (100%)	HNH_3 (66-112) [7.5e-10]

		Homologous protein			
ORF (aa)[a]	Annotation and comments (reference)	Protein ID Origin [E value][b]	aa identity	aa similarity	Pfam domain [E value][c]
orf21 (248)	Phage antirepressor KilAC domain-containing protein (Sandt et al., 2002)	DAW29715.1 *Siphoviridae sp. isolate ct06w1* [0.0]	247/248 (99%)	248/248 (100%)	AntA (17-85) [5.7e-19] ANT (124-236) [8.9e-32]
orf22 (73)	Restriction alleviation protein	DAW29742.1 *Siphoviridae sp. isolate ct06w1* [2e-55]	73/73 (100%)	73/73 (100%)	
tRNA-Met					
orf29 (176)	HNH endonuclease (Edgell, 2009)	DAW29714.1 *Siphoviridae sp. isolate ct06w1* [2e-138]	176/176 (100%)	176/176 (100%)	HNH (88-131) [8.1e-08]
orf30 (156)	Phage terminase, small subunit (Schouler & Ehrlich, 1994)	WP_060463559.1 *L.crispatus* [1e-55]	154/156 (99%)	155/156 (99%)	Terminase_4 (29-140) [4.3e-16]
orf31 (392)	IS*1201*, transposase, IS*256* family (de Los Reyes-Gavilán et al., 1992)	P35880 *L. helveticus* [0.0]	333/368 (90%)	352/368 (95%)	
orf33 (624)	Phage terminase, large subunit (Schouler & Ehrlich, 1994)	DAW29733.1 *Siphoviridae sp. isolate ct06w1* [0.0]	624/624 (100%)	624/624 (100%)	Terminase_1 (100-587) [1.5e-50]
orf35 (392)	Phage Portal protein (Moore & Prevelige, 2002)	WP_060464314.1 *L.crispatus* [3e-66]	392/392 (100%)	392/392 (100%)	Phage_portal (47-354) [4.9e-39]
orf36 (228)	ATP dependent Clp protease (Wang et al., 1997)	DAW29711.1 *Siphoviridae sp. isolate ct06w1* [7e-174]	228/228 (100%)	228/228 (100%)	CLP_protease (32-175) [2.0e-33]
orf37 (452)	Phage major capsid protein	DAW29706.1 *Siphoviridae sp. isolate ct06w1* [0.0]	452/452 (100%)	452/452 (100%)	Phage_capsid (132-421) [2.8e-17]

		Homologous protein			
ORF (aa)[a]	Annotation and comments (reference)	Protein ID Origin [E value][b]	aa identity	aa similarity	Pfam domain [E value][c]
orf38 (129)	Phage head-tail adaptor	DAW29705.1 *Siphoviridae sp. isolate ct06w1* [2e-98]	129/129 (100%)	129/129 (100%)	
orf39 (121)	Phage head closure knob	DAW29704.1 *Siphoviridae sp. isolate ct06w1* [8e-94]	121/121 (100%)	121/121 (100%)	Phage_H_join (12-110) [0.00065]
orf40 (132)	Phage type I neck protein	DAW29703.1 *Siphoviridae sp. isolate ct06w1* [4e-101]	132/132 (100%)	132/132 (100%)	HK97-gp10_like (6-98) [0.00044]
orf41 (124)	Phage tail completion protein	DAW29702.1 *Siphoviridae sp. isolate ct06w1* [4e-94]	124/124 (100%)	124/124 (100%)	
orf42 (258)	Phage major tail protein (Pell et al., 2013)	DAW29701.1 *Siphoviridae sp. isolate ct06w1* [0.0]	258/258 (100%)	258/258 (100%)	Phage_TTP_1 (8-211) [1.7e-42]
orf43 (137)	Tail assembly chaperone protein (Pell et al., 2013)	DAW29701.1 *Siphoviridae sp. isolate ct06w1* [4e-103]	137/137 (100%)	137/137 (100%)	Phage_TAC_3 (7-122) [1.5e-11]
orf45 (2339)	Phage minor tail protein (Pell et al., 2013)	DAW29700.1 *Siphoviridae sp. isolate ct06w1* [0.0]	2338/2339 (99%)	2339/2339(100%)	PhageMin_Tail (315-532) [1.1e-38]
orf46 (253)	Phage distal tail protein (Pell et al., 2013)	DAW29699.1 *Siphoviridae sp. isolate ct06w1* [0.0]	252/253 (99%)	253/253 (100%)	
orf47 (1135)	Phage tail protein (Pell et al., 2013)	DAW29698.1 *Siphoviridae sp. isolate ct06w1* [0.0]	1135/1135 (100%)	1135/1135(100%)	Prophage_tail (73-441) (744-824) [7.2e-13]
orf50 (410)	Phage tail protein, truncated (Upton & Buckley, 1995)	DAW29725.1 *Siphoviridae sp. isolate ct06w1* [0.0]	403/407 (99%)	405/407 (99%)	Bppu_N (3-164) [3.7e-08]

ORF (aa)[a]	Annotation and comments (reference)	Homologous protein Protein ID Origin [E value][b]	aa identity	aa similarity	Pfam domain [E value][c]
orf51 (392)	IS*1201E*, transposase, IS*256* family (de Los Reyes-Gavilán et al., 1992)	P35880 *L. helveticus* [0.0]	333/368 (90%)	352/368 (95%)	
orf52 (416)	Phage tail protein, truncated (Upton & Buckley, 1995)	DAW29725.1 *Siphoviridae sp. isolate ct06w1* [0.0]	397/398 (99%)	398/398 (100%)	Lipase_GDSL (196-401) [8.4e-12]
orf55 (142)	Holin family protein (Gindreau & Lonvaud-Funel, 1999)	DAW29722.1 *Siphoviridae sp. isolate ct06w1* [8e-107]	142/142 (100%)	142/142 (100%)	Phage_holin_6_1 (3-110) [1.3e-05]
orf56 (294)	Cpl1 lysin (Henrissat et al., 1995)	DAW29716.1 *Siphoviridae sp. isolate ct06w1* [0.0]	294/294 (100%)	294/294 (100%)	Glyco_hydro_25 (9-196) [1.8e-21]
orf57, 5'end of *msrA*	Peptide-methionine sulfoxide reductase MsrA, 97 nucleotides included in the *att*B (Wizemann et al., 1996)				

[a] The number of amino acids of each ORF is shown in parenthesis.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

***Table 4*. Real-time PCR quantification of Tn*7088* circular forms and reconstituted *att*B integration site[a].**

Strain	Circular Forms	Reconstituted *att*B site
M247	3.92×10^{-5} (± 2.17×10^{-7})	1.03×10^{-4} (± 3.27×10^{-5})
ATCC 33820	1.81×10^{-5} (± 2.14×10^{-6})	3.45×10^{-5} (± 1.21×10^{-6})

[a] Concentration was expressed as copies per chromosome.

***Table 5.* Insertion sequences found in the *L. crispatus* M247 genome.**

IS family	IS name (reference)	Intact transposase	Truncated or frameshift	Total number
IS*256*	IS*1201* (de Los Reyes-Gavilán et al., 1992)	74	9	83
IS*982*	IS*Lhe5* (Callanan et al., 2008)	7	2	9
	IS*Lh1* (D. Pridmore et al., 1994)	15	0	15
IS*3* sub-group IS*150*	IS*Lhe6* (Callanan et al., 2008)	5	2	7
	IS*Enfa5* (Y. Liu et al., 2012)	8	2	10
	IS*L6* (Lapierre et al., 2002)	1	0	1
	IS*Sau2* (Holden et al., 2004)	6	0	6
IS*110*	IS*Spn10* (Baldry S., 2010[a])	6	0	6
	IS*LHe4* (Callanan et al., 2008)	13	1	14
	IS*L4* (Lapierre et al., 2002)	1	0	1
IS*30*	IS*1139* (Lortie et al., 1994)	13	3	16
	IS*Sag3* (Tettelin et al., 2005)	2	0	2
	IS*Ljo1* (R. D. Pridmore et al., 2004)	1	1	2
	IS*1070* (Vaughan & de Vos, 1995)	1	0	1

IS family	IS name (reference)	Intact transposase	Truncated or frameshift	Total number
IS*4* sub-group IS*Pepr1*	IS*L5* (Lapierre et al., 2002)	9	0	9
	IS*Lre1* (De Palmenaer et al., 2008)	2	1	3
IS*Lre2*	IS*Lcr2* (Guerillot and Glaeser, 2012[a])	7	4	11
IS*4* sub-group IS*4*	IS*1675* (Rincé et al., 1997)	10	0	10
IS*66*	IS*Swo2* (Copeland et al., 2006[a])	4	0	4
	IS*Cth11* (Copeland et al., 2007[a])	4	0	4
IS*200*/IS*605* sub-group IS*1341*	IS*Lhe65* (Callanan et al., 2008)	2	0	2
	IS*Bth17* (Ziniu and Qiu, 2009[a])	1	0	1
IS*1182*	IS*Lac* (Altermann et al., 2005)	2	1	3
IS*200*/IS*605*	IS*Ljo5* (R. D. Pridmore et al., 2004)	2	0	2
IS*L3*	IS*Lhe2* (Callanan et al., 2008)	1	0	1
	IS*Sm4* (Boyd, 2013[a])	1	0	1
IS*NCY*	IS*H7A* (Ng et al., 1998)	2	0	2
Total		200	26	226

[a] Direct submission - no publication associated with the IS

***Table 6*. Annotated ORFs of the Tn*7088* of *L. crispatus* M247.**

		Homologous protein			
ORF (aa)[a]	**Annotation and comments (reference)**	**Protein ID Origin [E value][b]**	**aa identity**	**aa similarity**	**Pfam domain [E value][c]**
*orf*1 (254)	Transcriptional regulator, putative (Brennan & Matthews, 1989)				HTH_3 (6–66) [1.0e -12]
*orf*3 (183)	Cell division protein FtsK, putative (Begg et al., 1995)				
*orf*4 (264)	Cell division protein FtsK (Begg et al., 1995)				FtsK_SpoIIIE (2-109) [1.1e-05]
*orf*5 (273)	Relaxase (Balson & Shaw, 1990)				Rep_trans (133-273) [1.1e-19]
*orf*7 (59)	Putative excisionase, helix-turn-helix domain-containing protein				
*orf*8 (409)	DNA integrase (Kwon et al., 1997)				Phage_integrase (183-397) [9.8e-14]
*orf*9 (44)	Listeriolysin S family TOMM bacteriocin (S. W. Lee et al., 2008)	WP_180680548 *L. monocytogenes* [0.25] (47aa)	11/25(44%)	13/25(52%)	
*orf*10 (284)	ABC transporter: ATP-binding protein (S. W. Lee et al., 2008)				ABC_tran (21-150) [7.6e-22]
*orf*11 (251)	ABC transporter: permease (S. W. Lee et al., 2008)				ABC2_membrane (5-212) [3.2e-12]

ORF (aa)[a]	Annotation and comments (reference)	Homologous protein Protein ID Origin [E value][b]	aa identity	aa similarity	Pfam domain [E value][c]
*orf*12 (101)	Glucosyl transferase, LlsX family protein (Cotter et al., 2008)	WP_187990302 *L. monocytogenes* [1e-12]	40/100 (40%)	63/100 (63%)	Glucos_trans_II (15-99) [0.00018]
*orf*13 (292)	Peptide dehydrogenase, SagB family (S. W. Lee et al., 2008)	WP_117383294 *L. monocytogenes* [5e-95]	135/291 (46%)	192/291 (65%)	Nitroreductase (105-285) [5.8e-13]
*orf*14 (240)	Listeriolysin S biosynthesis cyclodehydratase, truncated (S. Lee, 2020)	WP_115905105 *L. monocytogenes* [1e-45]	88/219 (40%)	133/219 (60%)	
*orf*15 (285)	IS*Lhe5*, transposase, IS*982* family (Callanan et al., 2008)	WP_012211839 *L. helveticus* [0.0]	240/285 (84%)	259/285 (90%)	
*orf*16 (438)	Cyclodehydratase, YcaO-like family (S. W. Lee et al., 2008)	WP_003730945 *L. monocytogenes* [4e-172]	240/440 (55%)	312/440 (70%)	YcaO (70-407) [1.5e-28]
*orf*17 (170)	Intramembrane metalloprotease, CPBP family, putative bacteriocin leader cleavage (S. Lee, 2020)	EAG8006596 *L. monocytogenes* [4e-172]	65/161 (40%)	102/161 (63%)	CPBP (5-170) [0.045]
*orf*18 (408)	IS*1201*, transposase, IS*256* family (de Los Reyes-Gavilán et al., 1992)	P35880 *L. helveticus* [0.0]	333/368 (90%)	352/368 (95%)	

[a] The number of amino acids of each ORF is shown in parenthesis.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

Supplementary Table S1. **General statistics of M247 Nanopore and Illumina reads.**

	Nanopore reads[a] (SRR17479173)		**Illumina reads[b] (SRR17479172)**	
	Overall	**95x**	**R1**	**R2**
Reads (n)	153,079	5,275	268,663	268,663
Mean read length	13,483.4	41,433.7	209.4	194.4
Median read length	7,705.0	36,622.0	250.0	220.0
Read length N50[c]	25,587	44,883	251	225
Mean read quality (Q)[d]	11.8	13.9	33.8	30.2
Median read quality (Q)[d]	12.0	13.9	35.0	29.8
Sequencing output (no of bases)	2,064,031,022	218,562,625	56,246,560	52,216,531

[a] The overall Nanopore reads were filtered to obtain a 95x coverage of a 2.3 Mbp genome size

[b] R1 and R2 refer to forward and reverse Illumina reads, respectively

[c] N50 is the length of a sequence in a set for which all sequences of that length or greater sum to 50% of the set's total size.

[d] Phred quality score Q expresses the confidence in a particular base-call and is logarithmically related to the base-calling error probability P (Q= -10 log10 P).

Acknowledgements

Illumina genome sequencing was provided by MicrobesNG.

6. REFERENCES

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., & Wishart, D. S. (2016). PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, *44*(W1), W16–W21. https://doi.org/10.1093/nar/gkw387.

Balson, D. F., & Shaw, W. V. (1990). Nucleotide sequence of the *rep* gene of staphylococcal plasmid pCW7. *Plasmid*, *24*(1), 74–80. https://doi.org/10.1016/0147-619x(90)90027-a.

Begg, K. J., Dewar, S. J., & Donachie, W. D. (1995). A new *Escherichia coli* cell division gene, *ftsK.Journal of Bacteriology*, *177*(21), 6211–6222.https://doi.org/10.1128/jb.177.21.6211-6222.1995.

Bellanger, X., Payot, S., Leblond-Bourget, N., & Guédon, G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiology Reviews*, *38*(4), 720–760. https://doi.org/10.1111/1574-6976.12058.

Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif. *Journal of Biological Chemistry*, *264*(4), 1903–1906. https://doi.org/10.1016/S0021-9258(18)94115-3.

Callanan, M., Kaleta, P., O'Callaghan, J., O'Sullivan, O., Jordan, K., McAuliffe, O., Sangrador-Vegas, A., Slattery, L., Fitzgerald, G. F., Beresford, T., & Ross, R. P. (2008). Genome sequence of *Lactobacillus helveticus* , an organism distinguished by selective gene loss and insertion sequence element expansion. *Journal of Bacteriology*, *190*(2), 727–735. https://doi.org/10.1128/JB.01295-07

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., & Rajandream, M.-A. (2008). Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, *24*(23), 2672–2676. https://doi.org/10.1093/bioinformatics/btn529.

Castagliuolo, I., Galeazzi, F., Ferrari, S., Elli, M., Brun, P., Cavaggioni, A., Tormen, D., Sturniolo, G. C., Morelli, L., & PalÃ¹, G. (2005). Beneficial effect of auto-aggregating *Lactobacillus crispatus* on experimentally induced colitis in mice. *FEMS Immunology & Medical Microbiology*, *43*(2), 197–204. https://doi.org/10.1016/j.femsim.2004.08.011.

Cesena, C., Morelli, L., Alander, M., Siljander, T., Tuomola, E., Salminen, S., Mattila-Sandholm, T., Vilpponen-Salmela, T., & von Wright, A. (2001). *Lactobacillus crispatus* and its nonaggregating mutant in human colonization trials. *Journal of Dairy Science*, *84*(5), 1001–1010. https://doi.org/10.3168/jds.S0022-0302(01)74559-6.

Cotter, P. D., Draper, L. A., Lawton, E. M., Daly, K. M., Groeger, D. S., Casey, P. G., Ross, R. P., & Hill, C. (2008). Listeriolysin S, a novel peptide haemolysin associated with a subset of lineage i *Listeria monocytogenes*. *PLoS Pathogens*, *4*(9), e1000144. https://doi.org/10.1371/journal.ppat.1000144.

Cotter, P. D., Ross, R. P., & Hill, C. (2013). Bacteriocins—A viable alternative to antibiotics? *Nature Reviews Microbiology*, *11*(2), 95–105. https://doi.org/10.1038/nrmicro2937.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(W1), W246–W251. https://doi.org/10.1093/nar/gky425.

De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England)*, *34*(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.

de Los Reyes-Gavilán, C. G., Limsowtin, G. K., Tailliez, P., Séchaud, L., & Accolas, J. P. (1992). A *Lactobacillus helveticus*-specific DNA probe detects restriction fragment length polymorphisms in this species. *Applied and Environmental Microbiology*, *58*(10), 3429–3432. https://doi.org/10.1128/aem.58.10.3429-3432.1992.

De Palmenaer, D., Siguier, P., & Mahillon, J. (2008). IS*4* family goes genomic. *BMC Evolutionary Biology*, *8*, 18. https://doi.org/10.1186/1471-2148-8-18.

Donnarumma, G., Molinaro, A., Cimini, D., De Castro, C., Valli, V., De Gregorio, V., De Rosa, M., & Schiraldi, C. (2014). *Lactobacillus crispatus* L1: High cell density cultivation and exopolysaccharide structure characterization to highlight potentially beneficial effects against vaginal pathogens. *BMC Microbiology*, *14*(1), 137. https://doi.org/10.1186/1471-2180-14-137.

Edgell, D. R. (2009). Selfish DNA: homing endonucleases find a home. *Current Biology : CB*, *19*(3), R115-117. https://doi.org/10.1016/j.cub.2008.12.019.

Fagan, R. P., & Fairweather, N. F. (2014). Biogenesis and functions of bacterial S-layers. *Nature Reviews Microbiology*, *12*(3), 211–222. https://doi.org/10.1038/nrmicro3213.

Felden, B., Atkins, J. F., & Gesteland, R. F. (1996). TRNA and mRNA both in the same molecule. *Nature Structural Biology*, *3*(6), 494. https://doi.org/10.1038/nsb0696-494.

Garvey, P., Fitzgerald, G. F., & Hill, C. (1995). Cloning and DNA sequence analysis of two abortive infection phage resistance determinants from the lactococcal plasmid pNP40.*Applied and Environmental Microbiology*, *61*(12), 4321–4328. https://doi.org/10.1128/aem.61.12.4321-4328.1995.

Gindreau, E., & Lonvaud-Funel, A. (1999). Molecular analysis of the region encoding the lytic system from *Oenococcus oeni* temperate bacteriophage phi 10MC. *FEMS Microbiology Letters*, *171*(2), 231–238. https://doi.org/10.1111/j.1574-6968.1999.tb13437.

Guédon, G., Libante, V., Coluzzi, C., Payot, S., & Leblond-Bourget, N. (2017). The obscure world of integrative and mobilizable elements, highly widespread elements that pirate bacterial conjugative systems. *Genes*, *8*(11), 337. https://doi.org/10.3390/genes8110337.

Hammami, R., Zouhir, A., Ben Hamida, J., & Fliss, I. (2007). BACTIBASE: A new web-accessible database for bacteriocin characterization. *BMC Microbiology*, *7*(1), 89. https://doi.org/10.1186/1471-2180-7-89.

Heilbronner, S., Krismer, B., Brötz-Oesterhelt, H., & Peschel, A. (2021). The microbiome-shaping roles of bacteriocins. *Nature Reviews Microbiology*, *19*(11), 726–739. https://doi.org/10.1038/s41579-021-00569-w.

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences*, *92*(15), 7090–7094. https://doi.org/10.1073/pnas.92.15.7090.

Holden, M. T. G., Feil, E. J., Lindsay, J. A., Peacock, S. J., Day, N. P. J., Enright, M. C., Foster, T. J., Moore, C. E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S. D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., … Parkhill, J. (2004). Complete genomes of two clinical *Staphylococcus aureus* strains: Evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), 9786–9791. https://doi.org/10.1073/pnas.0402521101.

Hynönen, U., & Palva, A. (2013). *Lactobacillus* surface layer proteins: Structure, function and applications. *Applied Microbiology and Biotechnology*, *97*(12), 5225–5243. https://doi.org/10.1007/s00253-013-4962-2.

Iannelli, F., Giunti, L., & Pozzi, G. (1998). Direct sequencing of long polymerase chain reaction fragments. *Molecular Biotechnology*, *10*(2), 183–185. https://doi.org/10.1007/BF02760864.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., … McArthur, A. G.

(2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(D1), D566–D573. https://doi.org/10.1093/nar/gkw1004.

Kirjavainen, P. V., Ouwehand, A. C., Isolauri, E., & Salminen, S. J. (1998). The ability of probiotic bacteria to bind to human intestinal mucus. *FEMS Microbiology Letters*, 5.

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8.

Krishna, S. S. (2003). Structural classification of zinc fingers: SURVEY AND SUMMARY. *Nucleic Acids Research*, *31*(2), 532–550. https://doi.org/10.1093/nar/gkg161

Kwon, H. J., Tirumalai, R., Landy, A., & Ellenberger, T. (1997). Flexibility in DNA recombination: Structure of the lambda integrase catalytic core. *Science (New York, N.Y.)*, *276*(5309), 126–131. https://doi.org/10.1126/science.276.5309.126

Lapierre, L., Mollet, B., & Germond, J.-E. (2002). Regulation and Adaptive Evolution of Lactose Operon Expression in *Lactobacillus delbrueckii*. *Journal of Bacteriology*, *184*(4), 928–935. https://doi.org/10.1128/jb.184.4.928-935.2002

Lee, S. (2020). Bacteriocins of *Listeria monocytogenes* and their potential as a virulence factor. *Toxins*, *12*(2), 103. https://doi.org/10.3390/toxins12020103.

Lee, S. W., Mitchell, D. A., Markley, A. L., Hensler, M. E., Gonzalez, D., Wohlrab, A., Dorrestein, P. C., Nizet, V., & Dixon, J. E. (2008). Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the National Academy of Sciences*, *105*(15), 5879–5884. https://doi.org/10.1073/pnas.0801338105.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., & Ou, H.-Y. (2019). ICEberg 2.0: An updated database of bacterial integrative and conjugative elements. *Nucleic Acids Research*, *47*(D1), D660–D665. https://doi.org/10.1093/nar/gky1123.

Liu, Y., Wang, Y., Wu, C., Shen, Z., Schwarz, S., Du, X.-D., Dai, L., Zhang, W., Zhang, Q., & Shen, J. (2012). First report of the multidrug resistance gene *cfr* in *Enterococcus faecalis* of animal origin. *Antimicrobial Agents and Chemotherapy*, *56*(3), 1650–1654. https://doi.org/10.1128/AAC.06091-11.

Lopes, A., Tavares, P., Petit, M.-A., Guérois, R., & Zinn-Justin, S. (2014). Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics*, *15*(1), 1027. https://doi.org/10.1186/1471-2164-15-1027.

Lortie, L. A., Gagnon, G., & Frenette, M. (1994). IS1139 from *Streptococcus salivarius*: Identification and characterization of an insertion sequence-like element related to mobile DNA elements from Gram-negative bacteria. *Plasmid*, *32*(1), 1–9. https://doi.org/10.1006/plas.1994.1038.

Lutcke, H. (1995). Signal Recognition Particle (SRP), a ubiquitous initiator of protein translocation. *European Journal of Biochemistry*, *228*(3), 531–550. https://doi.org/10.1111/j.1432-1033.1995.tb20293.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J., & Koonin, E. V. (2011). Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, *9*(6), 467–477. https://doi.org/10.1038/nrmicro2577.

Marcotte, H., Ferrari, S., Cesena, C., Hammarstrom, L., Morelli, L., Pozzi, G., & Oggioni, M. R. (2004). The aggregation-promoting factor of *Lactobacillus crispatus* M247 and its genetic locus. *Journal of Applied Microbiology*, *97*(4), 749–756. https://doi.org/10.1111/j.1365-2672.2004.02364.

McAuliffe, O., Ross, R. P., & Hill, C. (2001). Lantibiotics: Structure, biosynthesis and mode of action. *FEMS Microbiology Reviews*, *25*(3), 285–308. https://doi.org/10.1111/j.1574-6976.2001.tb00579.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D., & Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. https://doi.org/10.1093/bib/bbs012.

Moore, S. D., & Prevelige, P. E. J. (2002). DNA packaging: A new class of molecular motors. *Current Biology : CB*, *12*(3), R96-98. https://doi.org/10.1016/s0960-9822(02)00670.

Nardini, P., Ñahui Palomino, R. A., Parolin, C., Laghi, L., Foschi, C., Cevenini, R., Vitali, B., & Marangoni, A. (2016). *Lactobacillus crispatus* inhibits the infectivity of *Chlamydia trachomatis* elementary bodies, in vitro study. *Scientific Reports*, *6*(1), 29024. https://doi.org/10.1038/srep29024.

Ng, W. V., Ciufo, S. A., Smith, T. M., Bumgarner, R. E., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J., Slagel, J., Hood, L., & DasSarma, S. (1998). Snapshot of a large dynamic replicon in a halophilic archaeon: Megaplasmid or minichromosome? *Genome Research*, *8*(11), 1131–1141. https://doi.org/10.1101/gr.8.11.1131.

Ojala, T., Kankainen, M., Castro, J., Cerca, N., Edelman, S., Westerlund-Wikström, B., Paulin, L., Holm, L., & Auvinen, P. (2014). Comparative genomics of *Lactobacillus crispatus* suggests novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. *BMC Genomics*, *15*(1), 1070. https://doi.org/10.1186/1471-2164-15-1070.

Pace, N. R., & Brown, J. W. (1995). Evolutionary perspective on the structure and function of ribonuclease P, a ribozyme. *Journal of Bacteriology*, *177*(8), 1919–1928. https://doi.org/10.1128/JB.177.8.1919-1928.1995.

Paetzel, M., Dalbey, R. E., & Strynadka, N. C. J. (1998). Crystal structure of a bacterial signal peptidase in complex with a b-lactam inhibitor. *Nature*, 396(6707):186-90. doi: 10.1038/24196.

Pan, M., Hidalgo-Cantabrana, C., & Barrangou, R. (2020). Host and body site-specific adaptation of *Lactobacillus crispatus* genomes. *NAR Genomics and Bioinformatics*, *2*(1), lqaa001. https://doi.org/10.1093/nargab/lqaa001.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114**.**

Pell, L. G., Cumby, N., Clark, T. E., Tuite, A., Battaile, K. P., Edwards, A. M., Chirgadze, N. Y., Davidson, A. R., & Maxwell, K. L. (2013). A conserved spiral structure for highly diverged phage tail assembly chaperones. *Journal of Molecular Biology*, *425*(14), 2436–2449. https://doi.org/10.1016/j.jmb.2013.03.035.

Petrova, M. I., Lievens, E., Malik, S., Imholz, N., & Lebeer, S. (2015). *Lactobacillus* species as biomarkers and agents that can promote various aspects of vaginal health. *Frontiers in Physiology*, *6*. https://doi.org/10.3389/fphys.2015.00081.

Pierro, F. D., Bertuccioli, A., Cattivelli, D., Soldi, S., & Elli, M. (2018). *Lactobacillus crispatus M247: A possible tool to counteract CST IV. Nutrafoods*, 17:169-172 https://doi.org/10.17470/NF-018-0001-4.

Pierro, F. D., Criscuolo, A. A., Giudici, A. D., Senatori, R., Sesti, F., Ciotti, M., & Piccione, E. (2021). Oral administration of *Lactobacillus crispatus* M247 to papillomavirus-infected women: Results of a preliminary, uncontrolled, open trial. *Minerva Obstet Gynecol.,*73(5):621-631. https://doi.org/10.23736/S2724-606X.21.04752-7.

Pridmore, D., Stefanova, T., & Mollet, B. (1994). Cryptic plasmids from *Lactobacillus helveticus* and their evolutionary relationship. *FEMS Microbiology Letters*, *124*(3), 301–305. https://doi.org/10.1111/j.1574-6968.1994.tb07300.

Pridmore, R. D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A.-C., Zwahlen, M.-C., Rouvet, M., Altermann, E., Barrangou, R., Mollet, B., Mercenier, A., Klaenhammer, T., Arigoni, F., & Schell, M. A. (2004). The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(8), 2512–2517. https://doi.org/10.1073/pnas.0307327101.

Quereda, J. J., Dussurget, O., Nahori, M.-A., Ghozlane, A., Volant, S., Dillies, M.-A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B., Cossart, P., & Pizarro-Cerda, J. (2016). Bacteriocin from epidemic *Listeria* strains alters the host intestinal microbiota to favor infection. *Proceedings of the National Academy of Sciences*, *113*(20), 5706–5711. https://doi.org/10.1073/pnas.1523899113.

Quereda, J. J., Meza-Torres, J., Cossart, P., & Pizarro-Cerdá, J. (2017). Listeriolysin S: A bacteriocin from epidemic *Listeria monocytogenes* strains that targets the gut microbiota. *Gut Microbes*, *8*(4), 384–391. https://doi.org/10.1080/19490976.2017.1290759.

Rincé, A., Dufour, A., Uguen, P., Le Pennec, J. P., & Haras, D. (1997). Characterization of the lacticin 481 operon: The *Lactococcus lactis* genes *lctF, lctE*, and *lctG* encode a putative ABC transporter involved in bacteriocin immunity. *Applied and Environmental Microbiology*, *63*(11), 4252–4260. https://doi.org/10.1128/aem.63.11.4252-4260.1997.

Sandt, C. H., Hopper, J. E., & Hill, C. W. (2002). Activation of prophage *eib* genes for immunoglobulin-binding proteins by genes from the IbrAB genetic island of *Escherichia coli* ECOR-9. *Journal of Bacteriology*, *184*(13), 3640–3648. https://doi.org/10.1128/JB.184.13.3640-3648.2002.

Santoro, F., Oggioni, M. R., Pozzi, G., & Iannelli, F. (2010). Nucleotide sequence and functional analysis of the *tet (*M)-carrying conjugative transposon Tn*5251* of *Streptococcus pneumoniae*: Tn*5251* of *Streptococcus pneumoniae*. *FEMS Microbiology Letters*, no-no. https://doi.org/10.1111/j.1574-6968.2010.02002.

Schouler, C., & Ehrlich, D. (1994). *Sequence and organization of the lactococcal prolate-headed blL67 phage genome. Microbiology (Reading)*, 140 (Pt11):3061-9. https://doi.org/10.1099/13500872-140-11-3061.

Shoemaker, N. B., Wang, G.-R., & Salyers, A. A. (2000). Multiple gene products and sequences required for excision of the mobilizable integrated *Bacteroides* element NBU1. *Journal of Bacteriology*, *182*(4), 928–936. https://doi.org/10.1128/JB.182.4.928-936.

Siciliano, R. A., Cacace, G., Mazzeo, M. F., Morelli, L., Elli, M., Rossi, M., & Malorni, A. (2008). Proteomic investigation of the aggregation phenomenon in *Lactobacillus crispatus*. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1784*(2), 335–342. https://doi.org/10.1016/j.bbapap.2007.11.007.

Sun, Z., Kong, J., Hu, S., Kong, W., Lu, W., & Liu, W. (2013). Characterization of a S-layer protein from *Lactobacillus crispatus* K313 and the domains responsible for binding to cell wall and adherence to collagen. *Applied Microbiology and Biotechnology*, *97*(5), 1941–1952. https://doi.org/10.1007/s00253-012-4044-x.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569.

Teodori, L., Colombini, L., Cuppone, A. M., Lazzeri, E., Pinzauti, D., Santoro, F., Iannelli, F., & Pozzi, G. (2021). Complete genome sequence of *Lactobacillus crispatus* Type Strain ATCC 33820. *Microbiology Resource Announcements*, *10*(32). https://doi.org/10.1128/MRA.00634-21.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of

Streptococcus agalactiae: Implications for the microbial 'pan-genome'. *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102.

Tisza, M. J., & Buck, C. B. (2021). A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences*, *118*(23), e2023202118. https://doi.org/10.1073/pnas.2023202118.

Upton, C., & Buckley, J. T. (1995). A new family of lipolytic enzymes? *Trends in Biochemical Sciences*, *20*(5), 178–179. https://doi.org/10.1016/s0968-0004(00)89002-7.

van Heel, A. J., de Jong, A., Song, C., Viel, J. H., Kok, J., & Kuipers, O. P. (2018). BAGEL4: A user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Research*, *46*(W1), W278–W281. https://doi.org/10.1093/nar/gky383.

Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V., & Chandler, M. (2011). ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*, *12*(3), R30. https://doi.org/10.1186/gb-2011-12-3-r30.

Vaughan, E. E., & de Vos, W. M. (1995). Identification and characterization of the insertion element IS*1070* from *Leuconostoc lactis* NZ6009. *Gene*, *155*(1), 95–100. https://doi.org/10.1016/0378-1119(94)00921-e.

Voltan, S., Castagliuolo, I., Elli, M., Longo, S., Brun, P., D'Incà, R., Porzionato, A., Macchi, V., Palù, G., Sturniolo, G. C., Morelli, L., & Martines, D. (2007). Aggregating phenotype in *Lactobacillus crispatus* determines intestinal colonization and TLR2 and TLR4 modulation in murine colonic mucosa. *Clinical and Vaccine Immunology*, *14*(9), 1138–1148. https://doi.org/10.1128/CVI.00079-07.

Voltan, S., Martines, D., Elli, M., Brun, P., Longo, S., Porzionato, A., Macchi, V., D'Incà, R., Scarpa, M., Palù, G., Sturniolo, G. C., Morelli, L., & Castagliuolo, I. (2008). *Lactobacillus* *crispatus* M247-Derived H2O2 Acts as a Signal Transducing Molecule Activating Peroxisome Proliferator Activated Receptor-γ in the Intestinal Mucosa. *Gastroenterology*, *135*(4), 1216–1227. https://doi.org/10.1053/j.gastro.2008.07.007.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963.

Walter, J. (2008). Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. *Applied and Environmental Microbiology*, *74*(16), 4985–4996. https://doi.org/10.1128/AEM.00753-08.

Wang, J., Hartling, J. A., & Flanagan, J. M. (1997). The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis. *Cell*, *91*(4), 447–456. https://doi.org/10.1016/S0092-8674(00)80431-6.

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Wizemann, T. M., Moskovitz, J., Pearce, B. J., Cundell, D., Arvidson, C. G., So, M., Weissbach, H., Brot, N., & Masure, H. R. (1996). Peptide methionine sulfoxide reductase contributes to the maintenance of adhesins in three major pathogens. *Proceedings of the National Academy of Sciences*, *93*(15), 7985–7990. https://doi.org/10.1073/pnas.93.15.7985.

FIGURE LEGEND

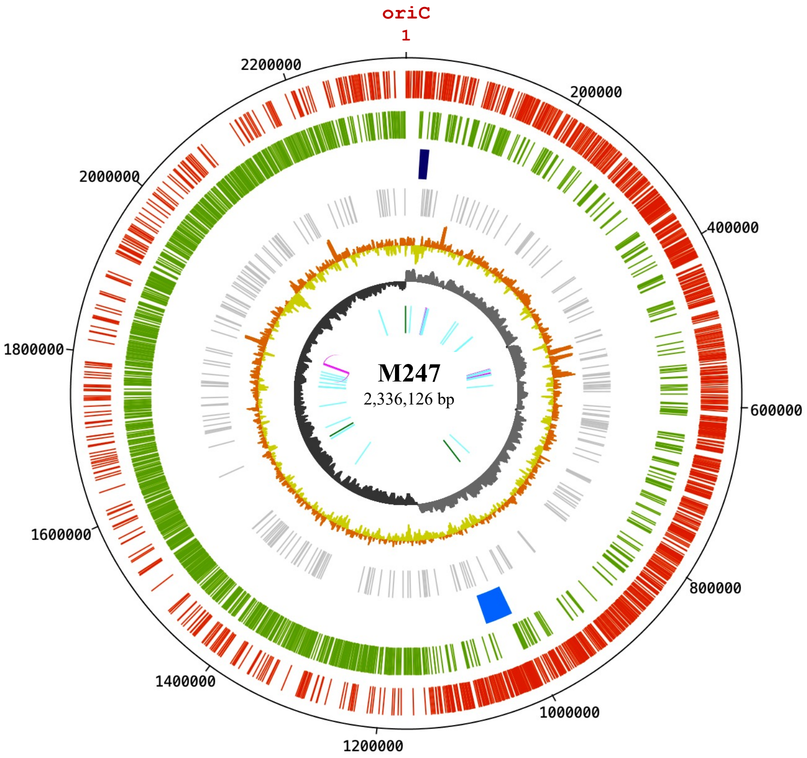


***Figure 1.* Circular representation of the *L. crispatus* M247 genome.** Circles range from 1 (outer circle) to 7 (inner circle). Circle 1 shows predicted coding regions located on the plus strand (orange). The second circle shows predicted coding regions on the minus strand (green). The dark blue and the blue blocks indicate the 14,184-bp long integrative and mobilizable element and the 42,649-bp long prophage, respectively. The fourth circle represents insertion sequences (grey ticks) distribution in the M247 genome. The fifth and the sixth circles show GC content (orange/yellow) and GC skew (black/grey), respectively. The innermost circle shows tRNA in light-blu, rRNA in pink and structural RNA in dark-green. Graphic genome representation was created using Artemis DNA-Plotter (v.17.0.1).

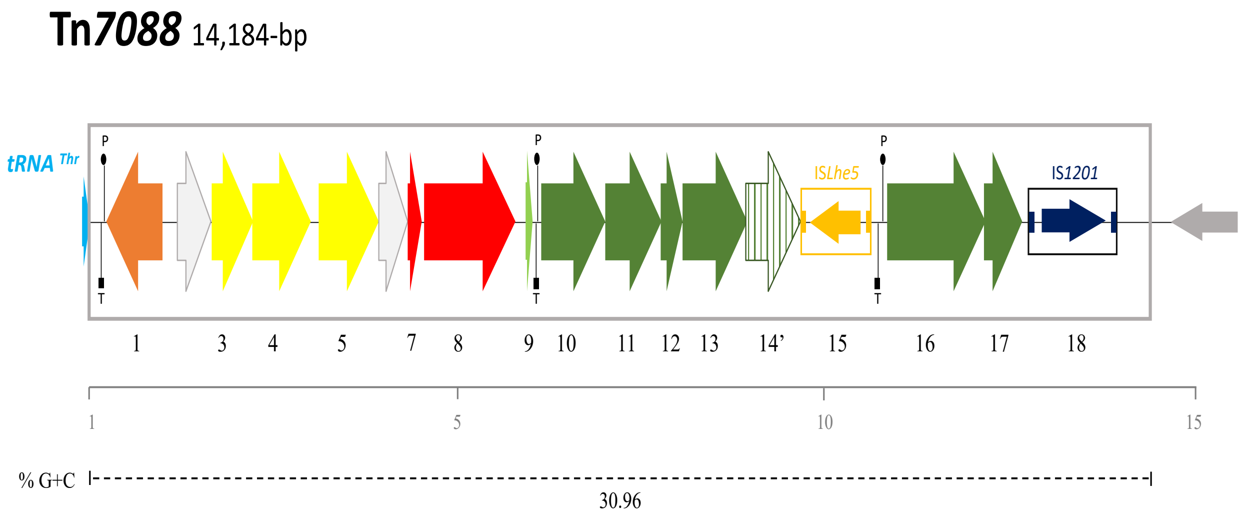
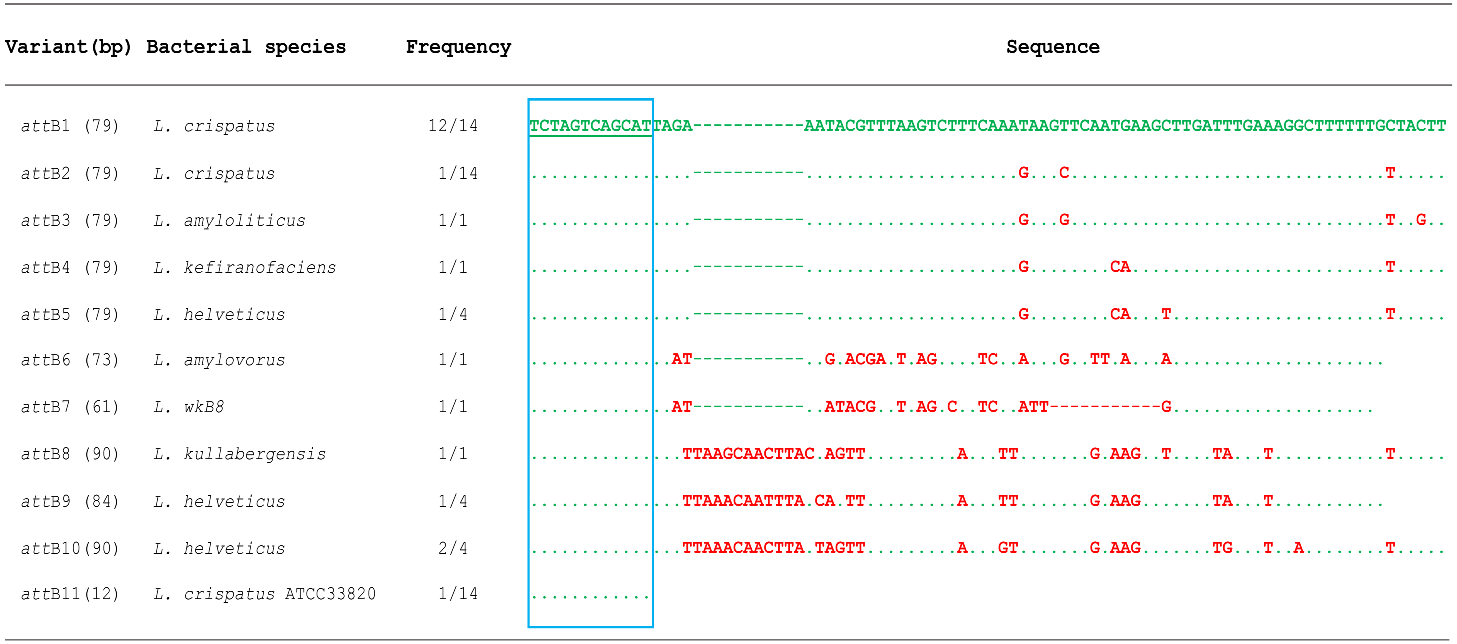


***Figure 2.* Structure of Tn*7088* of *L. crispatus* strain M247.** Tn*7088* is 14,184 bp-long and contains 18 ORFs. ORFs and their direction of transcription are represented by arrows, while annotated ORFs are indicated by sequential numbers. Insertion sequences are reported as boxed arrows and their inverted repeat are indicated by solid rectangles. All genes belonging to the bacteriocin biosynthetic gene-cluster are depicted in green. Genes involved in the putative intercellular mobilization of the element are represented in yellow, while genes for the integration/excision are highlighted in red. Pattern fill indicate truncated genes. In light blue is represented the tRNA^{Thr} gene of which the last 12 nucleotides at 3’ end are part of the Tn*7088* target site of integration. The GC content of the element is indicated by dotted bars. Scale, kilobases



Variant (bp)	Bacterial species	Frequency	Sequence
attB1 (79)	*L. crispatus*	12/14	TCTAGTCAGCATTAGA-----------AATACGTTTAAGTCTTTCAAATAAGTTCAATGAAGCTTGATTTGAAAGGCTTTTTTGCTACTT
attB2 (79)	*L. crispatus*	1/14	................-----------.....................G...C...............................T.....
attB3 (79)	*L. amyloliticus*	1/1	................-----------.....................G...G...............................T..G..
attB4 (79)	*L. kefiranofaciens*	1/1	................-----------.....................G........CA.........................T.....
attB5 (79)	*L. helveticus*	1/4	................-----------.....................G........CA...T.....................T.....
attB6 (73)	*L. amylovorus*	1/1	..............AT-----------..G.ACGA.T.AG....TC..A...G..TT.A...A.....................
attB7 (61)	*L. wkB8*	1/1	..............AT-----------..ATACG..T.AG.C..TC..ATT-----------G....................
attB8 (90)	*L. kullabergensis*	1/1	...............TTAAGCAACTTAC.AGTT.........A...TT.......G.AAG..T....TA...T...........T.....
attB9 (84)	*L. helveticus*	1/4	...............TTAAACAATTTA.CA.TT.........A...TT.......G.AAG.......TA...T...........
attB10 (90)	*L. helveticus*	2/4	...............TTAAACAACTTA.TAGTT.........A...GT.......G.AAG.......TG...T..A........T.....
attB11 (12)	*L. crispatus* ATCC33820	1/14	....

Figure 3. **Allelic variants of Tn*7088 att*B integration sites in other bacterial species.** Genome sequence analysis identified Tn*7088 att*B sites in the DNA sequence of other 22 complete genomes available in the microbial database (December 2021), of strains all belonging to genus *Lactobacillus*, with a size ranging from 12 nt of *L. crispatus* ATCC 33820 to 90 nt for *L. kullabergensis*. Inside *L. crispatus* and *L. helveticus* species, different strains can harbour different allelic variants (up to 3). Frequency expresses the number of strains per species displaying the associated *att*B variant. The 79-bp variant of M247 (*att*B1) is the most frequent among all variants and it is carried by the genome of 12 strains (52% of all strains analyzed), including FDAARGOS, CO3MRSI1, AB70, KT-11, Lc116, Lc1226, Lc1700, Lc 2029, PMC201, PRL2021, 1D, and used as reference for the alignment. Within the sequences identical nucleotides are indicated by periods, substitutions are in red. For better alignment, dashes are inserted. The 12 nucleotides belonging to the tRNA^{Thr} coding sequence are underlined in the reference sequence and boxed to highlight that are conserved among the strains.

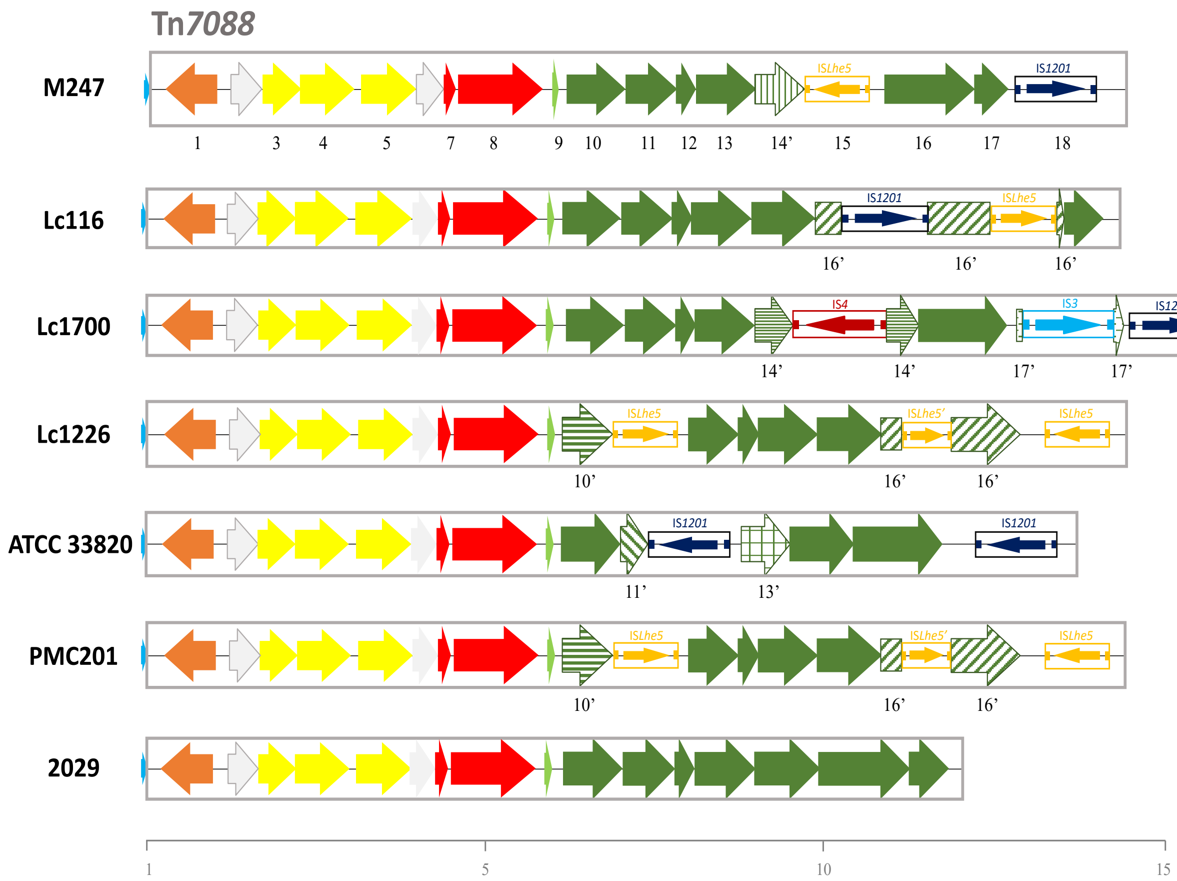
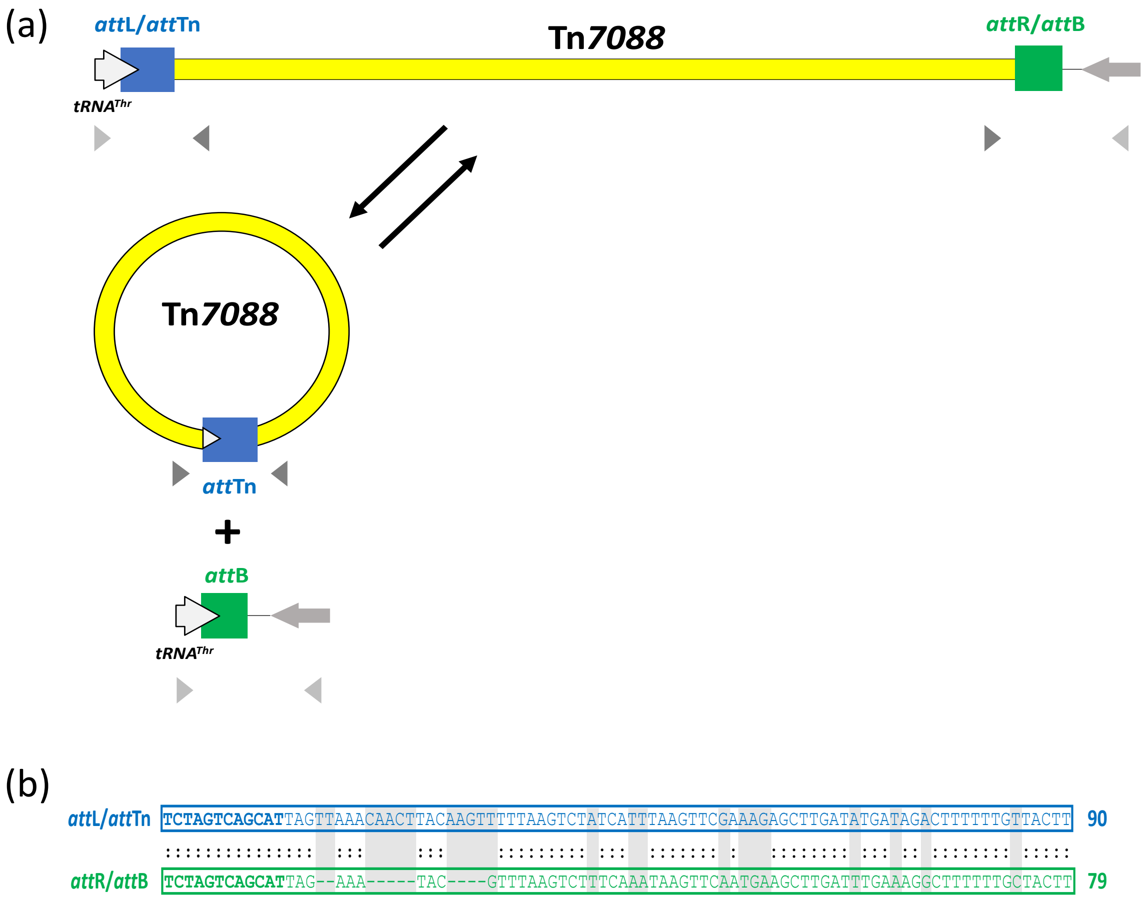


***Figure 4.* Schematic comparison of Tn*7088* with other *L. crispatus* Tn*7088*-like elements.** Tn*7088* is compared with six Tn*7088*-like elements identified in *L. crispatus* complete genomes available in the NCBI database. Strains names are reported on the left. ORFs are represented as arrows, insertion sequences are boxed arrows and their inverted repeat indicated by solid rectangles, whereas pattern fill arrows indicate truncated *orfs*. All Tn*7088*-like elements integrate in the 3' end of the threonine tRNA encoding gene (light blu arrow). DNA sequences vary in length from 11,757-bp for the Tn*7088*-like element of strain 2029 up to 16,080-bp for strain Lc1700. The 6,026-bp DNA sequence spanning from nucleotide 1 (*orf*1) to nucleotide 6,026 (*orf*9) of Tn*7088* is shared by all Tn*7088*-like elements, whereas the remaining DNA sequence containing the bacteriocin biosynthetic gene cluster vary among Tn*7088*-like elements. Variations include the integration of different ISs causing deletions of DNA sequences and disruptions of genes (*orf'*). The Tn*7088*-like element of strain 2029 harbors an undisrupted biosynthetic gene cluster devoid of additional inserted genetic material. Scale, kilobases.


(a)
attL/attTn
Tn7088
attR/attB
tRNA^{Thr}
Tn7088
attTn
+
attB
tRNA^{Thr}
(b)
attL/attTn
TCTAGTCAGCATTAGTTAAACAACTTACAAGTTTTTAAGTCTATCATTTAAGTTCGAAAGAGCTTGATATGATAGACTTTTTTGTTACTT
90
attR/attB
TCTAGTCAGCATTAG--AAA-----TAC----GTTTAAGTCTTTCAAATAAGTTCAATGAAGCTTGATTTGAAAGGCTTTTTTGCTACTT
79

Supplementary Figure S1. **(a) Mechanism of excision/integration of Tn*7088* and (b) sequence alignment of attachment sites in *L. crispatus* M247.** (a) Tn*7088* excises from the *L. crispatus* M247 chromosome producing a circular form and a reconstitution of *att*B insertion site. In the circular form the of Tn*7088* the left and right ends are joined by *att*Tn which is identical to attL whereas the reconstituted *att*B site is identical to *att*R. att sites are represented as filled rectangles, chromosomal genes as arrows, Tn*7088* as a yellow bar. Arrowheads represent PCR primers used for circular form (dark grey) and reconstituted *att*B site (light grey) detection. Not scaled. (b) *att*R-*att*B is 79-bp long and contains the last 12 nucleotides at 3’ end of a threonine encoding gene (bold letters). *att*L-*att*Tn contain 12 nucleotide changes and 11 nucleotides insertion compared to *att*R-*att*B (grey blocks).

CHAPTER 3

The unusual duplication of a 69.9-kb long chromosomal segment produces two copies of an inverted repeat and generates genomic instability in a laboratory strain of *Lactobacillus crispatus*

Lorenzo Colombini, Anna Cuppone, David Pinzauti, Francesco Santoro, Francesco Iannelli and Gianni Pozzi

Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy

Manuscript in preparation

1. ABSTRACT

Objectives: Laboratory bacterial strains tend to evolve according to the specific laboratory conditions in which have been sub-cultured for years since their first isolation. In this study, the unusual duplication of a 69.9-kb long chromosomal segment that produces two copies of an inverted repeat and generates genomic instability in a laboratory strain of *Lactobacillus crispatus*, was reported and analyzed.

Methods: The complete genome sequence of the laboratory strain of *L. crispatus*, carried in the University of Siena laboratory strain collection since 1990, namely M247_Siena, was determined combining both Nanopore and Illumina sequencing technologies. PCR genome mapping and coverage graph with Nanopore reads were used for genome structure analysis. Genome comparison analysis were carried out with bioinformatic tools.

Results: M247_Siena genome is organized in one circular chromosome of 2.38 Mb in length, containing 2,359 open reading frames with an average GC content of 37.07%. Compared to the original M247 strain, the genome of M247_Siena shows an unusual duplication of a 69.9-kb long chromosomal segment producing two copies of an inverted repeat located 224.4-kb apart. In M247_Siena the 69.9-kb repeat was found to replace a 15.4-kb sequence present at the same position on the M247 chromosome. Furthermore, in the M247_Siena genome a 224.4-kb inversion of the DNA sequence flanked by the 69.9-kb repeats and containing the origin of replication, occurred. Both the 69.9-kb and the 15.4-kb DNA regions are flanked by the same insertion sequences, namely IS*1201* and IS*Lcr2*, occurring in the form of inverted repeats on the chromosomal DNA sequence. PCR analysis indicated that the original M247 strain is subject to genomic instability consisting of chromosomal rearrangements that involve both the 69.9-kb and the 15.4-kb DNA sequences. These chromosomal rearrangements in the M247 bacterial population were rare. However, in the M247_Siena strain coverage graph with Nanopore reads indicated that the chromosomal rearrangements are favored by the presence of the 69.9-kb repeats. Multiple alignment of the 14 *L. crispatus* complete genomes showed that no other genome contains the

69.9-kb duplication, whereas the 15.4-kb region is always present, and that chromosomal rearrangements occur within the *L. crispatus* strains.

Conclusion: Our work reports the analysis of an unusual 69.9-kb DNA sequence duplication producing two copies of an inverted repeat in a laboratory strain of *L. crispatus*. The newly generated 69.9-kb repeats increase the genomic instability intrinsic to the strain.

2. INTRODUCTION

Bacterial genomes are remarkably stable from one generation to the next but are plastic on an evolutionary time scale, thus implying the existence of a delicate balance between genome maintenance and instability which depends on the type of bacteria, the cell cycle and the environment (Darmon & Leach, 2014). Genome instability is used by most bacteria as a driving force for survival, diversification, adaptation and evolution. Bacterial genomes are shaped by external agents such as mobile genetic elements and by internal events occurring mainly during DNA replication and DNA repair (Roth et al., 1996). Large-scale chromosomal rearrangements involving long stretches of DNA from few kilobases to sometimes up to millions of base pairs, have been detected in bacterial genomes (Darling et al., 2008; Sun et al., 2012). Chromosomal rearrangements include deletions, duplications, insertions and translocations resulting in loss, amplification, gain, change of location and orientation of a DNA segment, respectively (Periwal & Scaria, 2015). These types of structural variants may be important in evolution because they can alter the chromosome organization and gene expression in ways not possible through point mutations (Raeside et al., 2014). Chromosomal rearrangements can also markedly impact on the bacterial phenotype, inducing phase and antigenic variation that leads to the appearance of one or different bacterial subpopulation (Guérillot et al., 2019; Sousa et al., 1997). The detection of large-scale chromosomal rearrangements is subjected to bias due to i) the natural selective pressure which favors the maintenance of the relative replichores length and the distance of particular genes to the replication origin (*oriC*) or termination region (*ter*) (Eisen et al., 2000; Guinane et al., 2011; Suyama & Bork, 2001; Tillier & Collins, 2000); ii) the limit of the sequencing technology in producing reads that are enough long to span and describe the structural variants (Schmid et al., 2018). Genome instability also affects laboratory bacterial strains which evolves according to the specific laboratory conditions in which have been sub-cultured for decades since their first isolation (Fux et al., 2005). In this work we report the genome of a laboratory strain of *Lactobacillus crispatus*, named M247_Siena, which presents an unusual duplication of a 69.9-kb long chromosomal segment that produces two copies of an inverted repeat, located 224.4-kb apart. We also investigate i) the possible molecular mechanisms underlying the generation of such chromosomal structural variant and ii) the putative impact of these 69.9-kb long inverted repeats on genome stability.

3. MATERIALS AND METHODS

3.1. Bacterial strains and growth conditions

The *L. crispatus* M247 object of this study, isolated from the feces of a human newborn (Cesena et al., 2001) was carried in the University of Siena laboratory strain collection since 1990. Therefore, it was renamed M247_Siena to distinguish it from the previously described M247 strain (Colombini et al., unpublished). Frozen starter culture was grown in DeMan-Rogosa-Sharpe medium (MRS) broth (Oxoid LTD, Basingstoke, Hampshire, England) in anaerobic condition at 37°C.

3.2. DNA purification and quantification

Bacterial culture (500 ml) was harvested in exponential phase growth (OD_{590}=1.9) and centrifuged at 5,000 x *g* for 30 min at 4°C. *Lactobacillus* cells pellet was dry vortex-mixed for 2-3 min and incubated in 15 ml of protoplasting solution (20% Raffinose, 50 mM Tris-HCl [pH 8.0], 5 mM EDTA, 4 mg/ml lysozyme) at 37°C for 1 h to obtain protoplasts formation. Protoplast solution was centrifuged (5,000 x *g* for 5 min) and, to obtain osmotic lysis, the pellet was resuspended in 15 ml of ddH_2O and incubated at 37°C for 30 min adding 100 µg/ml proteinase K (Merck KGaA, Darmstadt, Germany) and SDS (after 15 min) at a final concentration of 0.5%. Then NaCl was added at a final concentration of 0.55 M and the mixture was incubated for additional 10 min. High molecular weight DNA was extracted with 1 volume of chloroform-isoamyl alcohol (24:1 [v:v]). DNA was precipitated in 0.6 volumes of ice-cold isopropanol and spooled on a glass rod. DNA was resuspended in 10-fold diluted saline-sodium citrate (SSC) 1x buffer, then adjusted to 1x SSC. The DNA solution was homogenized using a rotator mixer and maintained at +4°C. DNA was quantified with Qubit 2.0 Fluorometer (Invitrogen, Life Technologies, Carlsbad, CA, United States) using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific) and with spectrophotometer (Implen, Munich, Germany). DNA integrity and size were assessed by horizontal gel electrophoresis using 0.6% Seakem LE (Lonza, Rockland, ME USA) agarose in 0.5X Tris Borate EDTA running buffer.

3.3. Illumina Whole Genome Sequencing

Illumina sequencing was performed at MicrobesNG (University of Birmingham, United Kingdom) using Nextera library preparation kit (Illumina Inc., San Diego, USA) followed by sequencing on a NovaSeq 6000 device (Illumina Inc., San Diego, USA) (2x250 bp paired-end sequencing). Illumina reads were analyzed with NanoPlot v1.18.2 (De Coster et al., 2018). Illumina reads properties and accession numbers were reported in Supplementary Table S1.

3.4. Nanopore Whole Genome Sequencing

Sequencing reactions were carried out in 1.5 ml LoBind tubes (Sarstedt, Nümbrecht, Germany) using wide bore (∅1.2 mm) tips for DNA manipulation in order to reduce physical shearing. DNA size selection of the genomic DNA was obtained with 0.5 volumes of AMPure XP beads (Beckman Coulter, Milano, Italy) according to manufacturer's instructions. 2 μg of size-selected DNA were used for library construction by using the SQK-LSK 108 kit (Oxford Nanopore Technologies, Oxford, United Kingdom). Library preparation was performed following the manufacturer's protocol with the following modifications: (i) incubation on rotator mix for 15 min; (ii) the Library Loading Beads (LLB) were not added. Finally, 1 μg of DNA library was loaded onto a R9.4 MinION flow cell (Oxford Nanopore Technologies). A 48-h sequencing run was performed on GridION device (Oxford Nanopore Technologies). Real time base calling was performed with Guppy v3.2.6 (Oxford Nanopore Technologies), filtering out reads with a quality cutoff > Q7. Base called reads were analyzed with NanoPlot v1.18.2 (De Coster et al., 2018). Nanopore reads properties and accession numbers were reported in Supplementary Table S1.

3.5. Genome assembly and annotation

Nanopore reads were filtered using Filtlong v0.2.0 (https://github.com/rrwick/Filtlong) initially with the parameter --*min_length* 80000 to remove reads shorter than 80 kb, then with --*target_bases* 253000000 to obtain a 110x coverage of a 2.3 Mbp genome size. Filtered Nanopore reads were assembled with Unicycler v0.4.7 tool (Wick et al., 2017). The genome sequence was polished with Medaka v0.7.1 software (https://github.com/Nanoporetech/medaka) using the Nanopore reads longer than 80 kb. Two additional polishing rounds were carried out with the Pilon v1.22 tool using the Illumina reads (Walker et al., 2014). Assembly completeness was assessed with the Bandage v.0.8.1 tool (Wick et al., 2015), whereas assembly quality was evaluated with both Ideel (https://github.com/mw55309/ideel) and CheckM v1.1.3 tools (Parks et al., 2015). BWA v0.7.17 (Li, 2013) and minimap2 v2.13 (Li, 2018) programs were used to align the Illumina reads and the Nanopore reads to the assembled genome, respectively. Alignments were visualized with the Tablet v1.17.08.17 tool (Milne et al., 2013). Coverage graph obtained by genome mapping with all Nanopore reads longer than 80 kb, was used to investigate and validate the genomic structure. Location of the long inverted repeated DNA regions was further confirmed by PCR genome mapping. Genome was automatically annotated with NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.10 (Tatusova et al., 2016). Manual gene annotation of the DNA sequences of interest was carried out by BLAST homology searching of the databases available at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/sites/gquery). Protein domains were identified using the protein family database Pfam (https://pfam.xfam.org). Default parameters were used for all software unless otherwise specified.

3.6. Genome analysis

Chromosomal structural variants were investigated using the Sniffles v1.0.12 structural variation caller (Sedlazeck et al., 2018) and the npInv v.1.24 tool (Shao et al., 2018) for non-allelic homologous recombination mediated genomic inversion, then visualized with the integrative genomics viewer visualization tool (Robinson et al., 2011). Suggested alternative chromosomal assembly structures of M247_Siena were designed in silico and aligned to Nanopore reads longer than 80 kb which account for a 363x genome coverage (Supplementary Table S1). Aligned reads were processed with samclip v.0.4.0 (https://github.com/tseemann/samclip) with parameter *--max 100* for soft and hard clipping and then filtered with SAMtools (Li et al., 2009) using the command *view* with the parameter *-b* specifying the chromosomal region of interest maintaining only reads spanning the repeated regions that are useful to support the different chromosomal structures. The previously described *L. crispatus* M247 genome (Colombini et al., unpublished, GenBank accession no. CP088015) was used for genome comparison. Furthermore, the publicly available *L. crispatus* complete genomes in the NCBI Microbial Genome Database (https://www.ncbi.nlm.nih.gov/genome/genomes/1815/) at the moment of this study (December 2021) namely PRL2021 (CP058996), FDAARGOS_743 (CP046311), AB70 (CP026503), 1D (CP047415), CO3MRSI1 (CP033426), C25 (CP047142), B4 (CP059140), DC21.1 (CP039266), ATCC 33820 (CP072197), Lc1226 (CP083392), Lc1700 (CP083389), Lc116 (CP083393), PMC201 (CP076522), 2029 (CP079206), were downloaded and used for comparison. Genome comparison was obtained with the following tools: (i) Mauve (Darling et al., 2010); (ii) Blast https://blast.ncbi.nlm.nih.gov/Blast.cgi); (iii) Artemis and Artemis Comparison Tool (ACT) v17.0.1 (Carver et al., 2008); (iv) MUMmer v3.23 (Marçais et al., 2018).

3.7. Nucleotide sequence accession numbers

The complete genome sequence of *L. crispatus* M247_Siena is available under GenBank accession no. CP046589, whereas Nanopore and Illumina sequencing reads are available under Sequence Read Archive accession no. SRR10902282 and SRR10902283, respectively. The BioProject number is PRJNA594001.

3.8. PCR

PCR reactions were carried out following an already described protocol (Iannelli et al., 1998; Santoro et al., 2010). Long and short ranges PCR were performed to validate both the canonical and the rearranged genomic structures. Oligonucleotide primers are listed in Table 1.

3.9. Quantitative Real-Time PCR

Real-time PCR experiments were carried out with the KAPA SYBR FAST qPCR kit Master Mix Universal (2X) (Merck KGaA, Darmstadt, Germany) on a LightCycler 1.5 apparatus (Roche Diagnostics GmbH, Mannheim, Germany). Real-time PCR mixture contained, in a final volume of 20 µl, 1× KAPA SYBR FAST qPCR reaction mix, 5 pmol of each primer and 2 µl (20 ng) of bacterial gDNA as starting template. Thermal profile was an initial 4 min denaturation step at 95°C followed by 40 cycles of repeated denaturation (10 s at 95°C), annealing (15 s at 60°C), and polymerization (3 min at 72°C). The temperature transition rate was 20°C/s in the denaturation and annealing step and 5°C/s in the polymerization step. A standard curve for the *gyrB* gene of *L. crispatus* M247 was built plotting the threshold cycle against the number of chromosome copies using serial dilutions of chromosomal DNA with known concentration. This external standard curve was used to quantify in each sample the number of rearranged chromosomal structures. Primer pairs IF1118/IF1119 and IF111/IF1121 were used to detect and quantify the left and right DNA junctions of the rearranged 69.9-kb region, respectively, whereas IF1120/IF1122 and IF1110/1117 were used for left and right DNA junction of the rearranged 15.4-kb region, respectively. The frequency of M247 rearranged chromosomal structures was calculated as the average of the DNA junction frequencies. Primers are listed in Table 1 and primers pairs used for chromosomal rearrangements detection are represented in Figure 3. Melting curve analysis was performed to differentiate the amplified products from primer dimers. Electrophoresis gel run was performed to further verified the amplification products.

4. RESULTS

4.1. An unusual 69.9-kb long inverted repeat in the genome of the *L. crispatus* laboratory strain M247_Siena

Genome sequencing was performed on a *L. crispatus* laboratory strain carried in our laboratory for over 20 years, which we named M247_Siena to differentiate from the previously sequenced strain M247 (Colombini et al., unpublished, GenBank accession no. CP088015). The complete genome sequence was obtained combining Illumina sequencing reads with very long Nanopore sequencing reads (>80 kb) which were necessary to resolve the genome complexity of this laboratory strain. Sequence analysis showed that the M247_Siena genome is organized in one circular chromosome of 2,385,061 base pairs (bp) in length, containing 2,359 open reading frames (ORFs) with an average GC content of 37.07%. Compared to the previously sequenced strain M247, M247_Siena resulted 48,935-bp longer and characterized by an unusual duplication of a 69.9-kb long chromosomal segment producing two long inverted repeats (LIRs), 69,925 and 69,919 bp in length, located 38,619 bp downstream (LIR1) and 185,826 bp upstream (LIR2) of the chromosomal origin of replication, respectively (Figure 1). Interestingly, the 69.9-kb duplicated DNA sequence (LIR2) of M247_Siena was found i) replacing a 15,482-bp DNA sequence spanning nucleotides (nt) 2,133,454 to 2,148,935 in the original M247 genome, and ii) associated to the inversion of the 224,443-bp DNA region flanked by LIR1 and LIR2 and containing the replication origin (Figure 1). The absence of the 69.9-kb duplication and of the 15.4-bp region from strains M247 and M247_Siena, respectively, was further verified and confirmed by genome mapping with Nanopore reads (Supplementary Figure S1) and by PCR analysis.

4.2. Nucleotide sequence of the M247_Siena 69.9-kb long inverted repeat and the associated 15.4-kb deletion

Each copy of the 69.9-kb repeats of M247_Siena contains 72 ORFs, of which 46 have the same direction of transcription. By manual homology-based annotation, it was possible to attribute a putative function to the hypothetical gene product of 59 ORFs (Table 2). Hypothetical gene products were used to search public protein databases and the Pfam protein family database, taking into account significant homologies with functionally characterized proteins or good matches with Pfam domains. Annotated ORFs include 18 ORFs encoding for metabolic enzymes, 8 for genetic information processing factors, 18 for proteins involved in signaling and cellular processes, 8 for ribosomal proteins and transfer RNA and 7 for insertion sequences (ISs) (Table 2). Each repeat contains at the 5' end a copy of IS*1201* and at the 3' end a copy of IS*Lcr2* (Figure 2). Sequence comparison showed that the 2 LIRs contain 4 nucleotides changes and a 6-bp insertion all located in the 5' end copy of IS*1201*. Similarly, the 15,482-bp region is bounded at the 5' end by a copy of IS*1201* and at the 3' end by a copy of IS*Lcr2* (Figure 2). Therefore, these two ISs occur in both the M247_Siena and the M247 genomes in the form of inverted repeats (Figure 2). Manual homology-based annotation with functional prediction of the hypothetical gene product was possible for 15 out of the 17 ORFs contained in the 15.4-kb deleted region, whereas 2 ORFs encoded hypothetical proteins that showed no homology to other characterized sequences (Table 3). The GC content of the 69.9-kb and of the 15.4-kb DNA regions were 38.4% and 37.2%, respectively, comparable to the overall genome GC content of both M247 and M247_Siena strains (37.04 and 37.07%, respectively).

4.3. Strain M247 shows chromosomal rearrangements involving both the 69.9-kb and the 15.4-kb DNA regions

PCR analysis was conducted in the M247 original strain to investigate the presence of putative chromosomal rearrangements underlying the origin of the 69.9-kb repeat and of the 15.4-kb deletion. Divergent primers were designed on the ends of both the 69.9-kb and 15.4-kb DNA segments, whereas convergent primers were designed on the flanking chromosomal regions (Figure 3). PCR results indicated that the M247 genome is subject to chromosomal rearrangements leading to the exchange of the chromosomal position occupied by the 69.9-kb and the 15.4-kb DNA segments. The recombination events driving this type of rearrangements seem to occur at the level of the ISs, namely IS*1201* and IS*Lcr2* that flank both DNA segments (Figure 3). To obtain a quantitative estimate of bacterial chromosomes carrying the rearrangements, real-time PCR was used to quantify concentration of chromosomes in which the 69.9-kb DNA segment was exchanged with the 15.4-kb DNA region. Chromosomal rearrangements showed a low frequency rate of approximately 2.19±0.69 per 10^5 chromosomes (Table 4).

4.4. The presence of the 69.9-kb long inverted repeat favors chromosomal rearrangements in strain M247_Siena

We also evaluate the potential of the 69.9-kb repeats to generate chromosomal rearrangement within the M247_Siena strain. Genome structural variants were investigated with the Sniffles v1.0.12 structural variation caller (Sedlazeck et al., 2018) and with the npInv v.1.24 (Shao et al., 2018) tool for non-allelic homologous recombination mediated genomic inversions detection, using Nanopore reads longer than 80 kb which accounted for a 363x genome coverage (Supplementary Table S1). Genome structure analysis revealed the presence of two different chromosomal structures also for the M247_Siena characterized by the inversion of the 224,443-bp DNA region containing the *oriC* and flanked by the LIRs. A semi-quantitative estimation, obtained by reads counting, revealed that 118 out of 480 repeats-spanning reads (23.3%) contain the alternative inverted structure (Figure 4).

4.5. Chromosomal structure analysis of the other *L. crispatus* complete genomes

When we investigated the presence of the 69.9-kb repeats in the 14 *L. crispatus* complete genomes available in the NCBI database (December 2021) we found that a DNA fragment homologous to the LIR1 of M247_Siena is always present. No DNA fragment homologous to LIR2 was found, whereas the 15.4-kb region is always present. In strain Lc1226 the LIR1-homologous region and the 15.4-kb region are flanked by IS*1201* and IS*Lcr2* elements, located at the 5' end at the 3' end, respectively, and arranged in the same way as in the M247 and M247_Siena genomes. In the other strains one or both ISs elements are deleted or differentially arranged. Furthermore, whole genome alignment of the 14 *L. crispatus* complete genomes using M247_Siena as reference indicated that in all the other strains the 224,443-bp DNA region containing the *oriC* is arranged as in the less represented M247_Siena chromosomal structure (Figure 5). Some strains show further inversion: CO3MRSI1 displays an inversion of 191-kb around the origin of replication, whereas AB70, FDAARGOS, DC21.1 and 2029 have an additional DNA inversion of 610-kb, 619-kb, 1,3 Mb and 1,7 Mb in length respectively occurring in proximity of the putative termination sites.

5. DISCUSSION

In this work we reported the unusual genomic structure of a *L. crispatus* laboratory strain, named M247_Siena, which has been carried in our laboratory for over 20 years. The peculiarity of the M247_Siena genome is the presence of a duplication of a 69.9-kb DNA segment producing two copies of an inverted repeat, located about 224-kb apart (Figure 1). These types of large structural variations are thought to have mostly remained undetected with short reads next generation sequencing technologies (Chaisson et al., 2015, 2019; De Coster et al., 2019; Huddleston et al., 2017; Pendleton et al., 2015). However, the formation of very long inverted repeats in bacterial genomes is still considered to be an extremely rare event; it has been reported that for over 9600 prokaryotic genomes, only a small subset of strains (3%) harbor long near identical repeats above 30 kb in their genome (Schmid et al., 2018). The events involved in the generation of these long inverted repeated regions are not yet fully understood, but have been hypothesized to occur in the context of bacterial chromosomal rescue during double strand break repairing (El Kafsi et al., 2017). Genome comparison of the laboratory strain M247_Siena with the original M247 strain highlighted a 15.4-kb DNA region which was lost and replaced by the 69.9-kb repeat (Figure 1). Nucleotide sequence analysis of both 15.4- and 69.9-kb regions excluded the hypothesis of a mobile DNA origin because no *int/ xis* genes were found (Rocco & Churchward, 2006; Rudy et al., 1997) and GC content was comparable with the rest of the genome (Bohlin et al., 2017; Sueoka, 1962). However, both regions were flanked by the same ISs elements namely IS*Lcr2* and IS*1201* (Figure 2). ISs have been described as source of genomic instability within bacterial genome and also reported to induce large chromosomal inversion (Daveran-Mingot et al., 1998; Lee et al., 2016). PCR analysis revealed the presence of chromosomal rearrangements in the original M247 population, involving the exchange of the 69.9-kb and the 15.4-kb regions on the chromosome, thus suggesting that genome-scale rearrangements occur spontaneously (Figure 3). Real-time PCR quantification indicated that these types of chromosomal rearrangements were rare events occurring approximately in 2 out of 100,000 bacterial cells. The 69.9-kb LIRs of M247_Siena generate large regions of homology that may allow intrachromosomal homologous recombination during the DNA replication. Intrachromosomal homologous recombination can lead to deletions, duplications, translocations (for direct repeats), and inversions (for inverted repeats) (Achaz et al., 2003; Romero & Palacios, 1997; Roth et al., 1996; Smith, 1988). Long inverted repeats are reported to transiently stall DNA replication by forming hairpin structures on both the leading and lagging strands (Lai et al., 2016; Leach, 1994), contributing to the formation of a "X-shaped" symmetrical rearrangement involving the origin or terminus of replication (Eisen et al., 2000; Guinane et al., 2011; Suyama & Bork, 2001; Tillier & Collins, 2000). Indeed, genome mapping with Nanopore reads showed the presence of two alternative chromosomal structures in the M247_Siena population characterized by a 224,443-bp inversion of the region flanked by the LIRs and containing the origin of replication (Figure 4). The M247_Siena inversion does not alter significantly the replichores length, as observed for previously described large chromosomal inversions (Eisen et al., 2000; Tillier & Collins, 2000). A semi-quantitative estimation by sequencing reads counting, showed that 118 out of 480 (23.3%) repeats-spanning reads contain the inverted structure (Figure 4) suggesting that the chromosomal structure is rearranged approximately in 1 out of 5 bacterial cells. Quantitative PCR performed on M247 DNA showed that the frequencies of chromosomal rearrangements were significantly lower (2 out of 10^5 cells), thus suggesting that the newly generated long inverted repeats of M247_Siena increased the intrinsic genomic instability of the strain. Whole genome alignment with other available *L. crispatus* complete genomes revealed that i) no *L. crispatus* complete genome contains a repeated DNA sequence like M247_Siena, ii) in each genome a 69.9-kb and a 15.4-kb DNA homologous sequences are present as in M247 and that iii) the chromosomal structures of M247 and of the other *L. crispatus* strains examined resemble the M247_Siena structure containing the 224.4-kb inversion, which is the less represented form of M247_Siena (Figure 5). Furthermore, genomes alignment highlights the presence of chromosomal rearrangements also in other *L. crispatus* strains suggesting that genomic instability is shared within the *L. crispatus* species, although it is not known if ISs elements or other causative factors are involved.

In conclusion, we identified chromosomal rearrangements associated with the presence of insertion sequences in the original M247 strain which might explain the origin of the unusual 69.9-kb duplication in the laboratory strain M247_Siena. In addition, we have determined that the presence of such structural variant in the M247_Siena genome increases the genomic instability intrinsic of strain M247.

TABLES

Table 1. **Oligonucleotide primers.**

Name	Sequence (5' to 3')	Position on M247 chromosome (GenBank ID: CP088015)
IF1110	CTGGAATTATATTTATCTCTCGTA	38,489 - 38,512
IF1111	AAAAGGTATAGCAAAACGTACTT	41,626 - 41,604
IF1118	GGTAGTCCTTATCATAAGTAGAA	106,610 - 106,632
IF1119	TACCTGTAGTAGAAATCCAGTTA	2,132,588 - 2,132,610
IF1120	ACGATTCAGTGGTAGAAATACTA	108,608 - 108,586
IF1121	TAAACTGAGGATGCTATACTGA	2,149,162 - 2,149,141
IF1123	TTATGGATACTTACGAAGACGAA	74,750 - 74,728
IF1124	TTGTCGAATGCCTTAACAACCAT	74,576 - 74,598
IF1125	ATATTGATGCCAACGAGGTTAT	55,104 - 55,083
IF1126	AACTGCGTTTGTTTGGCACTAT	92,970 - 92,991
IF1502	GTTACCTCCTACGGATATTCAT	38,404 - 38,425
IF1503	TTGTGAAGTAGAGTTTAACCGAA	55,247 - 55,225
IF1504	GCACGCTCGTCATCAGTATA	92,919 - 92,938
IF1505	GTATTGAGTCTGGATCGGATT	108,705 - 108,685
IF1506	GTTGTAAAGTATGTGTTCTCAAG	2,132,825 – 2,132,847
IF1507	AGGAAGAGGACTAGGTAAGAA	2,141,741 - 2,141,721
IF1508	AGAGGACTAGGTAAGAAGACAT	2,141,737 - 2,141,716
IF1509	CACTGCTCCAGAAATGATACAA	2,140,801 - 2,140,822

***Table 2.* Annotated ORFs of the 69.9-kb long inverted repeated DNA sequence of M247_Siena.**

			Homologous protein			
ORF (aa)[a]	**Annotation and comments (reference)**	**Pfam domain[b] [E value]**	**Protein ID /Origin [E value][c]**	**aa identity**	**aa similarity**	**Functional group**
orf1 (392)	IS*1201*, transposase, IS*256* family (Tailliez et al., 1994)		L26311 *Lactobacillus helveticus.* [0.0]	332/368 (90%)	351/368 (95%)	
orf2 (469)	IS*L5*, transposase, IS*4* ssgr *ISPepr1* family (Lapierre et al., 2002)		AY040218 *Lactobacillus delbrueckii* [7e-156]	212/326 (65%)	260/326 (79%)	
orf6 (392)	IS*1201*, transposase, IS*256* family (Tailliez et al., 1994)		L26311 *Lactobacillus helveticus.* [0.0]	333/368 (90%)	352/368 (95%)	
orf7 (200)	Transcriptional regulator, truncated, putative (Brennan & Matthews, 1989)	HTH_19 (4-65) [5.8e-11]				genetic information processing factors
orf8 (270)	Transcriptional regulator, putative (Brennan & Matthews, 1989)	HTH_19 (4-67) [4.9e-07]				genetic information processing factors
orf10 (117)	Protease, truncated, putative					metabolic enzymes
orf11 (285)	IS*Lhe5*, transposase, IS*982* family (Callanan et al., 2008)		NC_010080 *Lactobacillus helveticus* [5e-168]	241/285 (85%)	259/285 (90%)	

			Homologous protein			
ORF (aa)[a]	**Annotation and comments (reference)**	**Pfam domain[b] [E value]**	**Protein ID /Origin [E value][c]**	**aa identity**	**aa similarity**	**Functional group**
orf12 (225)	Protease, truncated, putative					metabolic enzymes
orf13 (130)	CPBP family intramembrane metalloprotease (Pei et al., 2011)	CPBP (1-74) [1.5e-09]				metabolic enzymes
orf14 (66)	Transcriptional regulator, putative (Brennan & Matthews, 1989)	HTH_3 (4-58) [6.7e-16]				genetic information processing factors
orf16 (379)	Amino acid permease (Weber et al., 1988)	AA_permease_2 (10-377) [2.0e-22]				signaling and cellular processes
orf18 (257)	ABC transporter permease, putative (Rafii & Park, 2008)	ABC_trans_CmpB (10-164) [1.1e-44]	RYQ23281 *Bifidobacterium pseudolongum* [2e-24]	42/135 (31%)	72/135 (53%)	signaling and cellular processes
orf19 (54)	ATP:cob(I)alamin adenosyltransferase (Mera et al., 2009)	Cob_adeno_trans (1-33) [1.7e-06]				metabolic enzymes
orf22 (300)	YSIRK-type signal peptide-containing protein	YSIRK_signal(3-28) [1.6e-08]				signaling and cellular processes
orf23 (87)	YSIRK-type signal peptide-containing protein	Mub_B2 (1-63) [5.8e-10]				signaling and cellular processes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein			Functional group
			Protein ID /Origin [E value][c]	aa identity	aa similarity	
orf24 (189)	LPXTG-motif protein cell wall anchor domain protein, partial	Gram_pos_anchor (146-189) [3.3e-07]				signaling and cellular processes
orf25 (370)	IS*Lhe2*, transposase, IS*L3* family (Callanan et al., 2008)		CP000517 *L. helveticus.* [0.0]	282/363 (78%)	295/363 (81%)	
orf26 (586)	Oligopeptide ABC transporter substrate-binding protein (Kempf & Bremer, 1995)	SBP_bac_5 (100-509) [3.7e-62]				signaling and cellular processes
orf27 (344)	AI-2E family transporter (Herzberg et al., 2006)	AI-2E_transport (14-333) [5.3e-42]				signaling and cellular processes
orf28 (190)	Hydrolase SGNH/GDSL family (Upton & Buckley, 1995)	Lipase_GDSL (3-183) [2.1e-18]				metabolic enzymes
orf30 (444)	Solute carrier, family 45 major facilitator superfamily transporter (Pao et al., 1998)	MFS_1 (11-403) [5.6e-24]	PRO95402 *Lactiplantibacillus pentosus* [4e-167]	232/436 (53%)	310/436 (71%)	signaling and cellular processes
orf31 (253)	ATP-binding cassette domain-containing protein (Hung et al., 1998)	ABC_tran (25-180) [6.6e-29]				signaling and cellular processes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein Protein ID /Origin [E value][c]	aa identity	aa similarity	Functional group
orf32 (298)	ABC transporter permease (Woodson & Devine, 1994)	BPD_transp_2 (3-277) [2.1e-31]				signaling and cellular processes
orf33 (330)	ABC transporter substrate-binding protein (Hung et al., 1998)	ABC_sub_bind (34-325) [4.0e-98]				signaling and cellular processes
orf34 (237)	Noncanonical pyrimidine nucleotidase, YjjG family	HAD_2 (8-202) [4.5e-24]				genetic information processing factors
orf36 (285)	Aldo/keto reductase (Bohren et al., 1989)	Aldo_ket_red (20-271) [1.7e-49]				metabolic enzymes
orf37 (641)	Tetracycline resistance ribosomal protection protein (Lépine et al., 1993)	GTP_EFTU (2-224) [1.5e-49] EFG_C (509-594) [2.2e-19] EFG_IV (388-502) [5.6e-18] EFG_II (315-387) [5.7e-09]	Q08425 *Bacteroides fragilis* [9e-92]	185/616 (30%)	305/616 (49%)	signaling and cellular processes
orf38 (272)	Haloacid dehalogenase-like family hydrolase (Koonin & Tatusov, 1994)	Hydrolase_3 (9-265) [6.9e-56]				metabolic enzymes
orf39 (389)	Major facilitator superfamily transporter (Pao et al., 1998)	MFS_1 (7-275) [2.9e-16]				signaling and cellular processes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein Protein ID /Origin [E value][c]	aa identity	aa similarity	Functional group
orf40 (337)	D-lactate dehydrogenase (Dengler et al., 1997)	2-Hacid_dh_C (112-299) [1.8e50] 2-Hacid_dh (4-331) [8.0e-32]	2DLD_A *L. helveticus.* [0.0]	324/337 (96%)	335/337 (99%)	metabolic enzymes
orf41 (83)	YSIRK-type signal peptide-containing protein	YSIRK_signal (6-31) [1.1e-12]				signaling and cellular processes
orf42 (179)	LPXTG-motif cell wall anchor domain protein	Gram_pos_anchor (138-178) [5.8e-08]				signaling and cellular processes
orf43 (275)	Exodeoxyribonuclease III (Mol et al., 1995)	Exo_endo_phos (4-266) [1.1e-13]				genetic information processing factors
orf44 (480)	Amino acid permease (Weber et al., 1988)	AA_permease_2 (20-462) [1.6e-58]				signaling and cellular processes
orf45 (942)	DEAD/DEAH box helicase, putative (Tanner & Linder, 2001)	ResIII (204-356) [4.7e-31] PLDc_2 (41-170) [1.7e-13] Helicase_C (419-529) [4.7e-11]	AJT50415 *L. mucosae* LM1 [0.0]	438/977 (45%)	608/977 (62%)	genetic information processing factors
orf46 (387)	Serine hydrolase (Neu, 1969)	Beta-lactamase (38-363) [2.3e-30]				metabolic enzymes
orf48 (137)	Peptide deformylase (Meinnel et al., 1996)	Pep_deformylase (4-137) [7.3e-23]				metabolic enzymes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein			Functional group
			Protein ID /Origin [E value][c]	aa identity	aa similarity	
orf49	16S ribosomal RNA					ribosomal proteins and transfer RNA
orf50	tRNA-Ile					ribosomal proteins and transfer RNA
orf51	tRNA-Ala					ribosomal proteins and transfer RNA
orf52	23S ribosomal RNA					ribosomal proteins and transfer RNA
orf53	5S ribosomal RNA					ribosomal proteins and transfer RNA
orf54	tRNA-Asn					ribosomal proteins and transfer RNA
orf55 (73)	Steroid-binding protein (Lederer, 1994)	Cyt-b5 (5-69) [2.6e-07]				genetic information processing factors
orf56 (289)	1-acyl-sn-glycerol-3-phosphate acyltransferase (Lu et al., 2006)	Acyltransferase (77-210) [1.9e-08]				metabolic enzymes
orf57 (315)	Glycosyltransferase family 8 protein (Campbell et al., 1997)	Glyco_transf_8 (4-248) [4.1e-42]				metabolic enzymes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein Protein ID /Origin [E value][c]	 aa identity	 aa similarity	Functional group
orf58 (285)	Glycosyltransferase family 8 protein (Campbell et al., 1997)	Glyco_transf_8 (1-218) [3.1e-34]				metabolic enzymes
orf59 (392)	IS*1201*, transposase, IS*256* family (Tailliez et al., 1994)	Transposase_mut (3-374) [1.5e-107]	L26311 *Lactobacillus helveticus.* [0.0]	333/368 (90%)	352/368 (95%)	
orf60 (274)	Glycosyl transferase (Campbell et al., 1997)	Glyco_transf_8 (2-236) [4.5e-16]				metabolic enzymes
orf61 (254)	1-acyl-sn-glycerol-3-phosphate acyltransferase, putative (Lu et al., 2006)					metabolic enzymes
orf62	tRNA-Lys					ribosomal proteins and transfer RNA
orf64	tRNA-Lys					ribosomal proteins and transfer RNA
orf65 (238)	Response regulator transcription factor, YycF/WalR (Türck & Bierbaum, 2012)	Response_reg (5-114) [3.8e-30] Trans_reg_C (156-232) [2.2e-26]	ABD29209 *S. aureus* [3e-118]	155/234 (66%)	193/234 (82%)	genetic information processing factors
orf66 (617)	Cell wall metabolism sensor histidine kinase	HATPase_c (495-607) [3.7e-30] HisKA (376-443) [8.6e-19] HAMP	Q2G2U4 *S. aureus* [1e-169]	259/623 (42%)	395/623 (63%)	signaling and cellular processes

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein Protein ID /Origin [E value][c]	 aa identity	 aa similarity	Functional group
	YycG/WalK (Türck & Bierbaum, 2012)	(201-253) [1.6e-14] PAS (263-367) [2.6e-06]				
orf68 (274)	Auxiliary regulator of two-component sysem, YycI family (Türck & Bierbaum, 2012)	YycI (32-260) [4.6e-56]	YP_004888125 *L. plantarum* WCFS1 [4e-49]	87/272 (32%)	153/272 (56%)	signaling and cellular processes
orf69 (265)	Metallo-beta-lactamase fold metallo-hydrolase (Carfi et al., 1995)	Lactamase_B_2 (22-219) [1.4e-25]				metabolic enzymes
orf70 (420)	Serine protease (Yan et al., 1998)	Trypsin_2 (138-279) [2.0e-30] PDZ_2 (321-417) [1.1e-22]				metabolic enzymes
orf71 (159)	23S rRNA (pseudouridine(1915)-N(3))-methyltransferase RlmH (Tkaczuk et al., 2007)	SPOUT_MTase (1-158) [1.7e-55]				metabolic enzymes
orf72 (469)	IS*Lcr2*, transposase, IS*Lre2* family, partial		NZ_GL531739 *L. crispatus* [0.0]	289/301 (96%)	290/301 (96%)	

---

[a] The number of amino acids of each ORF is shown in parenthesis.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

***Table 3.* Annotated ORFs of the 15.4-kb DNA sequence of M247, found deleted and replaced by LIR2 in the M247_Siena genome.**

			Homologous protein			
ORF (aa)[a]	**Annotation and comments (reference)**	**Pfam domain [E value][b]**	**Protein ID /Origin [E value][c]**	**aa identity**	**aa similarity**	**Functional group**
orf1 (392)	IS*Lcr2*, transposase, IS*Lre2* family, partial		NZ_GL531739 *L. crispatus* [0.0]	289/301 (96%)	290/301 (96%)	
orf2 (470)	C69 family peptidase (Vesanto et al., 1996)	Peptidase_C69 (9-399) [9.2e-129]	CAA86210 *L. helveticus* [2e-89]	159/475 (33%)	253/475 (53%)	metabolism
orf3 (224)	ABC transporter ATP-binding protein	AAA_16 (22-199) [9.9e-05]				signaling and cellular processes
orf4 (366)	ABC transporter permease, pseudogene					signaling and cellular processes
orf5 (174)	TetR/AcrR family transcriptional regulator	TetR_N (22-54) [7.5e-05]				signaling and cellular processes
orf6 (199)	Histidine phosphatase family protein	His_Phos_1 (3-195) [7.1e-48]				
orf7 (251)	Protease, truncated, putative	Peptidase_S9 (51-247) [0.00027]				metabolism
orf8 (253)	Peroxide stress protein YaaA (Liu et al., 2011)	H2O2_YaaD (1-234) [8.3e-74]				
orf9 (214)	Alpha-fucosidase, putative	Alpha_L_fucos (12-113) [0.068]				

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain [E value][b]	Homologous protein Protein ID /Origin [E value][c]	 aa identity	 aa similarity	Functional group
orf11 (344)	IS*Ldl3*, transposase, IS*30* family (Ravin & Alatossava, 2003)		AJ316615 *L. delbrueckii.* [2e-106]	165/345 (48%)	223/345 (64%)	
orf12 (400)	L, D-transpeptidase family protein (Bielnicki et al., 2005)	YkuD (274-399) [2.4e-16] PG_binding_4 (125-246) [7.5e-06]	SDA61766 *L. kefiranofaciens* [7e-161]	216/398 (54%)	297/398 (74%)	
orf14 (251)	Membrane protein					
orf15 (150)	LytTR family transcriptional regulator (Nikolskaya, 2002)	LytTR (52-144) [1.2e-21]				
orf17 (390)	IS*1201*, transposase, IS*256* family (Tailliez et al., 1994)		L26311 *Lactobacillus helveticus.* [0.0]	332/368 (90%)	351/368 (95%)	

---

[a] The number of amino acids of each ORF is shown in parenthesis.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

Table 4. **Real-time PCR quantification of M247 chromosomes carrying chromosomal rearrangements[a].**

Primer pair	**Chromosomes with rearrangements**
IF1118 - IF1119	6.47×10^{-6} ($\pm 8.10 \times 10^{-7}$)
IF1120 - IF1122	3.46×10^{-5} ($\pm 4.01 \times 10^{-8}$)
IF1111 - IF1121	2.14×10^{-6} ($\pm 1.57 \times 10^{-6}$)
IF1110 - IF1117	4.44×10^{-5} ($\pm 3.71 \times 10^{-7}$)
Average	2.19×10^{-5} ($\pm 0.69 \times 10^{-6}$)

[a] Frequency is expressed as number of rearranged chromosomal structures per total number of chromosomes.

Supplementary Table S1. **General statistics of the M247_Siena Nanopore and Illumina reads.**

	Nanopore reads[a] (SRR10902282)			**Illumina reads[b] (SRR10902283)**	
	Overall (2,230x)	**>80 kb (363x)**	**>80 kb (110x)**	**R1**	**R2**
Reads (n)	335,898	8,280	2,425	280,486	280,486
Mean read length	15,275.0	104,794.3	104,383.0	225.3	213.9
Median read length	6,448.0	97,512.5	97,400.0	251.0	251.0
Read length N50[c]	36,990	102,516	102,207	251	251
Mean read quality (Q)[d]	12.1	12.1	14.0	33.1	29.1
Median read quality (Q) [d]	12.4	12.3	14.0	36.0	29.0
Sequencing output (no of bases)	5,130,846,027	867,696,545	253,128,657	63,191,714	60,009,977

[a] The overall Nanopore reads were filtered for length >80 kb and the output was further filtered to obtain a 110x coverage of a 2.3 Mbp genome size.

[b] R1 and R2 refer to illumina reads, forward and reverse, respectively.

[c] N50 is the length of a sequence in a set for which all sequences of that length or greater sum to 50% of the set's total size.

[d] Phred quality score Q expresses the confidence in a particular base-call and is logarithmically related to the base-calling error probability P ($Q= -10 \, _{\log 10} P$).

Acknowledgements

Illumina genome sequencing was provided by MicrobesNG.

6. REFERENCES

Achaz, G., Coissac, E., Netter, P., & Rocha, E. P. C. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics*, *164*(4), 1279–1289. https://doi.org/10.1093/genetics/164.4.1279.

Bielnicki, J., Devedjiev, Y., Derewenda, U., Dauter, Z., Joachimiak, A., & Derewenda, Z. S. (2005). *B. subtilis* ykuD protein at 2.0 A resolution: Insights into the structure and function of a novel, ubiquitous family of bacterial enzymes. *Proteins: Structure, Function, and Bioinformatics*, *62*(1), 144–151. https://doi.org/10.1002/prot.20702.

Bohlin, J., Eldholm, V., Pettersson, J. H. O., Brynildsrud, O., & Snipen, L. (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics*, *18*(1), 151. https://doi.org/10.1186/s12864-017-3543-7.

Bohren, K. M., Bullock, B., Wermuth, B., & Gabbay, K. H. (1989). The aldo-keto reductase superfamily. CDNAs and deduced amino acid sequences of human aldehyde and aldose reductases. *The Journal of Biological Chemistry*, *264*(16), 9547–9551.

Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif. *Journal of Biological Chemistry*, *264*(4), 1903–1906. https://doi.org/10.1016/S0021-9258(18)94115-3.

Callanan, M., Kaleta, P., O'Callaghan, J., O'Sullivan, O., Jordan, K., McAuliffe, O., Sangrador-Vegas, A., Slattery, L., Fitzgerald, G. F., Beresford, T., & Ross, R. P. (2008). Genome Sequence of *Lactobacillus helveticus* , an Organism Distinguished by Selective Gene Loss and Insertion Sequence Element Expansion. *Journal of Bacteriology*, *190*(2), 727–735. https://doi.org/10.1128/JB.01295-07.

Campbell, J. A., Davies, G. J., Bulone, V., & Henrissat, B. (1997). A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *The Biochemical Journal*, *326 (Pt 3)*(Pt 3), 929–939. https://doi.org/10.1042/bj3260929u.

Carfi, A., Pares, S., Duee, E., Galleni', M., Duez, C., Frere, J. M., & Dideberg, O. (1995). The 3-D structure of a zinc metallo-1-lactamase from *Bacillus cereus* reveals a new type of protein fold. *EMBO J.* 1995, 16;14(20):4914-21. PMID: 7588620; PMCID: PMC394593.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., & Rajandream, M.-A. (2008). Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, *24*(23), 2672–2676. https://doi.org/10.1093/bioinformatics/btn529.

Cesena, C., Morelli, L., Alander, M., Siljander, T., Tuomola, E., Salminen, S., Mattila-Sandholm, T., Vilpponen-Salmela, T., & von Wright, A. (2001). *Lactobacillus crispatus* and its nonaggregating mutant in human colonization trials. *Journal of Dairy Science*, *84*(5), 1001–1010. https://doi.org/10.3168/jds.S0022-0302(01)74559-6.

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907.

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, *5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147.

Darling, A. E., Miklós, I., & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genetics*, *4*(7), e1000128. https://doi.org/10.1371/journal.pgen.1000128

Darmon, E., & Leach, D. R. F. (2014). Bacterial genome instability. *Microbiology and Molecular Biology Reviews*, *78*(1), 1–39. https://doi.org/10.1128/MMBR.00035-13.

Daveran-Mingot, M.-L., Campo, N., Ritzenthaler, P., & Le Bourgeois, P. (1998). A natural large chromosomal inversion in *Lactococcus lactis* is mediated by homologous recombination between two insertion sequences. *Journal of Bacteriology*, *180*(18), 4834–4842. https://doi.org/10.1128/JB.180.18.4834-4842.1998.

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*(7), 1178–1187. https://doi.org/10.1101/gr.244939.118.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics (Oxford, England)*, *34*(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.

Dengler, U., Niefind, K., & Kie, M. (1997). Crystal structure of a ternary complex of D-2-hydroxy-isocaproate dehydrogenase from *Lactobacillus casei,* NAD+ and 2-oxoisocaproate at 1.9 A Resolution. *J Mol Biol*;267(3):640-60. doi: 10.1006/jmbi.1996.0864. PMID: 9126843.

Eisen, J. A., Heidelberg, J. F., White, O., & Salzberg, S. L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *E. Coli*, 9.

El Kafsi, H., Loux, V., Mariadassou, M., Blin, C., Chiapello, H., Abraham, A.-L., Maguin, E., & van de Guchte, M. (2017). Unprecedented large inverted repeats at the replication terminus of circular bacterial chromosomes suggest a novel mode of chromosome rescue. *Scientific Reports*, *7*(1), 44331. https://doi.org/10.1038/srep44331.

Fux, C. A., Shirtliff, M., Stoodley, P., & Costerton, J. W. (2005). Can laboratory reference strains mirror 'real-world' pathogenesis? *Trends in Microbiology*, *13*(2), 58–63. https://doi.org/10.1016/j.tim.2004.11.001.

Guérillot, R., Kostoulias, X., Donovan, L., Li, L., Carter, G. P., Hachani, A., Vandelannoote, K., Giulieri, S., Monk, I. R., Kunimoto, M., Starrs, L., Burgio, G., Seemann, T., Peleg, A. Y., Stinear, T. P., & Howden, B. P. (2019). Unstable chromosome rearrangements in *Staphylococcus aureus* cause phenotype switching associated with persistent infections. *Proceedings of the National Academy of Sciences*, *116*(40), 20135–20140. https://doi.org/10.1073/pnas.1904861116.

Guinane, C. M., Kent, R. M., Norberg, S., Hill, C., Fitzgerald, G. F., Stanton, C., & Ross, R. P. (2011). Host specific diversity in *Lactobacillus johnsonii* as evidenced by a major chromosomal inversion and phage resistance mechanisms. *PLoS ONE*, *6*(4), e18740. https://doi.org/10.1371/journal.pone.0018740.

Herzberg, M., Kaye, I. K., Peti, W., & Wood, T. K. (2006). YdgG (TqsA) Controls biofilm formation in *Escherichia coli* K-12 through autoinducer 2 transport. *Journal of Bacteriology*, *188*(2), 587–598. https://doi.org/10.1128/JB.188.2.587-598.2006.

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T. A., Munson, K. M., Kronenberg, Z. N., Vives, L., Peluso, P., Boitano, M., Chin, C.-S., Korlach, J., Wilson, R. K., & Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, *27*(5), 677–685. https://doi.org/10.1101/gr.214007.116.

Hung, L. W., Wang, I. X., Nikaido, K., Liu, P. Q., Ames, G. F., & Kim, S. H. (1998). Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature*, *396*(6712), 703–707. https://doi.org/10.1038/25393.

Iannelli, F., Giunti, L., & Pozzi, G. (1998). Direct sequencing of long polymerase chain reaction fragments. *Molecular Biotechnology*, *10*(2), 183–185. https://doi.org/10.1007/BF02760864.

Kempf, B., & Bremer, E. (1995). OpuA, an osmotically regulated binding protein-dependent transport system for the osmoprotectant glycine betaine in *Bacillus subtilis. The Journal of Biological Chemistry*, *270*(28), 16701–16713. https://doi.org/10.1074/jbc.270.28.16701

Koonin, E. V., & Tatusov, R. L. (1994). Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *Journal of Molecular Biology*, *244*(1), 125–132. https://doi.org/10.1006/jmbi.1994.1711.

Lai, P. J., Lim, C. T., Le, H. P., Katayama, T., Leach, D. R. F., Furukohri, A., & Maki, H. (2016). Long inverted repeat transiently stalls DNA replication by forming hairpin structures on both leading and lagging strands. *Genes to Cells*, *21*(2), 136–145. https://doi.org/10.1111/gtc.12326.

Lapierre, L., Mollet, B., & Germond, J.-E. (2002). Regulation and adaptive evolution of lactose operon expression in *Lactobacillus delbrueckii*. *Journal of Bacteriology*, *184*(4), 928–935. https://doi.org/10.1128/jb.184.4.928-935.2002.

Leach, D. R. F. (1994). Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays*, *16*(12), 893–900. https://doi.org/10.1002/bies.950161207.

Lederer, F. (1994). The cytochrome bs-fold: An adaptable module. *Biochimie*, 76(7):674-92. https://10.1016/0300-9084(94)90144-9.

Lee, H., Doak, T. G., Popodi, E., Foster, P. L., & Tang, H. (2016). Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Research*, gkw647. https://doi.org/10.1093/nar/gkw647.

Lépine, G., Lacroix, J. M., Walker, C. B., & Progulske-Fox, A. (1993). Sequencing of a *tet(Q)* gene isolated from *Bacteroides fragilis* 1126. *Antimicrobial Agents and Chemotherapy*, *37*(9), 2037–2041. https://doi.org/10.1128/AAC.37.9.2037.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Liu, Y., Bauer, S. C., & Imlay, J. A. (2011). The YaaA protein of the *Escherichia coli* OxyR regulon lessens hydrogen peroxide toxicity by diminishing the amount of intracellular unincorporated iron. *Journal of Bacteriology*, *193*(9), 2186–2196. https://doi.org/10.1128/JB.00001-11

Lu, Y.-J., Zhang, Y.-M., Grimes, K. D., Qi, J., Lee, R. E., & Rock, C. O. (2006). Acyl-phosphates initiate membrane phospholipid synthesis in Gram-positive pathogens. *Molecular Cell*, *23*(5), 765–772. https://doi.org/10.1016/j.molcel.2006.06.030.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944.

Meinnel, T., Blanquet, S., & Dardel, F. (1996). A new subclass of the zinc metalloproteases superfamily revealed by the solution structure of peptide deformylase. *Journal of Molecular Biology*, *262*(3), 375–386. https://doi.org/10.1006/jmbi.1996.0521.

Mera, P. E., St Maurice, M., Rayment, I., & Escalante-Semerena, J. C. (2009). Residue Phe112 of the human-type corrinoid adenosyltransferase (PduO) enzyme of *Lactobacillus reuteri* Is critical to the formation of the four-coordinate Co(II) corrinoid substrate and to the activity of the enzyme,. *Biochemistry*, *48*(14), 3138–3145. https://doi.org/10.1021/bi9000134.

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., Shaw, P. D., & Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, *14*(2), 193–202. https://doi.org/10.1093/bib/bbs012.

Mol, C. D., Kuo, C. F., Thayer, M. M., Cunningham, R. P., & Tainer, J. A. (1995). Structure and function of the multifunctional DNA-repair enzyme exonuclease III. *Nature*, *374*(6520), 381–386. https://doi.org/10.1038/374381a0.

Neu, H. C. (1969). Effect of 3-lactamase location in *Escherichia coli* on penicillin synergy. *Applied microbiology*, 17(6), 783–786. https://doi.org/10.1128/am.17.6.783-786.1969.

Nikolskaya, A. N. (2002). A novel type of conserved DNA-binding domain in the transcriptional regulators of the AlgR/AgrA/LytR family. *Nucleic Acids Research*, *30*(11), 2453–2459. https://doi.org/10.1093/nar/30.11.2453.

Pao, S. S., Paulsen, I. T., & Saier, M. H. (1998). Major facilitator superfamily. *Microbiology and Molecular Biology Reviews*, *62*(1), 1–34. https://doi.org/10.1128/MMBR.62.1.1-34.1998.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114.

Pei, J., Mitchell, D. A., Dixon, J. E., & Grishin, N. V. (2011). Expansion of type II CAAX proteases reveals evolutionary origin of γ-secretase subunit APH-1. *Journal of Molecular Biology*, *410*(1), 18–26. https://doi.org/10.1016/j.jmb.2011.04.066.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., … Bashir, A. (2015).

Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*(8), 780–786. https://doi.org/10.1038/nmeth.3454.

Periwal, V., & Scaria, V. (2015). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, *31*(1), 1–9. https://doi.org/10.1093/bioinformatics/btu600.

Raeside, C., Gaffé, J., Deatherage, D. E., Tenaillon, O., Briska, A. M., Ptashkin, R. N., Cruveiller, S., Médigue, C., Lenski, R. E., Barrick, J. E., & Schneider, D. (2014). Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *MBio*, *5*(5). https://doi.org/10.1128/mBio.01377-14.

Rafii, F., & Park, M. (2008). Detection and characterization of an ABC transporter in *Clostridium hathewayi*. *Archives of Microbiology*, *190*(4), 417–426. https://doi.org/10.1007/s00203-008-0385-3.

Ravin, V., & Alatossava, T. (2003). Three new insertion sequence elements IS*Ldl2*, IS*Ldl3*, and IS*Ldl4* in *Lactobacillus delbrueckii:* Isolation, molecular characterization, and potential use for strain identification. *Plasmid*, *49*(3), 253–268. https://doi.org/10.1016/S0147-619X(03)00018-0.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754.

Rocco, J. M., & Churchward, G. (2006). The integrase of the conjugative transposon Tn*916* directs strand- and sequence-specific cleavage of the origin of conjugal transfer, *oriT* , by the endonuclease orf20. *Journal of Bacteriology*, *188*(6), 2207–2213. https://doi.org/10.1128/JB.188.6.2207-2213.2006.

Romero, D., & Palacios, R. (1997). Gene amplification and genomic plasticity in prokaryotes. *Annual Review of Genetics*, *31*(1), 91–111. https://doi.org/10.1146/annurev.genet.31.1.91.

Roth, J. R., Benson, N., Galitski, T., Haack, K., Lawrence, J. G., & Miesel, L. (1996). Rearrangements of the Bacterial Chromosome: Formation and Applications. 37.

Rudy, C. K., Scott, J. R., & Churchward, G. (1997). DNA binding by the Xis protein of the conjugative transposon Tn*916*. *Journal of Bacteriology*, *179*(8), 2567–2572. https://doi.org/10.1128/jb.179.8.2567-2572.1997.

Santoro, F., Oggioni, M. R., Pozzi, G., & Iannelli, F. (2010). Nucleotide sequence and functional analysis of the tet (M)-carrying conjugative transposon Tn*5251* of *Streptococcus pneumoniae*: Tn*5251* of *Streptococcus pneumoniae*. *FEMS Microbiology Letters*, no-no. https://doi.org/10.1111/j.1574-6968.2010.02002.

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J. E., Remus-Emsermann, M. N. P., & Ahrens, C. H. (2018). Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Research*, *46*(17), 8953–8965. https://doi.org/10.1093/nar/gky726.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7.

Shao, H., Ganesamoorthy, D., Duarte, T., Cao, M. D., Hoggart, C. J., & Coin, L. J. M. (2018). npInv: Accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics*, *19*(1), 261. https://doi.org/10.1186/s12859-018-2252-9.

Smith, G. R. (1988). Homologous recombination in prokaryotes. *Microbiol Rev.,* 52(1):1-28. https://doi: 10.1128/mr.52.1.1-28.1988.

Sousa, C., de Lorenzo, V., & Cebolla, A. (1997). Modulation of gene expression through chromosomal positioning in *Escherichia coli*. *Microbiology*, *143*(6), 2071–2078. https://doi.org/10.1099/00221287-143-6-2071.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, *48*(4), 582–592. https://doi.org/10.1073/pnas.48.4.582.

Sun, S., Ke, R., Hughes, D., Nilsson, M., & Andersson, D. I. (2012). Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS ONE*, *7*(8), e42639. https://doi.org/10.1371/journal.pone.0042639.

Suyama, M., & Bork, P. (2001). Evolution of prokaryotic gene order: Genome rearrangements in closely related species. *Trends in Genetics*, *17*(1), 10–13. https://doi.org/10.1016/S0168-9525(00)02159-4.

Tailliez, P., Ehrlich, S. D., & Chopin, M. C. (1994). Characterization of IS*1201*, an insertion sequence isolated from *Lactobacillus helveticus. Gene*, *145*(1), 75–79. https://doi.org/10.1016/0378-1119(94)90325-5.

Tanner, N. K., & Linder, P. (2001). DExD/H box RNA helicases: From generic motors to specific dissociation functions. *Molecular Cell*, *8*(2), 251–262. https://doi.org/10.1016/s1097-2765(01)00329.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569.

Tillier, E. R. M., & Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nature Genetics*, *26*(2), 195–197. https://doi.org/10.1038/79918.

Tkaczuk, K. L., Dunin-Horkawicz, S., Purta, E., & Bujnicki, J. M. (2007). Structural and evolutionary bioinformatics of the SPOUT superfamily of methyltransferases. *BMC Bioinformatics*, *8*(1), 73. https://doi.org/10.1186/1471-2105-8-73.

Türck, M., & Bierbaum, G. (2012). Purification and activity testing of the full-length YycFGHI proteins of *Staphylococcus aureus*. *PLoS ONE*, *7*(1), e30403. https://doi.org/10.1371/journal.pone.0030403.

Upton, C., & Buckley, J. T. (1995). A new family of lipolytic enzymes? *Trends in Biochemical Sciences*, *20*(5), 178–179. https://doi.org/10.1016/s0968-0004(00)89002-7.

Vesanto, E., Peltoniemi, K., Purtsi, T., Steele, J. L., & Palva, A. (1996). Molecular characterization, over-expression and purification of a novel dipeptidase from *Lactobacillus helveticus*. *Applied Microbiology and Biotechnology*, *45*(5), 638–645. https://doi.org/10.1007/s002530050741.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963.

Weber, E., Chevallier, M.-R., & Jund, R. (1988). Evolutionary relationship and secondary structure predictions in four transport proteins of *Saccharomyces cerevisiae*. *Journal of Molecular Evolution*, *27*(4), 341–350. https://doi.org/10.1007/BF02101197.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of *de novo* genome assemblies: Fig. 1. *Bioinformatics*, *31*(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Woodson, K., & Devine, K. M. (1994). Analysis of a ribose transport operon from *Bacillus subtilis*. *Microbiology*, *140*(8), 1829–1838. https://doi.org/10.1099/13500872-140-8-1829

Yan, Y., Li, Y., Munshi, S., Sardana, V., Cole, J. L., Sardana, M., Kuo, L. C., Chen, Z., Steinkuehler, C., Tomei, L., & Francesco, R. D. (1998). Complex of NS3 protease and

NS4A peptide of BK strain hepatitis C virus: A 2.2 Å resolution structure in a hexagonal crystal form: 2.2 A crystal structure of the NS3 protease of hepatitis C virus. *Protein Science*, *7*(4), 837–847. https://doi.org/10.1002/pro.5560070402.

FIGURE LEGEND

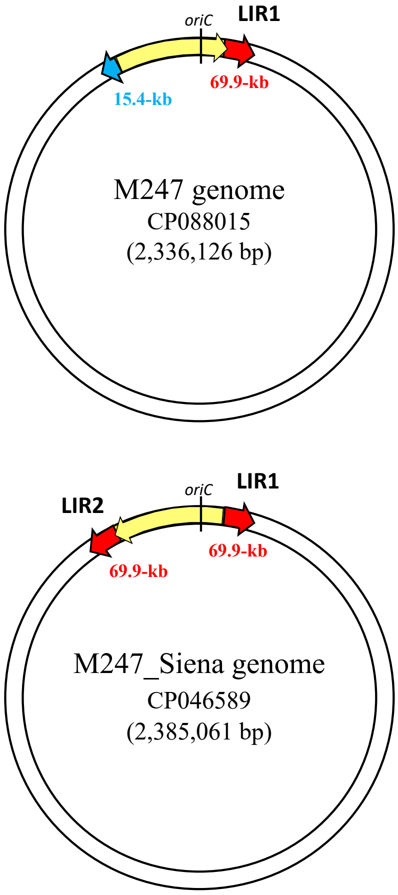


***Figure 1.* Schematic representation of the *L. crispatus* laboratory strain M247_Siena genome, in comparison with the original *L. crispatus* M247 genome.** The *L. crispatus* M247_Siena genome, 2,385,061 bp in length, contains a duplication of a 69.9-kb long chromosomal segment that produces two long inverted repeats (LIRs, red arrows) 224.4-kb apart. LIR1 and LIR2 are located 38,619 nucleotides downstream and 185,826 nucleotides upstream of the chromosomal origin of replication (*oriC*), respectively. Comparison with the *L. crispatus* original M247 genome indicated that i) the LIR2 was replacing a 15.4-kb DNA sequence (blue arrow) which is deleted in the M247_Siena genome and that ii) the LIR2 is associated to a 224.4-kb inversion of the genomic region flanked by LIR1 and LIR2, containing the *oriC* (yellow arrow).

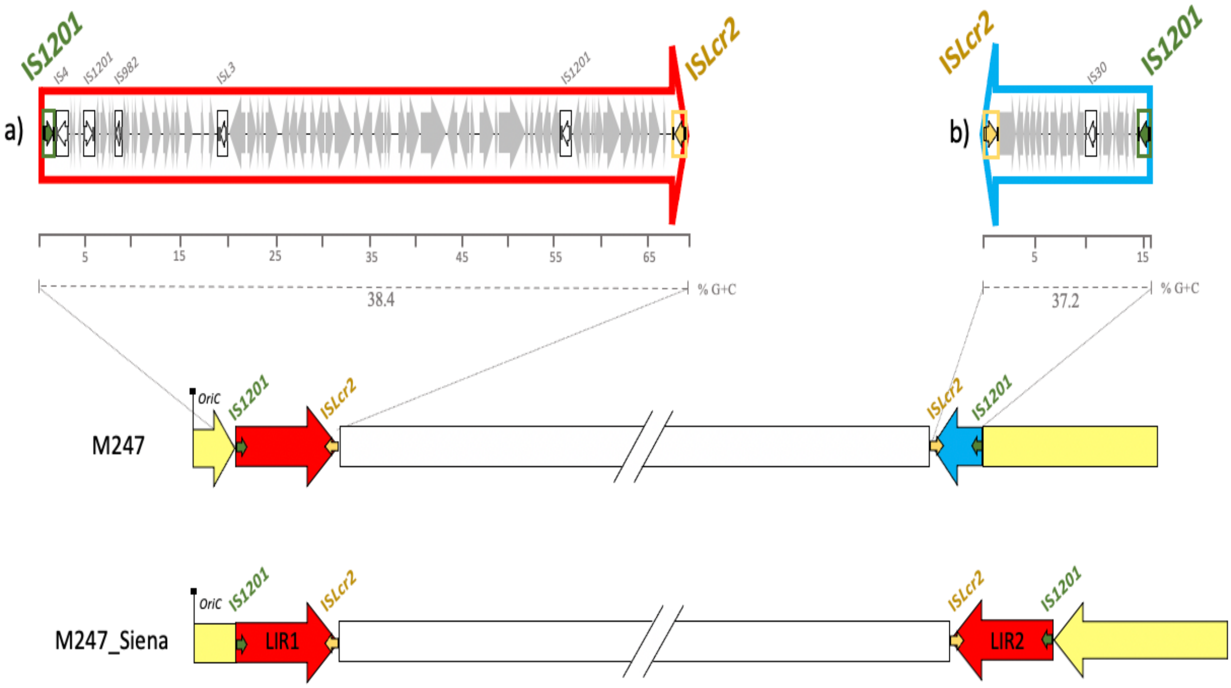


Figure 2. **(a) Structure of the M247_Siena 69.9-kb repeat and (b) of the 15.4-kb DNA sequence of M247 which is deleted and replaced by LIR2 in M247_Siena.** (a) Sequence analysis of the M247_Siena LIR1 and LIR2 showed that each copy of the repeat contains 72 ORFs, of which 46 have the same direction of transcription. Each repeat contains at the 5’ end a copy of IS*1201* and at the 3’ end a copy of IS*Lcr2*. Sequence comparison showed that the 2 LIRs contains 4 nucleotides changes and a 6-bp insertion all located in the 5’ end copy of IS*1201* (not shown). LIRs have an overall GC content of 38.4%. ORFs and their direction of transcription are represented by grey arrows, whereas boxed arrows are ISs. Scale, kilobases. (b) The same ISs elements flanking the LIR, namely IS*1201* and IS*Lcr2* were found at 5’ end and 3’ end, respectively, of the deleted 15.4-kb DNA sequence. The 15.4-kb region contains 17 ORFs and has a GC content of 37.2%. The linear representation of M247 and M247_Siena genomes highlights that in both genomes IS*1201* and IS*Lcr2* (green and orange small arrows, respectively) occur in the form of inverted repeats; the figure is not scaled.

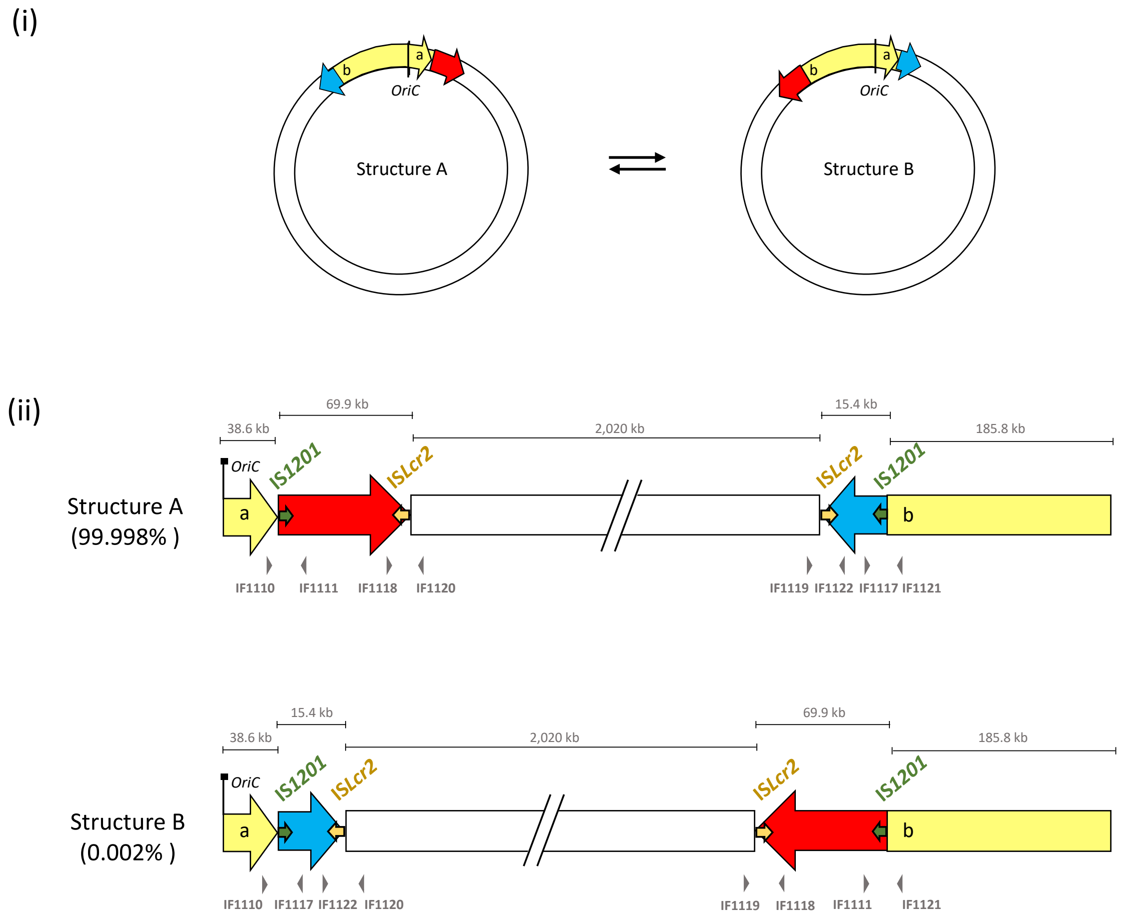


Figure 3. **Detection of chromosomal rearrangements within the original M247 population by PCR analysis.** (i) Circular and (ii) linear representations of the two chromosomal structures detected within the M247 population (structure A and structure B). PCR analysis indicated the presence of M247 chromosomes in which the 69.9-kb and the 15.4-kb DNA sequences were exchanged (structure B) in comparison to the assembled chromosomal structure (structure A). PCR mapping also suggested the IS elements namely IS*1201* and IS*Lcr2* (green and orange small arrows, respectively) flanking the exchanged DNA sequences as breakpoints of rearrangements. The PCR real-time quantification of the two M247 chromosomal structures estimated that chromosomal rearrangements involving the 69.9-kb and the 15.4-kb DNA sequences occurred in approximately 2 out of 100,000 M247 chromosomes (0.0002%). Red and blue arrow blocks represent the 69.9-kb region (duplicated in M247_Siena) and the 15.4-kb DNA segment (deleted in M247_Siena), respectively. Grey arrowheads indicate primers used for chromosomal structure analysis. The figure is not scaled.

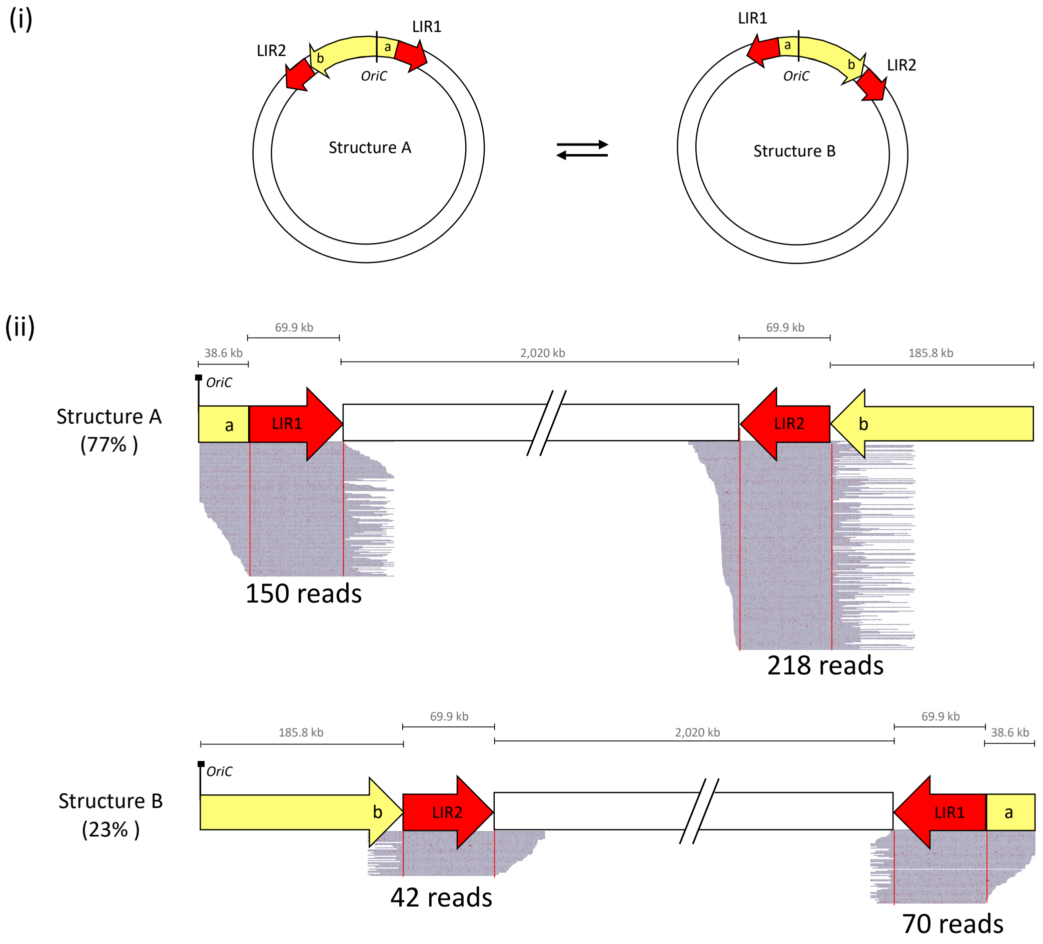


***Figure 4.* Detection of chromosomal rearrangements in the M247_Siena genome by genome mapping with Nanopore reads.** Analysis of Nanopore reads longer than 80-kb using npInv and Sniffles tools indicated the presence of reads supporting two chromosomal structures in M247_Siena (structure A and structure B). (i) Homologous recombination between the 2 long inverted repeats LIR1 and LIR2 produces the inversion of the 224,443-bp DNA region (yellow arrow), flanked by the LIRs (red arrows), which characterizes the two chromosomal structures A and B. The figure is not scaled. (ii) Linear representation of the 2 genomic structures aligned to the Nanopore reads (grey) longer than 80-kb spanning LIR1 and LIR2, which are useful in structural identification. The distribution images of the reads were copied by Tablet v1.17.08.17 interface panel and have been slightly modified to improve readability. A semi-quantitative estimation, obtained by reads counting, revealed that 112 out 480 useful reads (23.3%) contain the inverted structure (structure B).

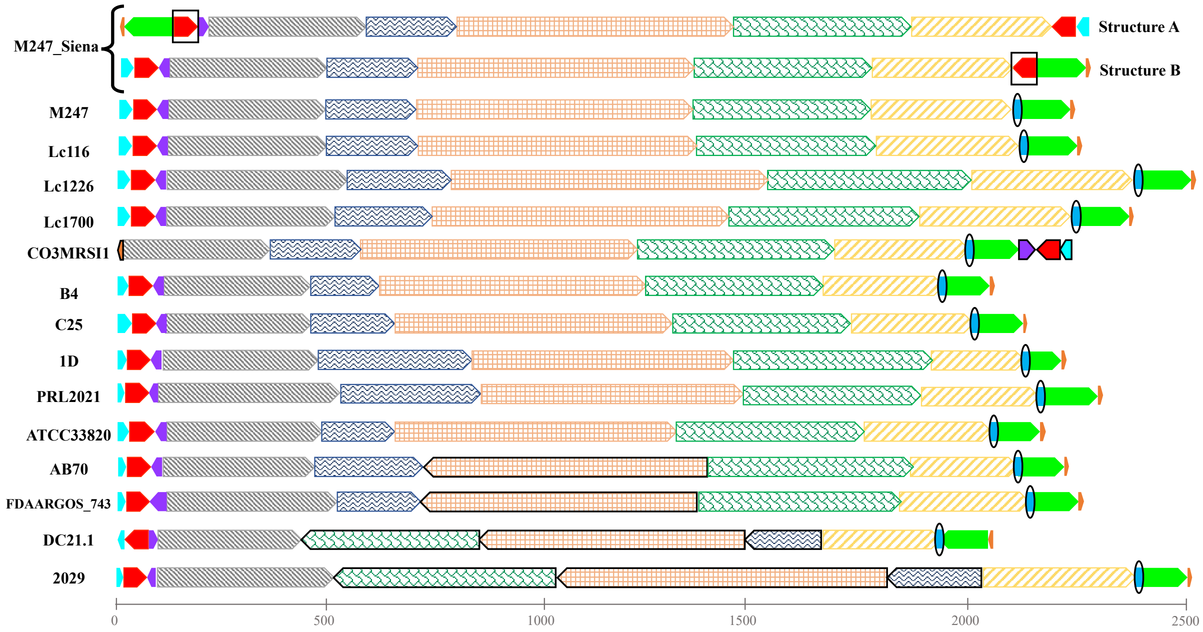
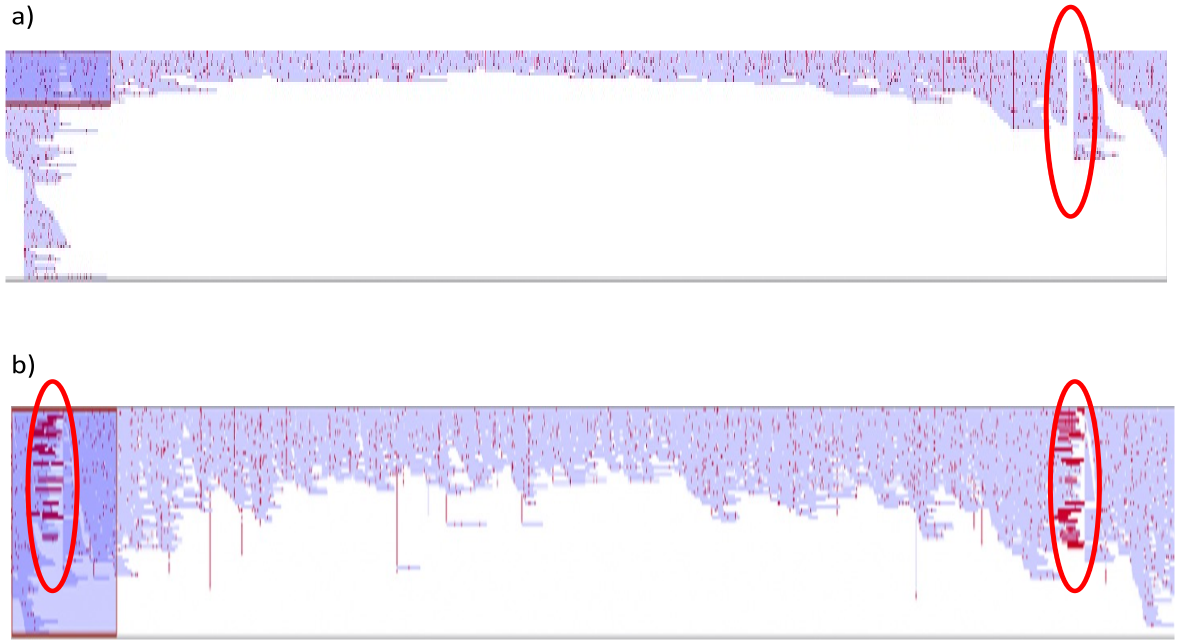


Figure 5. **Multiple alignment of the 14 *L. crispatus* complete genomes using M247_Siena as reference.** Each chromosome has been laid out horizontally: identically colored blocks indicate similar nucleotide sequences, left oriented blocks indicate reverse complement orientation relative to the reference chromosome (M247_Siena). Names of strains carrying the sequences are reported on the left of the figure. Each *L. crispatus* genome is covered by eleven blocks of homology, of which light blue and orange containing the replication origin and the termination sites, respectively; red indicates the 69.9-kb region duplicated in M247_Siena, light blue indicates the 15.4-kb region, whereas remaining blocks (light green, green, yellow, violet, grey, dark blue and brown) are conserved syntenic sequences. Other *L. crispatus* strains display a chromosomal structure resembling the M247 genome and therefore, characterized by the presence of the 15.4-kb region (circled light blue block) and the absence of LIR2 (boxed red block). In all strains the 224.4-kbp DNA segment (light blue, light green and orange blocks) is arranged as in the structure B of M247_Siena. Some strains show further inversion (black outlined blocks): CO3MRSI1 displays an inversion of 191-kb around the origin of replication, whereas AB70, FDAARGOS, DC21.1 and 2029 have an additional DNA inversion of 610-kb, 619-kb, 1,3 Mb and 1,7 Mb in length respectively occurring in proximity of the putative termination sites. Scale, kilobases.



Supplementary Figure S1. **M247_Siena and M247 genome mapping with Nanopore reads.** (a) Alignment of the M247_Siena Nanopore reads (2,151x coverage) to the M247 genome highlights the 15.4-kb deletion (red circle), whereas (b) the alignment of the M247 Nanopore reads (883x coverage) to the laboratory strain M247_Siena genome highlights the absence of the 69.9-kb repeats (red circles). Figure was generated using Tablet v1.17.08.17.

CHAPTER 4

Sequence typing and antimicrobial susceptibility testing of infertility-associated *Enterococcus faecalis* reveals clonality of aminoglycoside resistant strains

Stefano De Giorgi[a,#], Susanna Ricci[a,*], Lorenzo Colombini[a], David Pinzauti[a], Francesco Santoro[a,b], Francesco Iannelli[a], Stefania Cresti[b], Paola Piomboni[c,d], Vincenzo De Leo[c,d] and Gianni Pozzi[a]

[a]*Laboratory of Molecular Microbiology and Biotechnology (LA.M.M.B.), Department of Medical Biotechnologies, University of Siena, Italy*

[b]*Microbiology and Virology Unit, Siena University Hospital, Siena, Italy*

[c]*Department of Molecular and Developmental Medicine, University of Siena, Siena, Italy*

[d]*UOSA Medically Assisted Reproduction, Siena University Hospital, Siena, Italy*

[#] Present address: Department of Molecular Medicine, University of Padova, Italy.

[*] Corresponding author. e-mail: susanna.ricci@unisi.it (S. Ricci)

Short title: Infertility-associated *Enterococcus faecalis*

Submitted to Journal of Global Antimicrobial Resistance

1. ABSTRACT

Objectives: Infertility affects 9-12% of reproductive-aged couples. Both symptomatic and asymptomatic genital infections can contribute to infertility. We have recently shown that asymptomatic genital infection of infertile couples by *Enterococcus faecalis*, *Mycoplasma hominis* and *Ureaplasma urealyticum* was predictive of *in vitro* fertilization failure. This study aims at characterizing antibiotic susceptibility and population structure of a collection of infertility-associated *E. faecalis* strains.

Methods: Antibiotic susceptibility testing included VITEK-2, MIC and disk-diffusion assays. Oxford Nanopore and Illumina sequences were employed for hybrid genome assemblies. Genomes were used for multilocus sequence typing and identification of antimicrobial resistance genes.

Results: All 41 strains were susceptible to β-lactams, glycopeptides, tigecyclin, linezolid and nitrofurantoin, whereas 8/41 isolates were resistant to at least one antimicrobial. All the 8 strains showed resistance to high-level aminoglycosides, of which 7 were resistant to gentamicin. Only one strain was resistant to gentamycin, streptomycin, ciprofloxacin and levofloxacin. Simpson's diversity index indicated genotypic diversity of the infertility-associated *E. faecalis* population. Seventeen sequence types (STs) were identified and assigned to 3 clonal complexes (CCs) and 14 singletons. CC40 was the most predominant, followed by ST81 and CC16. Interestingly, 6/7 gentamicin resistant isolates clustered in CC16/ST480. The *aac(6')-aph(2")* and *ant(6)* genes encoding aminoglycoside modifying enzymes were identified in the gentamicin and streptomycin resistant strains, respectively, whereas quinolone resistance was mediated by point-mutations in the *gyrA* and *parC* genes.

Conclusions: All isolates were susceptibile to most clinically relevant antimicrobials. Twenty percent of the strains showed resistance to high-level aminoglycosides and 75% of those clustered in CC16/ST480.

Keywords: Infertility, asymptomatic genital infection, infertility-associated *Enterococcus faecalis*, antibiotic susceptibility, sequence type, clonal complex.

2. INTRODUCTION

According to the International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO), infertility is a disease of the reproductive system which generates disability as an impairment of function [1]. Infections of the genital tract are accounted amongst the factors contributing to infertility [2–4]. We have recently reported that asymptomatic infections of the genital tract had a negative impact on couple fertility, and the presence of *Enterococcus faecalis*, *Mycoplasma hominis* and *Ureaplasma urealyticum* in genital samples was predictive of an adverse outcome of *in vitro* fertilization (IVF) [5]. In particular, *E. faecalis* was significantly associated to reduced motility and abnormal morphology of spermatozoa in semen specimens and lower levels of lactobacilli in vaginal swabs [5]. Association of *E. faecalis* to altered semen parameters in infertile males with no symptoms of genital tract infections has also been described by other authors [6–8]. Couples seeking medical help due to infertility seldom present with genital infections with overt symptoms, such as pain, discomfort or discharge. Nonethless, numerous studies have associated asymptomatic or poorly symptomatic genital infections to impaired fertility [3,7,9–19]. Given that asymptomatic genital tract infections can threaten fertility and antimicrobial therapy may improve reproduction efficiency, the present study was conducted to characterize the *E. faecalis* clinical isolates previously associated to couple infertility [5] with the prospect of a potential antibiotic treatment of genital tract infections caused by *E. faecalis* in infertile couples. *E. faecalis* is the species responsible for the majority of enterococcal infections in humans, including urinary tract infections (UTIs), sepsis, endocarditis, peritonitis, abdominal/pelvic and soft tissue infections [20]. The most frequent clinical manifestation is UTI, of which *E. faecalis* is the second most common agent worldwide after *Escherichia coli* [21]. *E. faecalis* is also the leading pathogen among Gram-positive bacteria of catheter-associated UTIs (CAUTIs) in healthcare settings [22]. Ascending UTIs and intra-abdominal infections can lead to bacteremia and endocarditis. Both *Enterococcus faecium* and *E. faecalis* have a remarkable tropism for the endocardium and/or the heart valves, but *E. faecalis* alone accounts for about 90% of enterococcal endocarditis cases, especially in risk groups [23]. Treatment of asymptomatic enterococcal bacteriuria is not recommended by recent guidelines [24]. Uncomplicated enterococcal UTIs are generally managed with drugs in monotherapy (*i.e.*, ampicillin, fosfomycin, nitrofurantoin or fluoroquinolones). Antimicrobial agents for complicated UTIs and pyelonephritis by *E. faecalis* include penicillin/ampicillin,vancomycin, linezolid and daptomycin [25]. Treatment of enterococcal endocarditis, blood and deep-tissue infections requires the synergic combination of β-lactams or glicopeptides together with aminoglycosides, typically gentamicin [23,25]. Management of endocarditis and sepsis caused by aminoglycoside resistant enterococci necessitates alternative combination antimicrobial therapies [26]. Despite *E. faecalis* is one of the most commonly isolated uropathogen from semen samples of infertile men [5,8,27,28], therapeutic management of *E. faecalis* bacteriospermia is still controversial among reproductive medicine specialists. Antimicrobial therapy of enterococcal infections is complicated by their intrinsic resistance to several antibiotic classes, including cephalosporins, sulphonamides and low concentrations of aminoglycosides [29]. In addition, acquired antibiotic resistance may limit the number of therapeutic options especially for severe infections. Of special concern is the acquisition by horizontal transfer of genes coding for aminoglycoside modifying enzymes (AMEs) and conferring resistance to high concentrations of aminoglycosides [29,30]. As a result, the synergistic bactericidal effect between β-lactams and aminoglycosides is eliminated prompting the need for different combination therapies. According to the latest european surveillance report, the population-weighted mean percentage of high-level aminoglycoside resistance in *E. faecalis* in 2019 was 26.6 [31]. In the perspective of antimicrobial treatment of asymptomatic genital tract infections, it is important to investigate the epidemiology of antibiotic resistance of the bacterial pathogens isolated from genital samples of infertile couples. In the present study, antimicrobial susceptibility testing, whole-genome sequencing and multilocus sequence typing (MLST) were jointly used to characterize population structure and antibiotic resistance of a collection of *E. faecalis* clinical strains isolated from asymptomatic couples with infertility.

3. MATERIALS AND METHODS

3.1. Clinical isolates

A total of 41 clinical isolates of *E. faecalis* from 285 infertile couples attending the Centre for Diagnosis and Treatment of Couple Sterility at Siena University Hospital were analyzed. All couples were asymptomatic for genital tract infections. *E. faecalis* strains included 28 isolates from semen samples and 13 from vaginal swabs as described [5].

3.2. Bacterial growth conditions

Each *E. faecalis* isolate was grown on solid Brain Heart Infusion (BHI; Oxoid, Milan, Italy) enriched with 5% defibrinated horse blood (Liofilchem, Teramo, Italy) at 37°C overnight (o.n.). Four to six isolated bacterial colonies were suspended in 10 ml of liquid BHI medium (Oxoid) and incubated at 37°C until they reached the optical density at 590 nm (OD_{590}) of 0.5. Cultures were aliquoted, added with 10% glycerol (Baker, Bridgend, England) and stored at -80°C until use.

3.3. Antimicrobial susceptibility testing

For each isolate, antimicrobial susceptibility testing was initially performed using VITEK® 2 (Biomerieux Italia S.p.A., Florence, Italy) with the AST-P658 card (Biomerieux) covering the antibiotics recommended for enterococci by EUCAST (The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters, version 11.0, 2021). To confirm and implement the VITEK data, both MIC and disk diffusion (Kirby-Bauer and E-test) assays were carried out. *E. faecalis* OG1RF [32] was used as reference strain. Each isolate was cultured o.n. on blood-agar BHI plates, and then few isolated colonies were suspended in dH_2O to reach the turbidity of 0.5 McFarland. For MIC testing of gentamicin (GEN) and streptomycin (STR), bacterial suspensions were diluted (1:700) in liquid Mueller Hinton medium (Biomérieux) and distributed into the wells (50 µl/well) of a custom microtiter plate for gram-positive bacteria (Sensititre GPN3F; Thermo Fisher Scientific, Milano, Italy). The plate was incubated at 37°C o.n., and MICs were determined by manual reading. Kirby-

Bauer assays were used to test susceptibility to all antibiotics (Oxoid), while E-test was employed for linezolid. Results were assessed based on EUCAST breakpoints.

3.4. Genomic DNA preparation

Frozen stocks of the 41 enterococcal isolates were diluted (1:100) in 40 ml of BHI and grown until an OD_{590} of 2.0. Genomic DNA extraction was carried out as described [33]. DNA samples were resuspended in 0.9% NaCl and quantified with the Qubit 2.0 Fluorometer (Invitrogen, Life Technologies, Carlsbad, CA, USA) using the Qubit dsDNA BR assay kit (Thermo Fisher Scientific).

3.5. Oxford Nanopore Sequencing

Sequencing reactions were carried out in 1.5 ml LoBind tubes (Sarstedt, Nümbrecht, Germany) using wide bore (Ø 1.2 mm) tips to reduce DNA shearing. DNA size selection of genomic DNA was obtained with 0.5 vol of AMPure XP beads (Beckman Coulter S.r.l., Milano, Italy) according to the manufacturer's instructions. Approximately 2 µg of size-selected DNA was employed for library construction by using the Nanopore sequencing kit SQK-LSK 108 (Oxford Nanopore Technologies, Oxford, United Kingdom). Multiple samples were pooled using the Nanopore 'Native Barcoding Expansion 1-12 kit' (Oxford Nanopore Technologies). Library preparation was performed following the manufacturer's protocol with the following modifications: (i) each incubation step with the XP beads was done on a rotator mixer for 15 min; (ii) the 'Library Loading Beads' (LLB) were not employed. Finally, the pooled DNA library (at least 200 ng) was loaded onto a R9.4 MinION flow cell (Oxford Nanopore Technologies). Sequencing run was performed on the GridION X5 device (Oxford Nanopore Technologies) until a 100x genome coverage for each sample was reached (approximately 8-12 h). Real-time basecalling and analysis of basecalled reads was carried out as described [33]. Features of nanopore reads obtained from sequencing the 8 aminoglycoside resistant *E. faecalis* strains are reported in Table S2.

3.6. Illumina sequencing

Illumina sequencing was performed at Microbes NG (University of Birmingham, Birmingham, UK), using the Nextera library preparation kit (Illumina Inc., San Diego, USA) followed by HiSeq2500 sequencing (Illumina Inc.) (2x250 bp paired-end sequencing). Illumina reads were trimmed and analysed as reported [33]. Features of Illumina reads achieved from sequencing the 8 aminoglycoside resistant isolates are described in Table S3.

3.7. Genome assembly, annotation and analysis

Nanopore and Illumina reads of all 41 strains were assembled, polished and quality-assessed as described [33]. Genomes were automatically annotated using Prokka v1.14.5 ([34]; https://github.com/tseemann/prokka). Genome analysis was carried out using: (i) Artemis and Artemis Comparison Tool (ACT) v17.0.1 [35]; (ii) Blast (https://blast.ncbi.nlm.nih.gov/Blast.cgi); (iii) PlasmidFinder v2.0.1 [36]. *E. faecalis* OG1RF genome was obtained from the NCBI Microbial Genome Database (https://www.ncbi.nlm.nih.gov/genome/808?genome_assembly_id=168518, CP002621.1) and used as a reference strain.

3.8. Statistical analyses

Starting from the sequenced genomes, ST of each isolate was assigned based on the 7 housekeeping genes *gdh*, *gyd*, *pstS*, *gki*, *aroE*, *xpt* and *yqiL used for enterococcal typing* [37]. For each locus, a distinct allele number was assigned in accordance with the *E. faecalis* MLST database (https://pubmlst.org/organisms/enterococcus-faecalis/). Simpson's index of diversity (D) with 95% confidence interval was calculated [$0 \leq (1-D) \leq 1$, with values near zero corresponding to high diversity and values near one corresponding to more homogeneous populations]. The relatedness amongst different STs was investigated by the UPGMA agglomerative hierarchical clustering method using PHYLOViZ v2.0 [38]. The UPGMA method was used to construct a dendrogram from the matrix of pairwise allelic differences between the STs. The nearest two clusters were joined into a higher level cluster, and the distance (Hamming distance) between any two clusters is the mean distance between elements of each cluster. Clusters of related STs differing in ≤ 2 allelic loci and descending from a common ancestor were grouped into CCs by using goeBURST [38]. A singleton was defined as a ST unrelated to any other in the population at single-locus variant level [38]. Antimicrobial resistance genes were identified using ABRicate v1.0.1 (Seemann T, *Abricate*, https://github.com/tseemann/abricate) on the following databases ARG-ANNOT (Antibiotic Resistance Gene-ANNOTation) [39], CARD (Comprehensive Antibiotic Resistance Database) [40], MEGARes 2.00 [41] and ResFinder [42].

4. RESULTS

4.1. Antimicrobial susceptibility of infertility-associated *E. faecalis*

Susceptibility of all 41 *E. faecalis* isolates to 14 clinically relevant antimicrobial drugs was tested according to the EUCAST guidelines. Results obtained with the VITEK 2 automated system were confirmed by both disk diffusion and broth microdilution MIC methods. All the strains tested (41/41) were susceptible to β-lactams, glycopeptides, tigecycline, linezolid and nitrofurantoin, whereas 8/41 (19.5%) were resistant to at least one antimicrobial agent (Fig. 1). High-level aminoglycoside resistance was observed in all resistant isolates (8/8), whereas only one strain (1/8) was resistant to the fluoroquinolones ciprofloxacin (CIP) and levofloxacin (LVX). A total of 4 different phenotypic antimicrobial resistance patterns was defined (Table 1). Five isolates were resistant only to GEN, one only to STR, one to both GEN and STR, and eventually one to GEN, STR, CIP and LVX (Table 1).

4.2. Sequence types and identification of clonal complexes

Complete genomes of all isolates were obtained by using both short- and long-read sequencing techniques followed by hybrid assembly. Whole genome sequences were used to perform MLST on the 7 genes employed for *E. faecalis* typing. MLST allowed to assign 17 different sequence types (STs) (Fig. 2). All STs were present in the *E. faecalis* database. The most frequently found types were ST40 (11/41 isolates), ST81 (7/41) and ST179 (5/41) (Fig. 2 and Table S1). The other

14 STs were identified in ≤3 isolates (Fig. 2 and Table S1). Calculation of the Simpson's index of diversity (1-D, D = 0.889; 95% confidence interval = 0.83-0.95) showed a high level of diversity of the infertility-associated *E. faecalis* population. The goeBURST algorithm was then used to group the STs with allelic variants in one or two loci into clonal complexes (CCs). Analysis resolved 3 groups and 11 singletons with single- and double-locus variants. Blast with *E. faecalis* MLST database clustered the 17 STs into 14 distinct CCs, of which 3 (CC40, CC16 and CC21) comprised strains belonging to at least two different STs and 11 were singletons. The most prevalent cluster was CC40 (12/41 isolates), followed by ST81 (7/41) and CC16 (6/41) (Table S1). CC40 comprised ST40 and the single-locus variant ST268, CC16 covered ST16 and the single-locus variant ST179, while CC21 included ST21 and ST117 (Table S1).

4.3. Phylogenetic relatedness of high-level aminoglycoside resistant *E. faecalis* strains

Analysis of the distribution of aminoglycoside resistance among the STs showed that 6 out of the 8 resistant isolates were closely related (Fig. 2). In particular, 4 isolates (strains 5245, 2819, 4638 and 5034) belonged to ST179, while the other 2 were part of ST16 (strain 5410) and ST480 (strain 4774). The remaining 2 isolates belonged to the more distant ST211 (strain 4153) and ST40 (strain 4953) (Fig. 2). Construction of the minimum spanning tree containing the allelic variants in just 1 locus (n-1, n=7) indicated that ST179 and ST16 belonged to the same CC, of which ST16 is the group founder (CC16) (Fig. 3). Further inclusion of allelic variants in 4 gene loci (n-4) allowed to comprise also ST480 in the group (Fig. 3). Interestingly, 6 out the 7 isolates (85.7%) resistant to high-level GEN clustered in CC16/ST480, suggesting clonality of high-level aminoglycoside resistant *E. faecalis* isolates. In contrast, strains belonging to ST211 and ST40 presented allelic variants in 6 and 7 loci, respectively, indicating higher phylogenetic distance to the other isolates (Fig. 3).

4.4. Resistance to high-level aminoglycosides is mediated by aminoglycoside modifying enzymes

Complete genomes of the 8 high-level aminoglycoside resistant strains were searched for the presence of genes encoding AME and conferring high-level resistance to GEN and STR, using ABRicate. All the 7 GEN resistant strains were found to carry one copy of the *aac(6')-aph(2'')* gene coding for the bifunctional 6'-aminoglycoside acetyltransferase-2''-aminoglycoside phosphotransferase enzyme, whose presence in gram-positive bacteria is known to confer resistance to GEN and most other aminoglycosides, except for STR [30] (Fig. 4). A single copy of the *ant(6)* gene, conferring high-level STR resistance, was found in all the 3 STR resistant strains (Fig. 4). The ANT(6) enzyme is an aminoglycoside O-nucleotidyltransferases with streptomycin as a unique substrate [43]. Strains 5034 and 4774 harbored both *aac(6')-aph(2'')* and *ant(6)* genes (Fig. 4). Analysis of the genomic location of AME genes showed that they were placed in the *E. faecalis* chromosome in 6 out of 8 isolates, whereas in two cases (strains 4153 and 4953) they were carried by plasmids (Fig. 4).

4.5. Resistance to fluoroquinolones is due to mutations

Strain 4774 is the only isolate resistant to both CIP and LVX. The genome of 4774 was searched for the presence of both acquired resistance genes and point mutations in the chromosomal genes *gyrA* and *parC*. The strain was found to carry two point mutations in *gyrA* (Ser83Tyr) and *parC* (Ser80Ile) that conferred fluoroquinolone resistance, as also confirmed by MIC results (4 μg/ml for both CIP and LVX).

5. DISCUSSION

Infections of the genital tract can negatively impact on couple fertility [5]. Both subclinical and chronic infections are considered as a potential threaten to human fertility [3,44], since low-grade but persistent inflammation in the genital tract can affect reproductive efficiency. Specifically, inflammatory mediators such as cytokines, chemokines and reactive oxygen species can harm the functions of Sertoli cells resulting in reduced spermatogenesis and failed acrosome reaction [45,46]. In the female, dysbiosis of the vaginal microbiome [*i.e.,* bacterial vaginosis (BV) and aerobic vaginitis] is often accompanied by increased levels of proinflammatory cytokines and reactive oxygen species at the genital mucosa, which in turn can alter the estrous cycle, ovulation, oocyte and embryo quality [47–49]. Because these infections often remain asymptomatic, they can ascend along the reproductive tract and also be transmitted to the uninfected partner during natural intercourse or assisted reproductive technology (ART) procedures, thereby increasing the risk of couple infertility [2,4,49]. Yet, as genital infections are preventable and curable causes of infertility, efforts should be made to diagnose and treat potential infections caused by microbial pathogens associated to couple infertility. Interestingly, recent studies have shown that antibiotic treatment of asymptomatic genital infections caused by *Chlamydia trachomatis*, *U. urealyticum*, *M. hominis* and *Mycoplasma genitalium* improved semen parameters [12–15,17], possibly enhancing male fertility. Moreover, despite treatment of asymptomatic BV is not currently recommended [50], however, antimicrobial therapy may prevent sexual transmission of BV-associated pathogens and reduce BV complications, including tubal factor infertility and early spontaneous abortion in patients subjected to ART procedures [51–53]. In the present study, we have characterized a collection of *E. faecalis* clinical strains previously isolated from infertile couples [5]. Although *E. faecalis* did not cause a symptomatic infection in either partner, it negatively affected sperm parameters and levels of vaginal lactobacilli, likely contributing to the observed IVF failure [5]. Starting from those findings, we have investigated the antibiotic susceptibility and population structure of 41 infertility-associated *E. faecalis* clinical isolates for a prospective antimicrobial treatment of asymptomatic genital infections in infertile couples prior to being subjected to IVF cycles. The majority (33/41) of isolates were susceptible to clinically relevant antimicrobials, and almost all (40/41) were susceptible to antibiotics used to treat *E. faecalis* UTIs (Fig. 1). The 8 remaining strains were resistant to high-level aminoglycosides, and 1 of those (strain 4774) was also resistant to fluoroquinolones (Table 1). Several epidemiological studies reported the prevalence of high-level aminoglycoside resistant *E. faecalis* strains isolated from different geographic areas and human body sites [54–60], however, to our knowledge, no previous work has specifically described the antibiotic resistance profile of *E. faecalis* isolates associated to human infertility. Here, the prevalence of high-level GEN and STR resistance was 17.1% and 7.3%, respectively. The 2 strains displaying high-level resistance to both GEN and STR were vaginal isolates (5034 and 4774), in accordance with a report by Quinones *et al.* describing aminoglycoside resistance in community-acquired vaginal isolates over a 5 year-period in Cuba [55]. Compared to the most recent european surveillance report [31], high-level GEN resistance rate in infertility-associated *E. faecalis* strains was lower than both the european (26.6%) and italian (35.2%) mean rate percentages, which however, did not specifically refer to UTI or genital tract infections. MLST of enterococci has shown the emergence of specific CCs of antibiotic resistant isolates responsible for hospital and community outbreaks [56,60–63]. In contrast to *E. faecium*, *E. faecalis* has an overall non-clonal population structure due to high level of genetic recombination [60–62]. Also in this study, MLST analysis showed a highly diverse *E. faecalis* population associated to infertility. However, a homogeneous group was constituted by 5 out the 8 high-level aminoglycoside resistant strains that all belonged to CC16 (Figs 2 and 3), which is widely spread among both hospital and community isolates throughout Europe [56,62]. Moreover, the 4774 vaginal strain belonged to the closely related ST480, indicating that 75% of high-level aminoglycoside resistant *E. faecalis* isolates clustered in CC16 and ST480. Consistent with our data, a recent genomic analysis of *E. faecalis* strains collected from 16 middle-east and african countries showed that CC16 was the most predominant complex and almost all the ST16/ST480 strains originated from UTI and CAUTI and were resistant to high-level GEN [64]. Association of high-level GEN resistance with ST16 was also observed in european community-acquired isolates [56]. In contrast, mono-resistance to STR (strain 4953, CC40) was clonally distant from CC16/ST480. Of note, an *E. faecalis* epidemic clone circulating in Poland over a 10 year-period, was enriched with multi-drug resistant strains mostly belonging to ST40 (CC40) [54]. In our case, CC40 was the largest complex of the population gathering 29.2% of the isolates, which were instead susceptible (except for strain 4953) to the antimicrobials tested. Recently, a genomic analysis of a large collection of 2027 *E. faecalis* isolates from different sources showed that ST6, ST16 and ST40 were the largest STs, of which ST16 and ST40 contained strains of both human and animal origin, while ST6 was exclusively human-associated [65]. None of the 8 high-level aminoglycoside resistant strains belonged to the high-risk enterococcal complexes CC2 and CC9, which are well adapted to the hospital environment and capable of global dissemination [56,61–63]. High-level aminoglycoside resistance in enterococci is generally mediated by enzymatic drug modification by AMEs, including phosphotransferases (APH), acetyltransferases (AAC), and nucleotidiltransferases (ANT) [30,66]. The most common gene is *aac(6')-aph(2'')*, encoding the bifunctional enzyme AAC(6')-APH(2''), that confers resistance to virtually all clinically available aminoglycosides except for STR [67] and partially arbekacin [30]. Also in our case, 7 out of 8 isolates resistant to high-level GEN (MIC ≥1024 µg/ml) carried the *aac(6')-aph(2'')* gene (Fig 4). On the other hand, the *ant(6)* gene was identified in the 2 vaginal isolates (strains 4774 and 5034) and in strain 4953 which all showed resistance to high-level STR (MIC ≥2048 µg/ml) (Fig 4). As previously reported [68], the *ant(6)* gene was found to be part of the described gene cluster *ant(6)-sat4A-aph(3')*, which mediates resistance to STR, kanamycin and streptothricin. Indeed, the 3 high-level STR resistant isolates were also resistant to kanamycin and streptothricin (De Giorgi *et al.*, data not shown). In addition, strain 4774 was also resistant to CIP and LVX due to two previously described point-mutations in both *parC* and *gyrA* [69]. Whole genome sequencing analysis of the 8 resistant isolates showed that the AME genes were located either on the chromosome (6/8 strains) or on plasmids (2/8 strains). Interestingly, all the 6 strains with chromosomally-located AME genes clustered in CC16/ST480, revealing a clonal structure and suggesting the presence of a common genetic element mediating high-level aminoglycoside resistance in the infertility-associated *E. faecalis* collection (Colombini *et al.*, data not shown).

Asymptomatic infections of the genital tract in infertile couples often remain undetected and consequently untreated, possibly affecting natural fertility [3,5,44]. There is still a substantial lack of clinical data on the impact of antimicrobial therapy on natural conception and reproductive outcomes. However, as antibiotic therapy may improve semen quality and health of vaginal microbiota, species identification and drug susceptibility testing of microbial pathogens associated to human infertility should be pursued and provided to clinicians to treat genital infections. In this study, most infertility-associated *E. faecalis* isolates were susceptible to antibiotics commonly employed to manage *E. faecalis* UTI. Nevertheless, antibiotic resistance to high-level aminoglycosides was still observed in 20% of the population, which mostly showed clonality in CC16/ST480. High-level aminoglycoside resistance, a key marker of enterococcal antibiotic resistance worldwide, needs to be carefully monitored as it impedes the synergistic effect of aminoglycosides with cell-wall active agents. Therefore, understanding the molecular epidemiology and antimicrobial resistance of *E. faecalis* isolated from the urogenital tract is important not only to treat UTI and possible systemic infections, but also to provide a tailored antimicrobial treatment to infertile couples before they approach the cumbersome ART procedures.

TABLES

***Table 1.* Antibiotic resistance patterns in infertility-associated *E. faecalis* isolates**

Isolates (n)	**Resistance pattern[a]**			
5	Gm^R			
1	Gm^R	Sm^R	Cip^R	Lvx^R
1	Gm^R	Sm^R		
1		Sm^R		
8	**Total**			

[a] Gm^R, high-level gentamicin resistance; Sm^R, high-level streptomycin resistance; Cip^R, ciprofloxacin resistance; Lvx^R, levofloxacin resistance.

***Table S1.* Sequence type and clonal complex distribution in infertility-associated *E. faecalis* isolates.**

ST	**Frequency (%)**	**CC[a]**
40	11 (26.8)	40
81	7 (17.1)	81*
179	5 (12.2)	16
44	3 (7.3)	44*
34	2 (4.9)	34*
191	2 (4.9)	191*
16	1 (2.4)	16
19	1 (2.4)	19*
21	1 (2.4)	21
25	1 (2.4)	25*
53	1 (2.4)	53*
55	1 (2.4)	55*
72	1 (2.4)	72*
117	1 (2.4)	21
211	1 (2.4)	211*
268	1 (2.4)	40
480	1 (2.4)	480*

[a] CCs were obtained by blasting the isolates with the *E. faecalis* MLST database (https://pubmlst.org/organisms/enterococcus-faecalis) using goeBURST. Different STs were grouped into the same CC when they differed in 1 or 2 out of the 7 gene loci used for typing *E. faecalis* strains.

*, Singletons. Singletons are defined as STs unrelated to any other in the population at single-locus variant level [38].

Table S2. **Statistics of the nanopore reads obtained from sequencing the 8 high-level aminoglycoside resistant *E. faecalis* strains associated to infertility.**

Strains	Reads (n)	Lenght (no of bases)			Quality (Q)[b]		Sequencing output (bp)
		Mean	Median	N50[a]	Mean	Median	
2819	99,212	21,887.3	15,893.0	37,366	9.6	9.7	2,171,480,426
4153	80,292	12,374.3	6,296.5	26,882	12.3	12.6	993,556,414
4638	1,859	17,411.7	11,306.0	31,361	9.3	9.4	32,368,292
4774	49,490	18,367.5	13,517.0	29,310	9.7	9.7	909,007,230
4953	32,764	20,964.0	16,261.5	32,716	10.4	10.4	686,865,305
5034	28,647	20,878.0	14,031.0	35,873	9.6	9.6	598,093,079
5245	40,517	17,466.4	13,689.0	25,715	10.4	10.4	707,686,757
5410	163,553	6,193.7	3,319.0	11,921	12.3	12.6	1,012,992,166

[a] N50 is the length of a sequence in a set for which all sequences of that length or greater sum to 50% of the set's total size.

[b] Phred quality score Q expresses the confidence in a particular base-call and is logarithmically related to the base-calling error probability P ($Q = -10 \log_{10} P$).

Table S3. **Statistics of the trimmed Illumina reads obtained from sequencing the 8 high-level aminoglycoside resistant *E. faecalis* strains associated to infertility.**

Strain	**Reads (n)**[a]	**Quality**[b]		**Sequencing output (bp)**
		Mean	**Median**	
2819	R1 (2,115,916)	33.5	34.7	434,674,786
	R2 (2,115,916)	31.1	31.3	411,540,257
4153	R1 (1,199,574)	31.5	31.6	260,978,599
	R2 (1,199,574)	28.4	27.4	227,373,392
4638	R1 (252,341)	30.3	30.1	59,754,920
	R2 (252,341)	26.5	25.1	52,419,311
4774	R1 (239,572)	29.6	29.2	58,100,485
	R2 (239,572)	24.6	23.9	49,177,827
4953	R1 (57,071)	31.6	31.8	12,503,658
	R2 (57,071)	28.4	27.4	10,882,249
5034	R1 (668,018)	23.0	22.8	155,126,281
	R2 (668,018)	25.8	25.7	159,583,694
5245	R1 (491,176)	32.3	32.8	102,697,037
	R2 (491,176)	29.6	28.9	91,011,815
5410	R1 (428,302)	30.7	30.7	102,947,546
	R2 (428,302)	26.0	24.9	89,371,420

[a] For each strain, the Illumina paired-end run generates two read files R1 (forward) and R2 (reverse).

[b] Phred quality score Q expresses the confidence in a particular base-call and is logarithmically related to the base-calling error probability P ($Q = -10 \log_{10} P$).

Acknowledgements

We would like to acknowledge Davide Amodeo for help with the antimicrobial susceptibility testing. Funding: the study was supported by the Italian Ministry of Education, University and Research, grant ‘Progetti di Rilevante Interesse Nazionale’ (PRIN) 2017 no. 20177J5Y3P_003 “Next-generation antibacterials: new targets for old drugs and new drugs for old targets” (to GP). Competing interests: none declared. Ethical approval: not required

Contributor roles

SDG, investigation, data curation, formal analysis, writing (original draft); SR, conceptualization, data curation, supervision, writing (original draft, review & editing); LC, investigation, data curation, formal analysis; DP, investigation; FS, data curation, writing (original draft); FI, data curation, writing (original draft); SC, investigation, data curation; PP, data curation, writing (original draft); VDL, conceptualization; GP, conceptualization, supervision, funding acquisition, and writing (original draft, review & editing). All authors read and approved the manuscript.

6. REFERENCES

[1] Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouzon J, Sokol R, et al. The international glossary on infertility and fertility care, 2017. Fertil Steril 2017;108:393–406. https://doi.org/10.1016/j.fertnstert.2017.06.005.

[2] Pellati D, Mylonakis I, Bertoloni G, Fiore C, Andrisani A, Ambrosini G, *et al*. Genital tract infections and infertility. Eur J Obstet Gynecol Reprod Biol 2008;140:3–11. https://doi.org/10.1016/j.ejogrb.2008.03.009.

[3] Gimenes F, Souza RP, Bento JC, Teixeira JJV, Maria-Engler SS, Bonini MG, *et al*. Male infertility: a public health issue caused by sexually transmitted pathogens. Nat Rev Urol 2014;11:672–87. https://doi.org/10.1038/nrurol.2014.285.

[4] Tsevat DG, Wiesenfeld HC, Parks C, Peipert JF. Sexually transmitted diseases and infertility. Am J Obstet Gynecol 2017;216:1–9. https://doi.org/10.1016/j.ajog.2016.08.008.

[5] Ricci S, De Giorgi S, Lazzeri E, Luddi A, Rossi S, Piomboni P, *et al*. Impact of asymptomatic genital tract infections on *in vitro* Fertilization (IVF) outcome. PLoS One 2018;13:e0207684. https://doi.org/10.1371/journal.pone.0207684.

[6] Mehta RH, Sridhar H, Vijay Kumar BR, Anand Kumar TC. High incidence of oligozoospermia and teratozoospermia in human semen infected with the aerobic bacterium *Streptococcus faecalis*. Reprod Biomed Online 2002;5:17–21. https://doi.org/10.1016/s1472-6483(10)61591-x.

[7] Rodin DM, Larone D, Goldstein M. Relationship between semen cultures, leukospermia, and semen analysis in men undergoing fertility evaluation. Fertil Steril 2003;79:1555–8. https://doi.org/10.1016/S0015-0282(03)00340-6.

[8] Moretti E, Capitani S, Figura N, Pammolli A, Federico MG, Giannerini V, *et al*. The presence of bacteria species in semen and sperm quality. J Assist Reprod Genet 2009;26:47–56. https://doi.org/10.1007/s10815-008-9283-5.

[9] Hamdad-Daoudi F, Petit J, Eb F. Assessment of *Chlamydia trachomatis* infection in asymptomatic male partners of infertile couples. J Med Microbiol 2004;53:985–90. https://doi.org/10.1099/jmm.0.45641-0.

[10] Sanocka-Maciejewska D, Ciupińska M, Kurpisz M. Bacterial infection and semen quality. J Reprod Immunol 2005;67:51–6. https://doi.org/10.1016/j.jri.2005.06.003.

[11] Bezold G, Politch JA, Kiviat NB, Kuypers JM, Wolff H, Anderson DJ. Prevalence of sexually transmissible pathogens in semen from asymptomatic male infertility patients with and without leukocytospermia. Fertil Steril 2007;87:1087–97. https://doi.org/10.1016/j.fertnstert.2006.08.109.

[12] Ahmadi MH, Mirsalehian A, Sadighi Gilani MA, Bahador A, Afraz K. Association of asymptomatic *Chlamydia trachomatis* infection with male infertility and the effect of

antibiotic therapy in improvement of semen quality in infected infertile men. Andrologia 2018; 50:e12944. https://doi.org/10.1111/and.12944.

[13] Ahmadi MH, Mirsalehian A, Sadighi Gilani MA, Bahador A, Talebi M. Asymptomatic infection with *Mycoplasma hominis* negatively affects semen parameters and leads to male infertility as confirmed by improved semen parameters after antibiotic treatment. Urology 2017;100:97–102. https://doi.org/10.1016/j.urology.2016.11.018.

[14] Ahmadi MH, Mirsalehian A, Gilani MAS, Bahador A, Talebi M. Improvement of semen parameters after antibiotic therapy in asymptomatic infertile men infected with *Mycoplasma genitalium*. Infection 2018;46:31–8. https://doi.org/10.1007/s15010-017-1075-3.

[15] Zhang Q-F, Zhang Y-J, Wang S, Wei Y, Li F, Feng K-J. The effect of screening and treatment of *Ureaplasma urealyticum* infection on semen parameters in asymptomatic leukocytospermia: a case-control study. BMC Urol 2020;20:165. https://doi.org/10.1186/s12894-020-00742-y.

[16] Veiga E, Treviño M, Romay AB, Navarro D, Trastoy R, Macía M. Colonisation of the male reproductive tract in asymptomatic infertile men: effects on semen quality. Andrologia 2020;52:e13637. https://doi.org/10.1111/and.13637.

[17] Pajovic B, Radojevic N, Vukovic M, Stjepcevic A. Semen analysis before and after antibiotic treatment of asymptomatic *Chlamydia*- and *Ureaplasma*-related pyospermia. Andrologia 2013;45:266–71. https://doi.org/10.1111/and.12004.

[18] Babu G, Singaravelu BG, Srikumar R, Reddy SV, Kokan A. Comparative study on the vaginal flora and incidence of asymptomatic vaginosis among healthy women and in women with infertility problems of reproductive age. J Clin Diagn Res 2017;11:DC18–22. https://doi.org/10.7860/JCDR/2017/28296.10417.

[19] Piscopo RC, Guimarães RV, Ueno J, Ikeda F, Bella ZIJ-D, Girão MJ, *et al.* Increased prevalence of endocervical *Mycoplasma* and *Ureaplasma* colonization in infertile women

with tubal factor. JBRA Assist Reprod 2020;24:152–7. https://doi.org/10.5935/1518-0557.20190078.

[20] Agudelo Higuita NI, Huycke MM. Enterococcal disease, epidemiology, and implications for treatment. In: Gilmore MS, Clewell DB, Ike Y, Shankar N, eds. Enterococci: from commensals to leading causes of drug resistant infection. Boston: Massachusetts Eye and Ear Infirmary, 2014.

[21] Flores-Mireles AL, Walker JN, Caparon M, Hultgren SJ. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. Nat Rev Microbiol 2015;13:269–84. https://doi.org/10.1038/nrmicro3432.

[22] Peng D, Li X, Liu P, Luo M, Chen S, Su K, et al. Epidemiology of pathogens and antimicrobial resistance of catheter-associated urinary tract infections in intensive care units: a systematic review and meta-analysis. Am J Infect Control 2018;46:e81–90. https://doi.org/10.1016/j.ajic.2018.07.012.

[23] Fernández-Hidalgo N, Escolà-Vergé L, Pericàs JM. *Enterococcus faecalis* endocarditis: what's next? Future Microbiol 2020;15:349–64. https://doi.org/10.2217/fmb-2019-0247.

[24] Nicolle LE, Gupta K, Bradley SF, Colgan R, DeMuri GP, Drekonja D, *et al*. Clinical practice guideline for the management of asymptomatic bacteriuria: 2019 update by the Infectious Diseases Society of America. Clin Infect Dis 2019; 68:e83-e110. https://doi.org/10.1093/cid/ciy1121.

[25] Mercuro NJ, Davis SL, Zervos MJ, Herc ES. Combatting resistant enterococcal infections: a pharmacotherapy review. Expert Opin Pharmacother 2018;19:979–92. https://doi.org/10.1080/14656566.2018.1479397.

[26] Beganovic M, Luther MK, Rice LB, Arias CA, Rybak MJ, LaPlante KL. A review of combination antimicrobial therapy for *Enterococcus faecalis* bloodstream infections and infective endocarditis. Clin Infect Dis 2018;67:303–9. https://doi.org/10.1093/cid/ciy064.

[27] Monteiro C, Marques PI, Cavadas B, Damião I, Almeida V, Barros N, et al. Characterization of microbiota in male infertility cases uncovers differences in seminal hyperviscosity and oligoasthenoteratozoospermia possibly correlated with increased prevalence of infectious bacteria. Am J Reprod Immunol 2018;79:e12838. https://doi.org/10.1111/aji.12838.

[28] Farahani L, Tharakan T, Yap T, Ramsay JW, Jayasena CN, Minhas S. The semen microbiome and its impact on sperm function and male fertility: a systematic review and meta-analysis. Andrology 2021;9:115–44. https://doi.org/10.1111/andr.12886.

[29] Kristich CJ, Rice LB, Arias CA. Enterococcal infection - Treatment and antibiotic resistance. In: Gilmore MS, Clewell DB, Ike Y, Shankar N, eds. Enterococci: from commensals to leading causes of drug resistant infection. Boston: Massachusetts Eye and Ear Infirmary, 2014.

[30] Chow JW. Aminoglycoside resistance in enterococci. Clin Infect Dis 2000;31:586–9. https://doi.org/10.1086/313949.

[31] Antimicrobial resistance in the EU/EEA (EARS-Net) - Annual Epidemiological Report for 2019. European Centre for Disease Prevention and Control 2020. https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019.

[32] Bourgogne A, Garsin DA, Qin X, Singh KV, Sillanpaa J, Yerrapragada S, *et al*. Large scale variation in *Enterococcus faecalis* illustrated by the genome analysis of strain OG1RF. Genome Biol 2008;9:R110. https://doi.org/10.1186/gb-2008-9-7-r110.

[33] Teodori L, Colombini L, Cuppone AM, Lazzeri E, Pinzauti D, Santoro F, *et al*. Complete genome sequence of *Lactobacillus crispatus* type strain ATCC 33820. Microbiol Resour Announc 2021;10:e0063421. https://doi.org/10.1128/MRA.00634-21.

[34] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153.

[35] Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, *et al*. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 2008;24:2672–6. https://doi.org/10.1093/bioinformatics/btn529.

[36] Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, *et al*. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 2014;58:3895–903. https://doi.org/10.1128/AAC.02412-14.

[37] Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 2018;3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.

[38] Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 2012;13:87. https://doi.org/10.1186/1471-2105-13-87.

[39] Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, *et al*. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014;58:212–20. https://doi.org/10.1128/AAC.01310-13.

[40] Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, *et al*. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017;45:D566–73. https://doi.org/10.1093/nar/gkw1004.

[41] Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res 2020;48:D561–9. https://doi.org/10.1093/nar/gkz1010.

[42] Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, *et al.* Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012;67:2640–4. https://doi.org/10.1093/jac/dks261.

[43] Hormeño L, Ugarte-Ruiz M, Palomo G, Borge C, Florez-Cuadrado D, Vadillo S, *et al. ant(6)-I* genes encoding aminoglycoside O-nucleotidyltransferases are widely spread among streptomycin resistant strains of *Campylobacter jejuni* and *Campylobacter coli*. Front Microbiol 2018;9:2515. https://doi.org/10.3389/fmicb.2018.02515.

[44] Ochsendorf FR. Sexually transmitted infections: impact on male fertility. Andrologia 2008;40:72–5. https://doi.org/10.1111/j.1439-0272.2007.00825.x.

[45] Agarwal A, Rana M, Qiu E, AlBunni H, Bui AD, Henkel R. Role of oxidative stress, infection and inflammation in male infertility. Andrologia 2018;50:e13126. https://doi.org/10.1111/and.13126.

[46] Altmäe S, Franasiak JM, Mändar R. The seminal microbiome in health and disease. Nat Rev Urol 2019;16:703–21. https://doi.org/10.1038/s41585-019-0250-y.

[47] Weiss G, Goldsmith LT, Taylor RN, Bellet D, Taylor HS. Inflammation in reproductive disorders. Reprod Sci 2009;16:216–29. https://doi.org/10.1177/1933719108330087.

[48] Agarwal A, Aponte-Mellado A, Premkumar BJ, Shaman A, Gupta S. The effects of oxidative stress on female reproduction: a review. Reprod Biol Endocrinol 2012;10:49. https://doi.org/10.1186/1477-7827-10-49.

[49] Ravel J, Moreno I, Simón C. Bacterial vaginosis and its association with infertility, endometritis, and pelvic inflammatory disease. Am J Obstet Gynecol 2021;224:251–7. https://doi.org/10.1016/j.ajog.2020.10.019.

[50] Workowski KA, Bolan GA, Centers for Disease Control and Prevention. Sexually transmitted diseases treatment guidelines, 2015. MMWR Recomm Rep 2015;64:1–137.

[51] Muzny CA, Schwebke JR. Asymptomatic bacterial vaginosis: to treat or not to treat? Curr Infect Dis Rep 2020;22:32. https://doi.org/10.1007/s11908-020-00740-z.

[52] van Oostrum N, De Sutter P, Meys J, Verstraelen H. Risks associated with bacterial vaginosis in infertility patients: a systematic review and meta-analysis. Hum Reprod 2013;28:1809–15. https://doi.org/10.1093/humrep/det096.

[53] Haahr T, Jensen JS, Thomsen L, Duus L, Rygaard K, Humaidan P. Abnormal vaginal microbiota may be associated with poor reproductive outcomes: a prospective study in IVF patients. Hum Reprod 2016;31:795–803. https://doi.org/10.1093/humrep/dew026.

[54] Kawalec M, Pietras Z, Daniłowicz E, Jakubczak A, Gniadkowski M, Hryniewicz W, *et al*. Clonal structure of *Enterococcus faecalis* isolated from Polish hospitals: characterization of epidemic clones. J Clin Microbiol 2007;45:147–53. https://doi.org/10.1128/JCM.01704-06.

[55] Quiñones D, Kobayashi N, Nagashima S. Molecular epidemiologic analysis of *Enterococcus faecalis* isolates in Cuba by multilocus sequence typing. Microb Drug Resist 2009;15:287–93. https://doi.org/10.1089/mdr.2009.0028.

[56] Kuch A, Willems RJL, Werner G, Coque TM, Hammerum AM, Sundsfjord A, *et al*. Insight into antimicrobial susceptibility and population structure of contemporary human *Enterococcus faecalis* isolates from Europe. J Antimicrob Chemother 2012;67:551–8. https://doi.org/10.1093/jac/dkr544.

[57] Jabbari Shiadeh SM, Pormohammad A, Hashemi A, Lak P. Global prevalence of antibiotic resistance in blood-isolated *Enterococcus faecalis* and *Enterococcus faecium*: a systematic review and meta-analysis. Infect Drug Resist 2019;12:2713–25. https://doi.org/10.2147/IDR.S206084.

[58] Zalipour M, Esfahani BN, Havaei SA. Phenotypic and genotypic characterization of glycopeptide, aminoglycoside and macrolide resistance among clinical isolates of *Enterococcus faecalis*: a multicenter based study. BMC Res Notes 2019;12:292. https://doi.org/10.1186/s13104-019-4339-4.

[59] Haghi F, Lohrasbi V, Zeighami H. High incidence of virulence determinants, aminoglycoside and vancomycin resistance in enterococci isolated from hospitalized patients in Northwest Iran. BMC Infect Dis 2019;19:744. https://doi.org/10.1186/s12879-019-4395-3.

[60] Sparo M, Delpech G, García Allende N. Impact on public health of the spread of high-level resistance to gentamicin and vancomycin in enterococci. Front Microbiol 2018;9:3073. https://doi.org/10.3389/fmicb.2018.03073.

[61] Leavis HL, Bonten MJM, Willems RJL. Identification of high-risk enterococcal clonal complexes: global dispersion and antibiotic resistance. Curr Opin Microbiol 2006;9:454–60. https://doi.org/10.1016/j.mib.2006.07.001.

[62] Ruiz-Garbajosa P, Bonten MJM, Robinson DA, Top J, Nallapareddy SR, Torres C, *et al*. Multilocus sequence typing scheme for *Enterococcus faecalis* reveals hospital-adapted genetic complexes in a background of high rates of recombination. J Clin Microbiol 2006;44:2220–8. https://doi.org/10.1128/JCM.02596-05.

[63] Solheim M, Brekke MC, Snipen LG, Willems RJL, Nes IF, Brede DA. Comparative genomic analysis reveals significant enrichment of mobile genetic elements and genes encoding surface structure-proteins in hospital-associated clonal complex 2 *Enterococcus faecalis*. BMC Microbiol 2011;11:3. https://doi.org/10.1186/1471-2180-11-3.

[64] Farman M, Yasir M, Al-Hindi RR, Farraj SA, Jiman-Fatani AA, Alawi M, *et al*. Genomic analysis of multidrug-resistant clinical *Enterococcus faecalis* isolates for antimicrobial resistance genes and virulence factors from the western region of Saudi Arabia. Antimicrob Resist Infect Control 2019;8:55. https://doi.org/10.1186/s13756-019-0508-4.

[65] Pöntinen AK, Top J, Arredondo-Alonso S, Tonkin-Hill G, Freitas AR, Novais C, *et al*. Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. Nat Commun 2021;12:1523. https://doi.org/10.1038/s41467-021-21749-5.

[66] Ramirez MS, Tolmasky ME. Aminoglycoside modifying enzymes. Drug Resist Updat 2010;13:151–71. https://doi.org/10.1016/j.drup.2010.08.003.

[67] Leclercq R, Dutka-Malen S, Brisson-Noël A, Molinas C, Derlot E, Arthur M, et al. Resistance of enterococci to aminoglycosides and glycopeptides. Clin Infect Dis 1992;15:495–501. https://doi.org/10.1093/clind/15.3.495.

[68] Werner G, Hildebrandt B, Witte W. Aminoglycoside-streptothricin resistance gene cluster *aadE–sat4–aphA-3* disseminated among multiresistant isolates of *Enterococcus faecium*. Antimicrob Agents Chemother 2001;45:3267–9. https://doi.org/10.1128/AAC.45.11.3267-3269.2001.

[69] Kanematsu E, Deguchi T, Yasuda M, Kawamura T, Nishino Y, Kawada Y. Alterations in the GyrA subunit of DNA gyrase and the ParC subunit of DNA topoisomerase IV associated with quinolone resistance in *Enterococcus faecalis*. Antimicrob Agents Chemother 1998;42:433–5. https://doi.org/10.1128/AAC.42.2.433.

FIGURE LEGEND

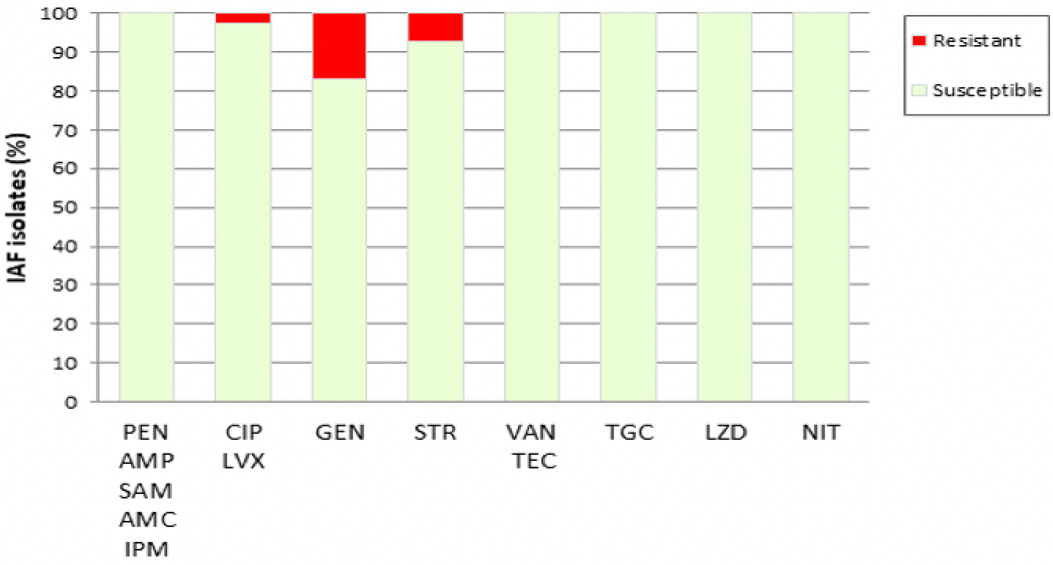


***Figure 1.* Antibiotic susceptibility profile of infertility-associated *E. faecalis*.** Forty-one *E. faecalis* clinical isolates were tested for their susceptibility to β-lactams (PEN, AMP, SAM, AMC, IPM), quinolones (CIP, LVX), high-level aminoglycosides (GEN, STR), glycopeptides (VAN, TEC), tetracyclines (TGC), oxazolidones (LZD) and nitrofurans (NIT). Results were obtained by VITEK-2 and confirmed by both MIC (Sensititre GPN3F plate) and diffusion-disk methods. Antibiotic susceptibility testing was assessed according to EUCAST guidelines. Resistant (red), Susceptible (light green).

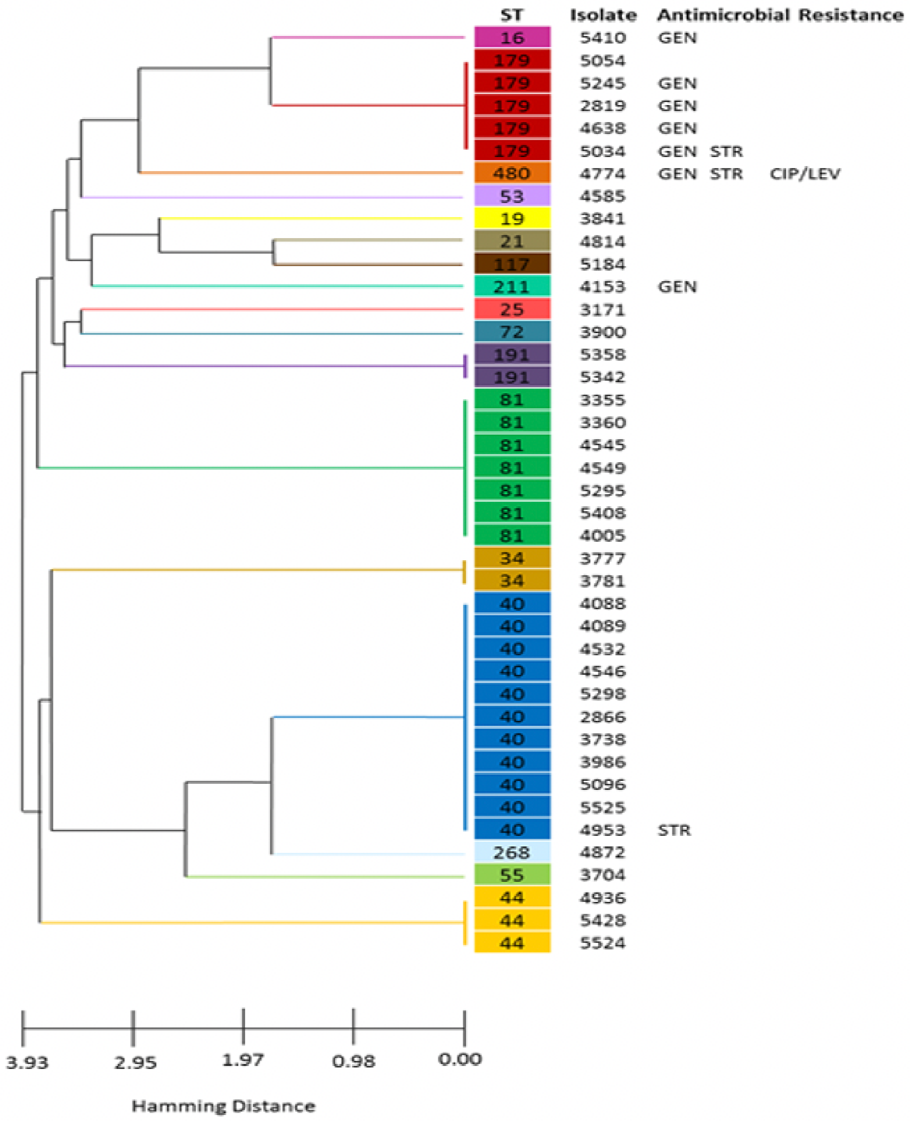


***Figure 2.* Unweighted pair group method with arithmetic mean (UPGMA) dendrogram of infertility-associated *E. faecalis* isolates based on ST.** The UPGMA dendogram was constructed starting from the matrix of pairwise allelic differences of the 7 loci defining *E. faecalis* ST by using PHYLOViZ v2.0 [38]. The two groups with the lowest number of allelic differences were combined into a higher level cluster, and the process was reiterated until the most distant groups were linked. The mean distance between any two clusters was measured by the Hamming distance, defined as the number of positions at which two aligned sequences differ. For each ST, isolate number and antimicrobial resistance are shown. GEN, gentamicin; STR, streptomycin; CIP, ciprofloxacin; LVX, levofloxacin.

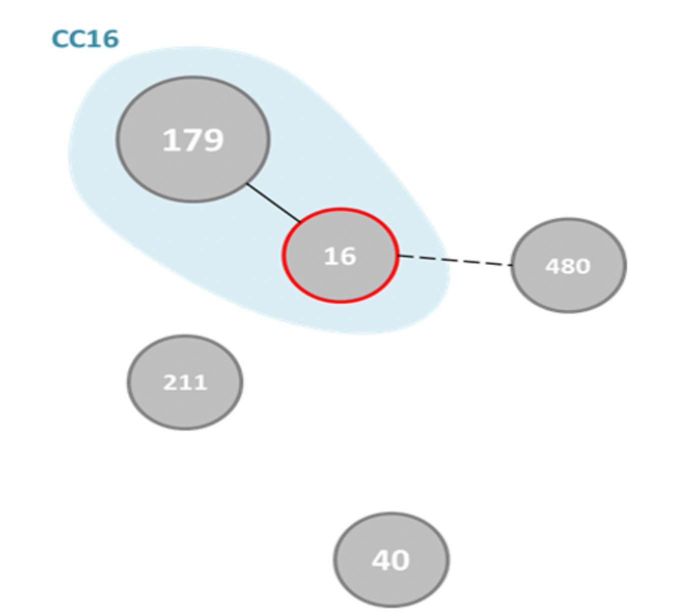
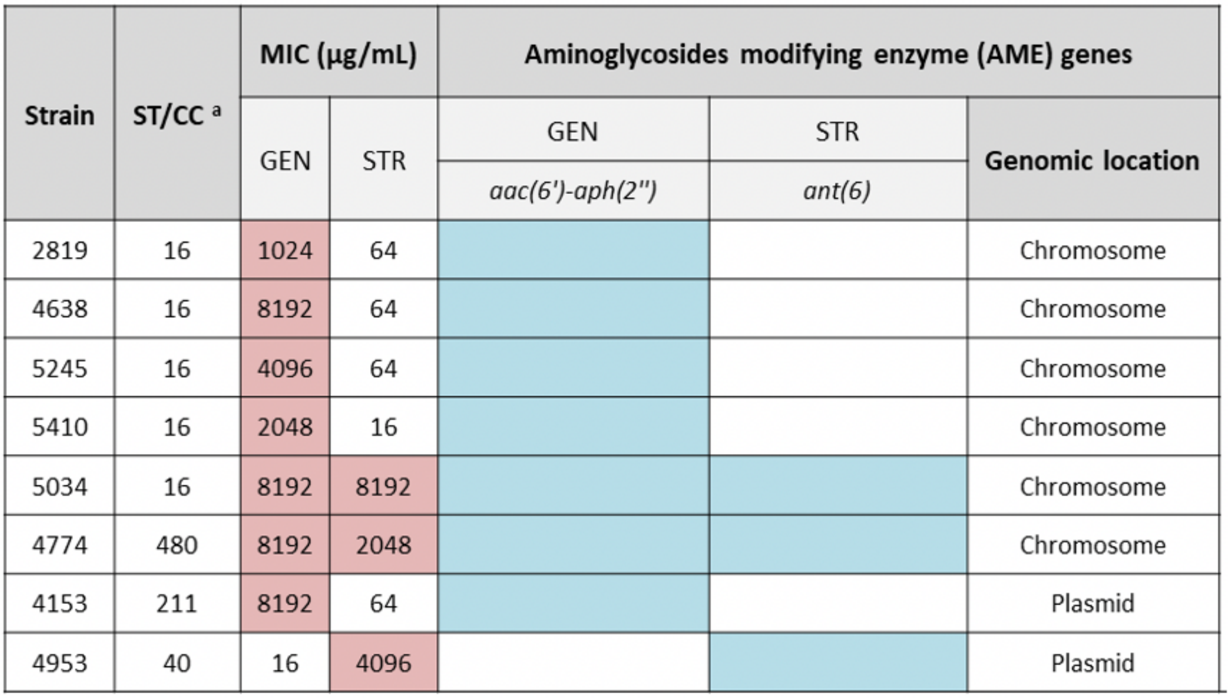


Figure 3. **Minimum spanning tree of infertility-associated *E. faecalis* strains resistant to high-level aminoglycosides.** The tree was constructed using PHYLOViZ v2.0 based on the goeBURST algorithm [38]. goeBURST divides large MLST data into nonoverlapping groups of related STs or CCs and then distinguishes the most parsimonious groups of descendants within each CC from the predicted founder. In *E. faecalis*, as the number of loci (n) defining ST is 7, the tree structure can be drawn at 7 levels of relatedness. The present diagram includes the 8 high-level aminoglycoside resistant isolates which are grouped into 5 different STs. Each ST is represented as a node whose size varies based on the number of isolates. ST179 contains 4 isolates (all Gm^R, except for one strain which is Gm^R Sm^R), while ST16 (Gm^R), ST480 (Gm^R Sm^R), ST211 (Gm^R) and ST40 (Sm^R) comprise 1 isolate each. ST16 and ST179 belong to the same CC with ST16 being the CC16 founder (red circle). The level of relatedness between ST16 and ST179 (n-1, solid line) and between ST16 and ST480 (n-4, dotted line) are also shown. ST211 and ST40 are more distantly related (>n-4, no connecting lines).



Strain	ST/CC [a]	MIC (µg/mL)		Aminoglycosides modifying enzyme (AME) genes		
		GEN	STR	GEN	STR	Genomic location
				aac(6')-aph(2")	*ant(6)*	
2819	16	1024	64	🟦		Chromosome
4638	16	8192	64	🟦		Chromosome
5245	16	4096	64	🟦		Chromosome
5410	16	2048	16	🟦		Chromosome
5034	16	8192	8192	🟦	🟦	Chromosome
4774	480	8192	2048	🟦	🟦	Chromosome
4153	211	8192	64	🟦		Plasmid
4953	40	16	4096		🟦	Plasmid

***Figure 4.* Genetic bases of resistance to aminoglycosides in infertility-associated *E. faecalis*.** For each strain, isolate number, CC/ST, MIC values of GEN and STR, and related AME genes are shown. GEN and STR were tested as recommended by EUCAST. MIC values >128 µg/ml (for GEN) and >512 µg/ml (for STR) were regarded as high-level aminoglycoside resistance (red boxes). Resistance to GEN conferred by *aac(6')-aph(2")* covers resistance to virtually all aminoglycosides, including tobramycin, amikacin, kanamycin, netilmicin and dibekacin. Search of AME genes conferring resistance to aminoglycosides was performed on the genomes of the 8 resistant strains by using the ABRicate tool on ARG-ANNOT, CARD, MEGARes and ResFinder databases. All the AME genes with a coverage and identity ≥99% are shown (light blue boxes). MIC values of GEN and STR of *E. faecalis* reference strain OG1RF were 1 µg/ml and 256 µg/ml, respectively.

[a] CC, clonal complex. ST, sequence type. Strains 4774 (ST480) and 4153 (ST211) are singletons.

CHAPTER 5

Nucleotide sequence analysis of the novel composite transposon Tn*7086* carrying aminoglycoside resistant genes in infertility-associated *Enterococcus faecalis* belonging to the clonal complex/sequence type CC16/ST480

Lorenzo Colombini[a], Stefano De Giorgi[b], Susanna Ricci[a], Francesco Santoro[a], Francesco Iannelli[a] and Gianni Pozzi[a]

[a]*Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy*

[b]*Department of Molecular Medicine, University of Padova, Italy.*

Manuscript in preparation

1. ABSTRACT

Background: A recent study has reported that aminoglycoside-resistant *Enterococcus faecalis* strains, isolated from infertile couples, showed clonality in the clonal complex (CC)/sequence type (ST) CC16/ST480. In this study, a novel mobile genetic element carrying aminoglycosides resistances genes in 6 infertility-associated and clonally-related *E. faecalis* clinical isolates was identified and characterized.

Methods: The complete genomes of the *E. faecalis* strains 2819, 4638, 4774, 5034, 5245, 5410, were obtained and compared to the genome of the reference strain OG1RF. The DNA sequence of a novel composite transposon was analyzed. PCR mapping was performed to investigate the excision mechanism, and quantitative real-time PCR was used to quantify the number of circular intermediates.

Results: A novel composite transposon carrying the resistance genes *aac(6')-aph(2''), sat4A, aphA3, erm(B)* and *aadE* flanked by identical copies of the insertion sequence IS*1216E* was identified and found to be integrated in the *panE* gene of *E. faecalis* strain 4638. The transposon was denominated Tn*7086*. DNA sequence analysis showed that Tn*7086* is 24,643 bp-long and contains 29 open reading frames, of which 27 were annotated. PCR analysis demonstrated that Tn*7086* excises from the bacterial chromosome leaving one copy of IS*1216E* in the site of integration and forms circular intermediates in which the ends are joined by the other IS*1216E* copy. Genome comparison with the other 5 aminoglycosides resistant *E. faecalis* strains belonging to CC16/ST480 identified 5 different Tn*7086*-like elements which: (i) are integrated in the chromosomal *panE* gene (ii) excise and form circular intermediates in the same way as Tn*7086*, (iii) show comparable DNA sequences comprehensive of antibiotic resistance gene clusters. Tn*7086*-like elements of strains 5034 and 5410 contain one DNA insert each carrying *lnu(B)-lsa(E)* and *ant9, cat, str* resistance genes, respectively, while both inserts are present in strain 4774. The Tn*7086*-like element of strain 2819 was found to contain a different DNA insert which constitutes another novel composite transposon flanked by two copies of a new IS*Lcr* element and denominated Tn*7087*. Circular forms of Tn*7086* in strain 4638 were present at a concentration of 1.54 x 10^{-5} copies per chromosome, whereas reconstituted integration sites in the bacterial chromosomes were at 6.72 x 10^{-5}. These values were comparable for all the Tn*7086*-like elements.

Conclusion: The present study reports the identification and characterization of the novel composite transposon Tn*7086*, representative of a family of mobile genetic elements which are flanked by *IS1216E* copies, contain aminoglycosides resistance genes and integrate in the chromosomal *panE* gene of infertility-associated *E. faecalis* isolates that cluster in CC16/ST480.

2. INTRODUCTION

Aminoglycosides are potent, broad-spectrum antibiotics which are used as single antimicrobial agents or in combination with other antibiotics in both empirical and definitive therapy (Avent et al., 2011; Jackson et al., 2013; Krause et al., 2016). Aminoglycosides are synergistically used in combination with a cell wall biosynthesis inhibitor to treat certain severe bacterial infections, including complicated enterococcal infections such as enterococcal endocarditis and/or sepsis (Mercuro et al., 2018). The acquisition via horizontal gene transfer of aminoglycoside-modifying enzymes (AMEs) encoding genes confers resistance to high concentrations of aminoglycosides and therefore, restricts the use of these antibiotics (Krause et al., 2016). AMEs are broadly categorized into three groups based on their ability to acetylate (N-acetyltransferase, AACs), phosphorylate (O-phosphotransferases, APHs), or adenylate (O-nucleotidyltransferases, ANTs) amino or hydroxyl groups found at various positions in the aminoglycoside core scaffold (Ramirez & Tolmasky, 2010). The ability of enterococci to acquire mobile genetic elements, conveying antibiotic resistance and virulence traits among both Gram-positive and -negative bacterial species, has contributed to their emergence as leading human pathogens (Courvalin, 1994; Palmer et al., 2010; Pöntinen et al., 2021). *Enterococcus faecalis* and *Enterococcus faecium* are the two enterococcal species most frequently associated with both hospital- and community-acquired infections, such as urinary tract infections, peritonitis, abscesses, endocarditis and bacteriemia (Hidron et al., 2008; Lebreton et al., 2014). Compared to their commensal enterococcal relatives, hospital-adapted *E. faecalis* and *E. faecium* have been reported to contain a larger mobilome which is estimated to account for over a quarter of the genome (Hegstad et al., 2010; Weaver, 2019). Transposable elements constitute the majority of the mobilome in enterococci (Lam et al., 2012; Paulsen et al., 2003; Qin et al., 2012), including: (i) composite transposon (class I transposons), (ii) Tn*3/21* family transposons (class II transposons), and (iii) integrative and conjugative elements (ICEs) comprehensive of the Tn*916*-family of conjugative transposons (Weaver, 2019). Composite transposons owe their intracellular mobility to the presence of flanking copies of insertion sequences (ISs) of the same family that act together to move the DNA between them, while Tn*3*/*21* family transposons are bounded by short inverted repeats and contain both genes needed for transposon movement and accessory genes (Harmer et al., 2020). IS*Ef1*, IS*256* and IS*1216* are the most frequent ISs of *E. faecalis*, with the latter two (IS*256* and IS1*216*) being described to be associated with both simple composite transposons and large, mosaic genetic elements (Weaver, 2019). Recently, epidemiological characterization of a collection of 41 clinical isolates of infertility-associated *E. faecalis* showed the presence of AME genes conferring resistance to high level aminoglycosides in almost 20% (8/41) of the strains (De Giorgi et al., unpublished). Whole genome sequencing analysis of all the 8 aminoglycoside resistant isolates located the AME genes either on the chromosome (6/8 strains) or on plasmids (2/8 strains). Interestingly, all the six strains with chromosomally-located AME genes clustered in the clonal complex (CC)/sequence type (ST) CC16/ST480, showing a clonal structure and suggesting the presence of a common genetic element mediating aminoglycoside resistance in the infertility-associated *E. faecalis* isolates. In the present study, we have identified and characterized a novel composite transposon denominated Tn*7086*. All the six infertility-associated *E. faecalis* strains here investigated carry either Tn*7086* or Tn*7086*-like elements, which share the same genomic traits, including being flanked by identical ISs elements, using the same chromosomal integration site and carrying AMEs genes and other resistance determinants.

3. MATERIALS AND METHODS

3.1. Bacterial strains

E. faecalis strains 2819, 4638, 4774, 5034, 5245 and 5410 were isolated from genital samples of infertile couples (Ricci et al., 2018) and epidemiologically characterized as previously reported (De Giorgi et al., unpublished). *E. faecalis* strain OG1RF was purchased from the American Type Culture Collection and used as a reference control strain. Bacterial strains were grown in liquid Brain Heart Infusion (BHI; Oxoid, Milan, Italy) medium at 37°C.

3.2. Genomic DNA preparation and genome sequencing

Bacteria were harvested by centrifugation at the end of the exponential phase of growth [optical density at 590 nm (OD_{590}) of ~2.0]. Genomic DNA extraction and purification were performed as previously described (De Giorgi et al., unpublished). DNA was precipitated in 2 volumes of ice-cold ethanol, washed with 70% ethanol, and resuspended in 0.9% NaCl. DNA was quantified with Qubit 2.0 Fluorometer (Invitrogen, Life Technologies, Carlsbad, CA, United States) by using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific) and results were confirmed by spectrophotometer measurement (Implen, Munich, Germany). DNA integrity and size were assayed by agarose (0.6%; Seakem LE, Lonza, Rockland, ME USA) gel electrophoresis 0.5X Tris Borate EDTA running buffer. DNA sequencing was performed using both Oxford Nanopore (Oxford Nanopore Technologies, Oxford, UK) and Illumina (Illumina Inc., San Diego, USA) technologies (De Giorgi et al., unpublished). Briefly, Nanopore sequencing library was prepared using the SQK-LSK 108 kit (Oxford Nanopore Technologies), and samples were sequenced on a GridION instrument (Oxford Nanopore Technologies) using an R9.4 flow cell (FLO-MIN106) (Oxford Nanopore Technologies). Illumina sequencing was performed at MicrobesNG (University of Birmingham, UK) using the Nextera library preparation kit (Illumina Inc.) followed by paired-end sequencing (2x250 bp paired-end sequencing) on a NovaSeq 6000 platform (Illumina Inc.).

3.3. Genome assembly and analysis

Hybrid genome assembly of Nanopore and Illumina reads was obtained using Unicycler v0.4.7 (Wick et al., 2017) with default parameters (De Giorgi et al., unpublished). Automatic genome annotation was carried out with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v4.10 (Tatusova et al., 2016). Comparative genomics analyses were performed with BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and Artemis Comparison Tool (ACT) v17.0.1 (Carver et al., 2008). The genome sequence of *E. faecalis* strain OG1RF (GenBank accession number CP002621) was downloaded and used for genome comparison.

3.4. DNA sequence analysis

DNA sequence analysis including coding sequence identification was performed with the software Artemis v. 17.0.1 (Carver et al., 2008). Manual gene annotation of each open reading frame (ORF) was carried out by BLAST homology searching of the databases available at the National Center for Biotechnology Information (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). Protein domains were identified using the protein family database Pfam (https://pfam.xfam.org). Transposon names were assigned by the Tn Registry website curators (https://transposon.lstmed.ac.uk/tn-registry).

3.6. PCR

PCR and sequencing of PCR products were carried as described before (Iannelli et al., 1998; Santoro et al., 2010). Convergent primers designed on the chromosomal regions flanking the insertion site were used to investigate the excision of the transposon from the chromosome, whereas divergent primers designed on the ends of the element served to evaluate its ability to form circular intermediates. Oligonucleotide primers are listed in Table 1.

3.7. Quantitative Real-Time PCR

Real-time PCR experiments were carried out using the KAPA SYBR FAST qPCR kit Master Mix Universal (2X) (Merck KGaA, Darmstadt, Germany) on a LightCycler 1.5 apparatus (Roche Diagnostics GmbH, Mannheim, Germany). Real-time PCR mixture contained, in a final volume of 20 µl, 1×KAPA SYBR FAST qPCR reaction mix, 5 pmol of each primer and 2 µl of bacterial DNA. Thermal profile was an initial 4 min denaturation step at 95°C followed by 40 cycles of repeated denaturation (10 sec at 95°C), annealing (15 sec at 61°C), and elongation (3 min at 72°C). DNA elongation time was increased up to 4 min for strains 4774, 5034 and 5410 and. The temperature transition rate was 20°C/sec in the denaturation and annealing steps and 5°C/sec in the elongation step. A standard curve for the *gyrB* gene was built by plotting the threshold cycle against the number of chromosome copies using serial dilutions of chromosomal *E. faecalis* OG1RF DNA with known concentration. The standard curve was used to quantify the number of:

(i) chromosome copies, (ii) circular intermediates, and (iii) reconstituted site of integration in the chromosome. Quantification of reconstituted integration site in the bacterial chromosome of each strain was performed with the primer pairs IF1344/1345, whereas circular intermediates were quantified using the primer pairs 1404/1405, 1396/1418, 1406/1418, 1482/1418, 1484/1418, 1400/1418 for strains 4638, 2819, 5245, 5410, 5034, 4774, respectively (Table 1). Melting curves were analyzed to differentiate the amplified products from primer dimers. Agarose gel electrophoresis was performed to control the amplification products.

4. RESULTS AND DISCUSSION

4.1. Identification of Tn*7086,* a novel composite transposon carrying aminoglycoside resistance genes

A recent study reported that the infertility-associated *E. faecalis* isolates 2819, 4638, 4774, 5034, 5245 and 5410 belonged to CC16/ST480 and carried the *aac(6')-aph(2'')* gene conferring resistance to gentamicin (De Giorgi et al., unpublished). To investigate the genetic bases of aminoglycoside resistance in these clonally-related strains, whole genomes were obtained and compared to the genome of OG1RF which does not possess any native antibiotic resistance genes (Dunny et al., 1978). Genome comparison analysis revealed that the *aac(6')-aph(2'')* gene was carried by a novel family of composite transposons sharing the same features in all the 6 *E. faecalis* strains: i) being flanked by identical ISs, ii) insertion in the same chromosomal gene and iii) presence of similar DNA sequences that comprise additional genes conferring resistance to other antimicrobials. Nucleotide sequence analysis suggested that the element of *E. faecalis* isolate 4638 was the reference genetic element of the family. Therefore, the transposon of strain 4638 was denominated Tn*7086* and described, whereas elements of the other *E. faecalis* strains were referred as Tn*7086*-like elements and compared to Tn*7086*.

4.2. Nucleotide sequence of Tn*7086*

DNA sequence analysis showed that the transposon Tn*7086* spans nucleotides (nt) 1,554,378 to 1,579,020 of *E. faecalis* 4638 chromosome and therefore it is 24,643 bp-long, with an overall GC content (34.87%) lower than the average of the whole genome (37.45%). The transposon contains 29 ORFs, of which 28 are transcribed in the same direction (Figure 1). Tn*7086* integrates 142 bp upstream the 5' end of the *rbgA* gene, into the *panE* gene which encodes for a 2-dehydropantoate 2-reductase enzyme (Zheng & Blanchard, 2000). Compared to the OG1RF reference strain, the integrated form of Tn*7086* is associated to a 555-bp deletion involving: (i) 427 bp at the 3' end of the *panE* gene ($panE_{516\text{-}942}$), and (ii) 128 bp downstream of *panE*. The transposon is flanked by two direct repeats of the IS*1216E* element, which belong to the IS*6* family (https://isfinder.biotoul.fr/scripts/ficheIS.php?ident=99). Each IS*1216E* is 808bp-long and contains a 681-bp transposase encoding gene flanked by two inverted repeats of 23 bp (5'-GGTTCTGTTGCAAAGTTTTAAAT-3').

4.3. Excision mechanism of Tn*7086*

PCR and PCR sequencing analysis showed that Tn*7086* is able to excise from the bacterial chromosome of strain 4638 producing circular intermediates (Figure 2). The recombination event leading to the excision of the transposon occurs between the two copies of IS*1216E.* Upon excision from the chromosome, a single copy of IS*1216E* remains at the insertion site in the chromosome, whereas the two ends of Tn*7086* circular intermediates are joined by the other IS*1216E* copy (Figure 2). The 555-bp deleted DNA sequence of *panE* was not retrieved in any circular intermediate of Tn*7086.*

4.4. Description of the ORFs in Tn*7086*

Manual homology-based annotation with functional prediction of the hypothetical gene products was possible for 27 out of the 29 predicted ORFs, whereas 2 ORFs, namely *orf*3 and *orf*19, encoded hypothetical proteins that showed no homology to other described sequences (Table 2). All ORFs begin with the ATG starting codon, except for *orf*6, *orf*11, *orf*17 and *orf*24 (GTG) and *orf*15 (TTG). *orf*2 and *orf*13 are identical IS*Ssu5* elements of the IS*1380* family (Chen et al., 2007), arranged in opposite directions. Tn*7086* contains different antimicrobial resistance genes (Table 2 and Figure 1) which are present on other characterized transposons: i) the macrolide-lincosamide-streptogramin resistance determinant *ermB* gene of Tn*917* (Shaw & Clewell, 1985), which is present in two copies in Tn*7086* (*orf*9 and *orf*23); ii) the aminoglycoside-streptothricin resistance gene cluster *ant6-1a'–sat4–aphA-3* (*orf18'-orf17-orf16*) present in Tn*5405*, which is also disseminated among *E. faecium* isolates (Werner et al., 2001); iii) the gentamicin resistance determinant *acc*(6')-*aph*(2'') (*orf20*) found in Tn*4001* (Rouch et al., 1987). *orf*8 and *orf*10 encode a signal peptide and a leader peptide, respectively, associated to *ermB* (*orf*9), whereas *orf*22 codes for a signal peptide associated to the second copy of *ermB* (*orf*23). *orf*11 and *orf*24 code for transcriptional regulators that possibly play a role in the *ermB* gene expression regulation. *orf*4 and *orf*5 constitute a toxin-antitoxin system which may contribute to the maintenance of the element ensuring the persistence of transposon within a bacterial population (Hayes, 2003; Meinhart & Alonso, 2003) and it is likely regulated by *orf*6 which encodes a transcriptional repressor. o*rf*28 is a recombinase that may assist Tn*7086* transposition, whereas *orf*7 codes for a plasmid partition protein which usually plays a role in accurate partitioning of plasmid during cell division (Pratto et al., 2008). *orf*14 encodes an adenine phosphoribosyltransferase essential for purine homeostasis in prokaryotes (Islam et al., 2007). Finally, the transposon also carries several truncated *orfs*. *orf*12, *orf*25, *orf*27 encoding DNA topoisomerase (TopB) are truncated: *orf*12 is a 1,474-bp duplication of the 3' end of *orf*25, whereas *orf*25 and *orf*27 contain the 3' and the 5' end DNA sequence of the undisrupted DNA topoisomerase gene, respectively. *orf*26 codes for a group II intron, which is reported to have activities of endonuclease, RNA maturase and reverse transcriptase (Lambowitz & Zimmerly, 2011) and disrupts the sequence of the DNA topoisomerase encoding genes (*orf*25 and *orf*27). Similarly, *orf*15, *orf*18 and *orf*21, encoding aminoglycoside 6-adenyltransferases, are also truncated: *orf*18 is a duplication of the 3' end of *orf*15, which in turn contains the 3' end sequence of the original *ant6-1a* gene, whereas *orf*21 contains the 5' end (Figure 1 and Table 2).

4.5. BLAST sequence analysis of Tn*7086*

BLAST analysis of DNA sequence similarity using as a query the 24,643 bp-long Tn*7086* showed that Tn*7086*-like elements are present in the genomes of many *E. faecalis* and *E. faecium* strains and also in few other bacterial species, such as *Streptococcus suis* and *Staphylococcus aureus*. Compared to previously described enterococcal mobile genetic elements, the DNA sequence of transposon Tn*7086* shows homology to i) plasmid p16-164 of *E. faecium* ((Dejoies et al., 2021); acc. no. CP065774) with a total of 76% of DNA sequence homology that covers all *orfs*, except for those encoding the toxin-antitoxin system and the two IS*su5* copies and to ii) *E. faecalis* conjugative plasmid pRE25 ((Schwarz et al., 2001); acc. no. NC_008445) and therefore to plasmids pE35048-oc ((Morroni et al., 2018); acc. no. MF580438) and pWZ909 ((Zhu et al., 2010); acc. no. GQ484954), which share the pRE25 backbone, with a total of 65% of DNA sequence homology involving *orf*1, *orfs*3 to 12*, orfs*15 to 19 and *orf*29; iii) homologous to the previously described Tn*6349* of *S. aureus* AOUC-0915 ((D'Andrea et al., 2019); acc. no. MH746818), with which Tn*7086* shares a common backbone that includes the boundary IS*1216E* elements, the toxin-antitoxin system, one undisrupted DNA topoisomerase TopB encoding gene, and *ermB*, thus suggesting that a common transposon ancestor may have diverged in *E. faecalis* and *S. aureus* acquiring different antimicrobial resistance genes; iv) homologous to $\text{MES}_{6272\text{-}2}$-like structure of *E. faecium* V19 ((Lin et al., 2020); acc. no. MT877068) with 42% of DNA sequence homology including all antibiotic resistance genes and the IS*1216E* elements (*orf*1, *orfs*8-10, *orfs*14-21, *orf*29) (Supplementary Table S1).

4.6. Identification and structure of Tn*7086*-like composite transposons in aminoglycoside-resistant infertility-associated *E. faecalis* isolates

Aminoglycoside-resistant infertility-associated *E. faecalis* strains 2819, 4774, 5034, 5245 and 5410 belonging to CC16/ST480 were searched for the presence of Tn*7086*-like mobile elements.

DNA sequence analysis of the transposon-chromosome junction region indicated that all the Tn*7086*-like elements were integrated within the chromosomal *panE* gene located between the *rbgA* and *cynR* genes. Upon integration, 3 cases were identified: (i) a 555-bp deletion involving the 3'-end of *panE* and 128 nucleotides of the downstream intergenic region was found in strains 5245 and 5034 as already described for Tn*7086* of strain 4638; (ii) an 8-bp (AGCCAGCG) target site duplication of nucleotides 509-516 of *panE* occurred in strains 4774 and 5410; (iii) a 624-bp deletion involving a larger portion of the 3'-end of *panE* that includes the 8-bp target site and the same 128 nucleotides downstream was identified in strain 2819 (Figure 3). Tn*7086*-like elements vary in length from 15.4-kb (strain 5410) up to 35.3-kb (strain 2819), but they are all flanked by two identical copies of IS*1216E* (Figure 4). Comparison of the Tn*7086* DNA sequence with that of the other five Tn*7086*-like transposons indicates that all five genetic elements have a 398-bp insertion at nt 809 carrying a truncated *orf* encoding a putative homologous to the chaperon protein *DnaJ* homologous. A 5'-end deletion of the IS*Ssu5* element (*orf*2 of Tn*7086*) is present in the transposons of both strains 4774 and 5034, whereas in strain 5410 there is a larger (8,072 bp) 5'-end deletion ranging from nt 809 of *orf*2 to nt 8,881 of *orf*12. A 3-end deletion of 3,188-bp from nt 20,645 to nt 23,833 (part of *orf26* and *orfs 27-28*) was found in strains 4774, 5034 and 5410, whereas the 3'-end deletion it was larger in strain 5245 (5,163 bp) spanning nt 18,670 to nt 23,833 thus including the entire *orf*26 and the 3' end of *orf25*. These 3'-end deletions may have been caused by the autocatalytic RNA activity of the group II intron transcript (Lambowitz & Zimmerly, 2011). A further 1,973-bp deletion was identified at nucleotide11,633 in the central region of *orf*15 spanning to *orf*17 in strain 5410. In addition to the aminoglycoside resistance genes carried by all the composite transposons of the Tn*7086* family, strains 4774 and 5034 also contain a 10,666-bp DNA sequence carrying the *lnu(B)-lsa(E)* and *ant9* resistant genes (Table 3), inserted at nt 11,633 of *orf15* and producing a 1,153-bp direct duplication involving *orf*15 and *orf*16. Moreover, strains 4774 and 5410 present a 4,891 bp-long DNA insert at nt 23,833 of *orf*29 carrying the *cat* and *str* genes (Table 3), which generates an 809-bp direct duplication involving *orf29* (Figure 4). Finally, a distinct genetic element was found inserted at nucleotide 23,833 (*orf*29) in strain 2819 and named Tn*7087* (Figure 4). Tn*7087* is 9,925 bp-long, has a 31.9% GC content and it is flanked by two copies of a novel IS arranged in the same orientation, but characterized by different length and nucleotide sequence and therefore referred as IS*Lcr*L and IS*Lcr*R (Supplementary Figure S1). IS*Lcr*L is 654 bp-long and contains 34 nucleotide changes and 1 nucleotide deletion compared to IS*Lcr*R, whereas both are bounded by two imperfect inverted repeats of 14 bp (5'-ATATTAAGTGCAAA-3' and 5'-TTTGCCATTTAAAT-3). PCR and sequencing analysis showed that Tn*7087* excises from Tn*7086* and produces a circular form and a deletion in the Tn*7086* of strain 2819. As previously described for Tn*7086*, excision of Tn*7087* also occurs by recombination between the flanking IS*Lcr* copies, leaving the IS*Lcr*L in Tn*7086*, whereas the ends of Tn*7087* in the circular forms are joined by IS*Lcr*R (Supplementary Figure S1).

4.7. Quantitative analysis of excision of Tn*7086* and Tn*7086*-like elements from *E. faecalis* strains

To analyze the frequency of excision from the *E. faecalis* chromosome of the six different transposons belonging to the Tn*7086* family, real-time PCR was used to quantify the number of copies of circular intermediates and reconstituted integration chromosomal sites (Table 4). The number of circular intermediates varied with the strain ranging from 1.3±0.16 (strain 5410) to 22.4±17.7 (strain 5245) copies per 10^6 chromosomes. Calculation of the number of reconstituted integration sites showed a range from 6.11±2.18 (strain 4774) to 67.2±15.1 (strain 4638) copies per 10^6 chromosomes, indicating a 3-4.7-fold higher copy numbers compared to circular forms and suggesting that some circular intermediates of Tn*7086* may be lost after excision from the chromosome (Table 4). Of note is the behavior of strain 5245, where the number of reconstituted integration site was exceedingly higher reaching 7280 copies per million of *E. faecalis* chromosomes (Table 4).

5. CONCLUSIONS

In this study, we described a novel family of composite transposons represented by Tn*7086* of strain 4638 which harbors the consensus DNA sequence of the family. Tn*7086* and Tn*7086*-like i) are flanked by two direct repeats of the IS*1216E*, ii) integrate in the chromosomal *panE* gene and iii) carry multiple antibiotic resistance determinants including *ermB*, *ant6-1a'–sat4–aphA-3*, *acc*(6')-*aph*(2''), *lnu(B)-lsa(E)*, *ant9*, *cat* and *str*. Upon transposon integration, 3 scenarios regarding the *panE* gene were identified: 2 cases involving a partial deletion of *panE* and 1 case in which *panE* was disrupted with a consequent 8-bp internal target site duplication (Figure 3). Tn*7086* and Tn*7086*-like elements were demonstrated to excise from the bacterial chromosome leaving a copy of the IS*1216E* in the integration site and producing circular intermediates in which the ends were joined by the other IS*1216E* copy. Thus, it can be concluded that the recombination event driving the excision of the transposons occurred between the two copies of IS*1216E* with no reconstitution of the *panE* gene. This type of movement has been previously described in both Gram-negative and -positive bacteria for mobile genetic elements bounded by IS*26* and IS*1216E* copies, respectively (Harmer et al., 2014, 2020; Harmer & Hall, 2016, 2017). Three types of DNA inserts were identified by DNA sequence comparison of Tn*7086* with other Tn*7086*-like elements: i) a 4.8-kb insert (strains 4774 and 5410) containing a chloramphenicol acetyltransferase and a streptomycin adenylyltransferase encoding resistance genes, ii) a 10.6-kb insert (strains 4774 and 5034) containing an aminoglycoside nucleotidyltransferase encoding gene and iii) a 9.9-kb insert (strain 2819) found to be a composite transposon itself and named Tn*7087*. Tn*7087* was flanked by two new ISs and was demonstrated to move like Tn*7086*. Interestingly, the 10.6-kb and the 4.8-kb inserts were found inserted at specific nt positions (orf*15* and orf*29*), respectively, suggesting that these positions may act as hotspots for insertions of additional genetic material.

TABLES

Table 1. **Oligonucleotide primers.**

Name	Sequence (5’ to 3’)	Strain: genome position
IF1344	GCCTGTTCACGAGCCAATTT	2819: 1,526,733 – 1,526,752
IF1345	CGCTATGGGCAGTCGCTTT	2819: 1,562,987 – 1,562,969
IF1396	GGCACAATCACGGTAACTCAA	2819: 1,560,295 – 1,560,315
IF1397	CCTATTCGTACACTCTATCGTT	2819: 1,553,452 – 1,553,431
IF1418	ACACCCGAACAGTTTAAGGATA	2819:1,528,105 – 1,528,084
IF1421	GCCATTTTCAACCAACCTCTAA	2819: 1,551,029 – 1,551,050
IF1422	ACAGAACCCTAATATCTCCTT	2819: 1,561,768 – 1,561,748
IF1400	GTGTGAGAGATAGCAATAGATTTA	4774: 1,571,687 – 1,571,710
IF1404	GCCATTTTCAACCAACCTCTAA	4638: 1,560,295 – 1,560,315
IF1405	CCTATTCGTACACTCTATCGTT	4638: 1,553,452 – 1,553,431
IF1406	TCCTGAAGTGATTACATCTGTA	5245: 1,547,193 – 1,547,214
IF1482	TTTGGAAGAAAGTATCTGCCTA	5410: 1,556,827 – 1,556,806
IF1484	TGCTTCTAAGTCTTATTTCCATAA	5034: 1,508,257 – 1,508,280
IF1496	GTTGCCACACTTAGGACATTT	5034: 1,508,805 – 1,508,825

***Table 2.* Annotated ORFs of Tn*7086*.**

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein		
			Protein ID /Origin [E value][c]	**aa identity**	**aa similarity**
*orf*1 (226)	IS*1216E*, transposase, IS*6* family (Arthur et al., 1997)		AAC44739 /*E. faecium* [3e-171]	223/225 (99%)	223/225 (99%)
*orf*2 (439)	IS*Ssu5*, transposase, IS*1380* family (Chen et al., 2007)		ABP92124 /*S.suis* [0.0]	439/439 (100%)	439/439 (100%)
*orf*4 (287)	Zeta toxin (Meinhart & Alonso, 2003)	Zeta_toxin (17-217) [1.6e-47]	1GVN_B /*S. pyogenes* [0.0]	278/287 (97%)	283/287 (98%)
*orf*5 (90)	Epsilon antitoxin (Meinhart & Alonso, 2003)	Epsilon_antitox (2-90) [3.5e-56]	1GVN_A /*S. pyogenes* [9e-65]	90/90 (100%)	90/90 (100%)
*orf*6 (69)	Omega transcriptional repressor (Murayama et al., 2001)	Omega_Repress (1-69) [6.5e-47]			
*orf*7 (298)	Plasmid partition protein A (Pratto et al., 2008)	AAA_31 (35-214) [3.3e-32]	2OZE_A /*S. suis* [0.0]	293/298 (98%)	297/298 (99%)
*orf*8 (43)	Signal peptide (Shaw & Clewell, 1985)		AAA27453 /*E. faecalis* Tn*917* [2e-31]	43/43 (100%)	43/43 (100%)
*orf*9 (245)	23S rRNA adenine N-6-methyltransferase (Shaw & Clewell, 1985)	RrnaAD (1-242) [2.7e-75]	AAA27452 /*E. faecalis* Tn*917* [0]	243/245 (99%)	244/245 (99%)
*orf*10 (31)	23S rRNA methylastransferase leader peptide (Shaw & Clewell, 1985)	ErmC (1-31) [6.7e-26]	AAA27451 /*E. faecalis* Tn*917* [9e-16]	26/31 (84%)	26/31 (83%)
*orf*11 (54)	Omega transcriptional repressor (Murayama et al., 2001)	Omega_Repress (1-53) [1.9e-27]			

ORF (aa)[a]	**Annotation and comments (reference)**	**Pfam domain[b] [E value]**	**Homologous protein**		
			Protein ID /Origin [E value][c]	**aa identity**	**aa similarity**
*orf*12 (495)	DNA topoisomerase, truncated (Lima et al., 1994)	Topoisom_bac (2-373) [4.1e-77] zf-C4_Topoisom (411-443). [3.9e-11]	EGQ1712264/*S. pseudointermedius* [0.0]	483/489 (99%)	486/489 (99%)
*orf*13 (439)	IS*Ssu5*, transposase, IS*1380* family (Chen et al., 2007)		ABP92124 /*S.suis* [0.0]	439/439 (100%)	439/439 (100%)
*orf*14 (175)	Adenine phosphoribosyltransferase (Islam et al., 2007)	Pribosyltran (25-173) [4.8e-17]			
*orf*15 (237)	Aminoglycoside 6-adenylyltransferase, truncated (Werner et al., 2001)	Adenyl_transf (59-227) [2.7e-60]	AAK62560 /*E. faecium* [3e-98]	136/164 (83%)	149/164 (90%)
*orf*16 (264)	Aminoglycoside 3'-phosphotransferase (Werner et al., 2001)	APH (25-257) [3.2e-26]	AAK62562 /*E. faecium* [0.0]	263/264 (99%)	264/264 (100%)
*orf*17 (180)	Streptogramin A acetyltransferase (Werner et al., 2001)	Acetyltransf_1 (35-153) [5.3e-20]	AAK62561 /*E. faecium* [1e-122]	180/180 (100%)	180/180 (100%)
*orf*18 (233)	Aminoglycoside 6-adenylyltransferase, truncated (Werner et al., 2001)	Adenyl_transf (1-208) [1.7e-82]	AAK62560 /*E. faecium* [2e-180]	233/233 (100%)	233/233 (100%)
*orf*20 (479)	Bifunctional AAC(6')-APH(2'') (Rouch et al., 1987)	APH (204-440) [1.4e-23] Acetyltransf_8 (13-159) [9.2e-23]	AAA88548 /*S. aureus* pSK1 [0.0]	478/479 (100%)	479/479 (100%)
*orf*21 (173)	Aminoglycoside 6-adenylyltransferase, truncated (Werner et al., 2001)	Acetyltransf_1 (36-154) [1.1e-17]	AAK62560 /*E. faecium* [2e-20]	37/37 (100%)	37/37 (100%)

ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein		
			Protein ID /Origin [E value][c]	aa identity	aa similarity
		Adenyl_transf (1-70) [1.4e-11]			
*orf*22 (43)	Signal peptide (Shaw & Clewell, 1985)		AAA27453 /*E. faecalis* Tn*917* [2e-31]	43/43 (100%)	43/43 (100%)
*orf*23 (245)	23S rRNA adenine N-6-methyltransferase (Shaw & Clewell, 1985)	RrnaAD (1-242) [2.7e-75]	AAA27452 /*E. faecalis* Tn*917* [0.0]	243/245 (99%)	244/245 (99%)
*orf*24 (77)	Omega transcriptional repressor (Murayama et al., 2001)	Omega_Repress (1-69) [2.0e-46]			
*orf*25 (558)	DNA topoisomerase, truncated (Lima et al., 1994)	Topoisom_bac (2-436) [9.9e-105]	EGQ1712264 /*S. pseudointermedius* [0.0]	557/557 (100%)	557/557 (100%)
*orf*26 (628)	Group II intron reverse transcriptase/maturase	RVT_1 (105-352) [8.1e-38]	WP_160459188 /*S.aureus* [0.0]	626/628 (99%)	626/628 (99%)
*orf*27 (120)	DNA topoisomerase, truncated (Aravind, 1998)	Toprim (3-119) [5.6e-15]	EGQ1712264 /*S. pseudointermedius* [9e-82]	116/117 (99%)	116/117 (99%)
*orf*28 (205)	DNA resolvase (Yang & Steitz, 1995)	Resolvase (3-142) [1.3e-33]			
*orf*29 (226)	IS*1216E*, transposase, IS*6* family (Arthur et al., 1997)		AAC44739 /*E. faecium* [3e-171]	223/225 (99%)	223/225 (99%)

[a] The number of amino acids of each ORF is shown in parenthesis. *orf*3 and *orf*19 are not reported because no homology with previously described sequences was found.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

Table 3. **Annotated ORFs of the DNA inserts of Tn*7086*-like elements of *E. faecalis* strains 2819, 4774, 5034 and 5410, identified by DNA sequence comparison with Tn*7086* of strain 4638.**

Strain	Length (bp) of DNA insert	ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein		
					Protein ID/ Origin [E value][c]	aa identity	aa similarity
2819	9,925[d]	*orf*1 (267)	IS*Lcr*				
		*orf*2 (475)	DNA integrase (Dyda et al., 1994)	Integrase core domain rve (154-284) [9.0e-23] Mu transposase, C-terminal (351-411) [1.9e-15] HTH_28 (19-70) [9.6e-10]			
		*orf*3 (194)	DNA resolvase (Yang & Steitz, 1995)	Resolvase (2-138) [1.6e-44] HTH_7 (140-183) [2.2e-15]			
		*orf*4 (680)	Histidinol-phosphatase, putative		MVH72510.1 *S.aureus* [0.0]	572/573 (99%)	573/573 (100%)
		*orf*5 (408)	Cell wall protein containing LPxTG motif, putative	Gram_pos_anchor (365-408) [0.00061]			
		*orf*6 (145)	IS*Lcr*				
4774, 5034	10,666[e]	*orf*1 (244)	Class I S-Adenosyl-l-methionine (SAM)-dependent methyltransferase	Methyltransf_11 (47-141) [3.8e-25]			

Strain	Length (bp) of DNA insert	ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein		
					Protein ID/ Origin [E value][c]	aa identity	aa similarity
		*orf*2 (289)	Nucleotidyltransferase (Holm & Sander, 1995)	NTP_transf_2 (4-116) [4.8e-12]			
		*orf*3 (74)	Transcriptional regulator, putative (Brennan & Matthews, 1989)	HTH_26 (6-63) [2.7e-10]			
		*orf*4 (359)	IS*Vlu1,* transposase, IS*L3* family		QZN88035 *Vagococcus lutrae* [0.0]	264/358 (74%)	307/358 (85%)
		*orf*5 (267)	Lincosamide nucleotidyltransferase Lnu(B)	Polbeta (9-96) [4.3e-05]			
		*orf*6 (494)	ABC-F type ribosomal protection protein Lsa(E)	ABC_tran (21-146) [1.6e-15] ABC_tran (325-451) [4.2e-19]			
		*orf*7 (144)	DNA recombinase, truncated (Singleton et al., 2001)	RecG_N (15-64) [0.5] RecG_N (42-130) [0.36]	WP_198518244 *E. faecium* [7e-97]	144/144 (100%)	144/144 (100%)
		*orf*9 (269)	Aminoglycoside nucleotidyltransferase	Polbeta (18-100) [3.7e-08]			
		*orf*10 (112)	Adenine phosphoribosyltransferase (Islam et al., 2007)	Pribosyltran (24-91) [0.19]			
		*orf*11 (237)	Aminoglycoside 6-adenylyltransferase,	Adenyl_transf (59-227) [2.7e-60]	AAK62560 *E. faecium* [3e-98]	136/164 (83%)	149/164 (90%)

Strain	Length (bp) of DNA insert	ORF (aa)[a]	Annotation and comments (reference)	Pfam domain[b] [E value]	Homologous protein		
					Protein ID/ Origin [E value][c]	aa identity	aa similarity
			truncated (Werner et al., 2001)				
4774, 5410	4,891[f]	*orf*1 (226)	IS*1216E*, transposase, IS*6* family (Arthur et al., 1997)	Methyltransf_11 (47-141) [3.8e-25]	AAC44739 *E. faecium* [3e-171]	223/225 (99%)	223/225 (99%
		*orf*2 (94)	Replication initiation protein, putative				
		*orf*3 (215)	Chloramphenicol acetyltransferase	CAT (3-205) [9.6e-90]			
		*orf*4 (264)	Replication initiation protein	Mob_Pre (1-115) [3.8e-33] Rep_trans (89-261) [3.0e-08]			
		*orf*5 (282)	Streptomycin adenyltransferase	Adenyl_transf (1-279) [1.8e-98]			

[a] The number of amino acids of each ORF is shown in parenthesis.

[b] Numbers in parentheses represent the part of the protein homologous to the Pfam domain.

[c] Determined by compositional matrix adjustment.

[d] A 9.9-kb-long DNA sequence was inserted at nucleotide 23,833 of strain 2819; this 9.9-kb-long is a composite transposon named Tn*7087*.

[e] A 10.6-kb-long DNA sequence was inserted at nucleotide 11,633 of strains 4774 and 5034.

[f] A 4.8-kb-long DNA sequence was inserted at nucleotide 23,833 of strains 4774 and 5410.

Table 4. **Frequency of excision of Tn*7086* in *E. faecalis* strain 4638 and of Tn*7086*-like composite transposons in the other *E. faecalis* isolates**[a]**.**

Strain	Circular forms	Reconstituted integration site
4638	1.54×10^{-5} ($\pm 4.05 \times 10^{-6}$)	6.72×10^{-5} ($\pm 1.51 \times 10^{-5}$)
2819	3.24×10^{-6} ($\pm 1.81 \times 10^{-7}$)	5.88×10^{-5} ($\pm 3.22 \times 10^{-6}$)
4774	6.98×10^{-6} ($\pm 3.62 \times 10^{-6}$)	6.11×10^{-6} ($\pm 2.18 \times 10^{-6}$)
5245	2.24×10^{-5} ($\pm 1.77 \times 10^{-5}$)	7.28×10^{-3} ($\pm 9.67 \times 10^{-4}$)
5410	1.30×10^{-6} ($\pm 1.64 \times 10^{-7}$)	7.23×10^{-6} ($\pm 6.98 \times 10^{-7}$)
5034	5.09×10^{-6} ($\pm 6.67 \times 10^{-8}$)	1.30×10^{-5} ($\pm 4.14 \times 10^{-6}$)

[a] The frequency of excision is expressed as number of copies of circular forms or reconstituted chromosomal integration sites per bacterial chromosome

***Supplementary Table S1.* Results of BLAST homology search to Tn*7086*[a].**

Sequence name	**GenBank accession no.**	**Origin**	**Reference**	**Percentage of query coverage (Tn*7086* ORFs)**
lsa(E)-carrying resistance gene cluster	MG765453	*E. faecalis* E512		87% (*orf*1-26, 29)
p16-164	CP065774	*E. faecium* 16-164	(Dejoies et al., 2021)	76% (1, 6, 8-12, 14-29)
IS*1216V*, Tn*551*	LC125351	*S. aureus* 6272		72% (1, 7, 7-11, 14-29)
pRE25	NC_008445	*E. faecalis* RE25	(Schwarz et al., 2001)	65% (1,3-12, 15-19, 29)
pE35048-oc	MF580438	*E. faecium* E35048	(Morroni et al., 2018)	57% (3-11, 22-29)
pWZ909	GQ484954	*E. faecalis*	(Zhu et al., 2010)	56% (3-12, 23-29)
Tn*6349*	MH746818	*S. aureus* AOUC-0915	(D'Andrea et al., 2019)	53% (1, 3-12, 26-29)
pVEF4	MG674582	*E. faecium* HL1	(Leinweber et al., 2018)	53% (1, 3, 6-7, 12, 27-29)
partial pVEF4	FN424376	*E. faecium* 399/F98/A4	(Sletvold et al., 2010)	53% (1, 3, 6-7, 12, 24-29)
IS*1216V*	LC125352	*S. aureus* 2250		52% (orf1, 6, 8-12, 22-29)
lsa(E)-carrying multidrug resistance gene cluster	KX156278	*E. faecalis* E533		46% (orf2, 8-9, 13-23)
pVEF1	AM296544	*E. faecium*	(Sletvold et al., 2007)	43%
pAMB1	GU128949	*E. faecalis* DS5		43%
MESPM1 and MES6272-2	MT877068	*E. faecium* V19 plasmid pV19	(Lin et al., 2020)	42% (1, 8-10, 14-21, 29)
pKM0218	MF477836	*Macrococcus canis*	(Chanchaithong et al., 2019)	42% (orf1, 8-11, 15-23, 29)

Sequence name	**GenBank accession no.**	**Origin**	**Reference**	**Percentage of query coverage (Tn*7086* ORFs)**
pRUM	AF507977	*E. faecium*	(Grady & Hayes, 2003)	34%
lsa(E)-carrying multidrug resistance gene cluster	KX712118	*Staphylococcus epidermidis*	(Deng et al., 2017)	34% (8-9, 14-23)
Tn*6215*	KC166248	*Clostridium difficile* CD80	(Goh et al., 2013)	33% (5-9, 11-12, 22-25)
ABC-F type ribosomal protection protein OptrA (optrA), HNH endonuclease, hypothetical protein, ABC transporter ATP-binding protein, ParA partitioning protein (parA), putative replication protein, RepS (rep), hypothetical protein, site-specific recombinase, resolvase family, and group II intron reverse transcriptase/maturase genes, complete cds	MT723949	*Enterococcus thailandicus*	(Fioriti et al., 2020)	28%
Plasmid pIlo8 omega2 gene, *ermL* gene, *ermIP* gene, ORF3, delta gene, omega gene, epsilon gene and ORFZ	AJ549242	pIlo8	(Zúñiga et al., 2003)	26%
Tn*6003*	AM410044	*S. pneumoniae* Ar4	(Cochetti et al., 2007)	26%
Integrative and conjugative element ICESsu05SC260	KX077888	*S. suis*	(Huang et al., 2016)	24%
ATP-binding protein gene, partial cds; and membrane protein, peptidase P60, conjugal transfer protein, ABC transporter ATP-binding protein, ABC-2 transporter	MN625138	*Clostridioides difficile* 1-11	(Zhao et al., 2020)	24%

Sequence name	**GenBank accession no.**	**Origin**	**Reference**	**Percentage of query coverage (Tn*7086* ORFs)**
permease, 23S rRNA-methyltransferase Erm(B), ParA family protein				
pSCBC1	CP038169	*E. faecium* SCBC1	(Lei et al., 2019)	23%
pYN2-1	CP038173	*E. faecium* YN2-1	(Fioriti et al., 2020)	23%
IS*Ssu5* composite transposon	KP998101	*S. aureus* LAR2682		22%
Tn*6194*	HG475346	*Clostridium difficile* CII7		22%
aadE-sat4-aphA-3 gene cluster	JQ655275	*Campylobacter coli* SX81	(S. Qin et al., 2012)	21%
pBEE99	GU046453	*E. faecalis* E99	(Tendolkar et al., 2006)	19%
Tn*1116*	AM411377	S. *pyogenes* A-3	(Brenciani et al., 2007)	17%
CVM N48037F	CP028720	*E. faecalis* CVM N48037F	(Tyson et al., 2018)	16%
pTEF2	CP046110	*E. faecalis* 133170041-3		4%
pAD1	CP046109	*E. faecalis* 133170041-3		3%

[a] Analysis was performed to search NCBI using the DNA sequence of Tn*7086* as the query. Only mobile genetic elements/gene clusters previously identified are reported (homology results with genome assemblies are not included)

Aknowledgments

Illumina genome sequencing was provided by MicrobesNG.

6. REFERENCES

Aravind, L. (1998). Toprim—A conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Research*, *26*(18), 4205–4213. https://doi.org/10.1093/nar/26.18.4205.

Arthur, M., Depardieu, F., Gerbaud, G., Galimand, M., Leclercq, R., & Courvalin, P. (1997). The VanS sensor negatively controls VanR-mediated transcriptional activation of glycopeptide resistance genes of Tn*1546* and related elements in the absence of induction. *Journal of Bacteriology*, *179*(1), 97–106. https://doi.org/10.1128/jb.179.1.97-106.1997.

Avent, M. L., Rogers, B. A., Cheng, A. C., & Paterson, D. L. (2011). Current use of aminoglycosides: Indications, pharmacokinetics and monitoring for toxicity: Aminoglycosides: review and monitoring. *Internal Medicine Journal*, *41*(6), 441–449. https://doi.org/10.1111/j.1445-5994.2011.02452.

Brenciani, A., Bacciaglia, A., Vecchi, M., Vitali, L. A., Varaldo, P. E., & Giovanetti, E. (2007). Genetic Elements Carrying *erm* (B) in *Streptococcus pyogenes* and Association with *tet* (M) Tetracycline Resistance Gene. *Antimicrobial Agents and Chemotherapy*, *51*(4), 1209–1216. https://doi.org/10.1128/AAC.01484-06.

Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif. *Journal of Biological Chemistry*, *264*(4), 1903–1906. https://doi.org/10.1016/S0021-9258(18)94115-3.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B. G., Parkhill, J., & Rajandream, M.-A. (2008). Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, *24*(23), 2672–2676. https://doi.org/10.1093/bioinformatics/btn529.

Chanchaithong, P., Perreten, V., & Schwendener, S. (2019). *Macrococcus canis* contains recombinogenic methicillin resistance elements and the *mecB* plasmid found in *Staphylococcus aureus*. *Journal of Antimicrobial Chemotherapy*, *74*(9), 2531–2536. https://doi.org/10.1093/jac/dkz260.

Chen, C., Tang, J., Dong, W., Wang, C., Feng, Y., Wang, J., Zheng, F., Pan, X., Liu, D., Li, M., Song, Y., Zhu, X., Sun, H., Feng, T., Guo, Z., Ju, A., Ge, J., Dong, Y., Sun, W.,Yu, J. (2007). A glimpse of streptococcal toxic shock syndrome from comparative genomics of *S. suis* 2 chinese isolates. *PLoS ONE*, *2*(3), e315. https://doi.org/10.1371/journal.pone.0000315.

Cochetti, I., Tili, E., Vecchi, M., Manzin, A., Mingoia, M., Varaldo, P. E., & Montanari, M. P. (2007). New Tn*916*-related elements causing *erm(B)*-mediated erythromycin resistance in tetracycline-susceptible pneumococci. *Journal of Antimicrobial Chemotherapy*, *60*(1), 127–131. https://doi.org/10.1093/jac/dkm120.

Courvalin, P. (1994). Transfer of antibiotic resistance genes between Gram-Positive and Gram-negative Bacteriat. *Antimicrob. Agents Chemother.*, *38*, 5.

D’Andrea, M. M., Antonelli, A., Brenciani, A., Di Pilato, V., Morroni, G., Pollini, S., Fioriti, S., Giovanetti, E., & Rossolini, G. M. (2019). Characterization of Tn*6349*, a novel mosaic transposon carrying *poxtA*, *cfr* and other resistance determinants, inserted in the chromosome of an ST5-MRSA-II strain of clinical origin. *Journal of Antimicrobial Chemotherapy*, *74*(10), 2870–2875. https://doi.org/10.1093/jac/dkz278.

Dejoies, L., Sassi, M., Schutz, S., Moreaux, J., Zouari, A., Potrel, S., Collet, A., Lecourt, M., Auger, G., & Cattoir, V. (2021). Genetic features of the *poxtA* linezolid resistance gene in human enterococci from France. *The Journal of Antimicrobial Chemotherapy*, *76*(8), 1978–1985. https://doi.org/10.1093/jac/dkab116.

Deng, F., Wang, H., Liao, Y., Li, J., Feßler, A. T., Michael, G. B., Schwarz, S., & Wang, Y. (2017). Detection and genetic environment of pleuromutilin-lincosamide-streptogramin a resistance genes in Staphylococci isolated from pets. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00234.

Dunny, G. M., Brown, B. L., & Clewell, D. B. (1978). Induced cell aggregation and mating in *Streptococcus faecalis*: Evidence for a bacterial sex pheromone. *Proceedings of the National Academy of Sciences*, *75*(7), 3479–3483. https://doi.org/10.1073/pnas.75.7.3479.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R., & Daviest, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integras: similarity to other polynucleotidyl transferases. *Science*, 23;266(5193):1981-6. https://doi:10.1126/science.7801124.

Fioriti, S., Morroni, G., Coccitto, S. N., Brenciani, A., Antonelli, A., Di Pilato, V., Baccani, I., Pollini, S., Cucco, L., Morelli, A., Paniccià, M., Magistrali, C. F., Rossolini, G. M., & Giovanetti, E. (2020). Detection of oxazolidinone resistance genes and characterization of genetic environments in *Enterococci* of swine origin, Italy. *Microorganisms*, *8*(12), 2021. https://doi.org/10.3390/microorganisms8122021.

Goh, S., Hussain, H., Chang, B. J., Emmett, W., Riley, T. V., & Mullany, P. (2013). Phage ϕC2 mediates transduction of Tn*6215*, encoding erythromycin resistance, between *Clostridium difficile* Strains. *MBio*, *4*(6). https://doi.org/10.1128/mBio.00840-13.

Grady, R., & Hayes, F. (2003). Axe-Txe, a broad-spectrum proteic toxin-antitoxin system specified by a multidrug-resistant, clinical isolate of *Enterococcus faecium*: Toxin-antitoxin system in *Enterococcus*. *Molecular Microbiology*, *47*(5), 1419–1432. https://doi.org/10.1046/j.1365-2958.2003.03387.

Harmer, C. J., & Hall, R. M. (2016). IS *26* -Mediated formation of transposons carrying antibiotic resistance genes. *MSphere*, *1*(2). https://doi.org/10.1128/mSphere.00038-16.

Harmer, C. J., & Hall, R. M. (2017). Targeted conservative formation of cointegrates between two DNA molecules containing IS *26* occurs via strand exchange at either IS end: IS*26* targeted conservative cointegrate formation. *Molecular Microbiology*, *106*(3), 409–418. https://doi.org/10.1111/mmi.13774.

Harmer, C. J., Moran, R. A., & Hall, R. M. (2014). Movement of IS *26* -Associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS *26* and preferentially inserts adjacent to another IS *26*. *MBio*, *5*(5). https://doi.org/10.1128/mBio.01801-14.

Harmer, C. J., Pong, C. H., & Hall, R. M. (2020). Structures bounded by directly-oriented members of the IS*26* family are pseudo-compound transposons. *Plasmid*, *111*, 102530. https://doi.org/10.1016/j.plasmid.2020.102530.

Hayes, F. (2003). Toxins-Antitoxins: Plasmid maintenance, programmed cell death, and cell cycle arrest. *Science*, *301*(5639), 1496–1499. https://doi.org/10.1126/science.1088157.

Hegstad, K., Mikalsen, T., Coque, T. M., Werner, G., & Sundsfjord, A. (2010). Mobile genetic elements and their contribution to the emergence of antimicrobial resistant *Enterococcus faecalis* and *Enterococcus faecium*. *Clinical Microbiology and Infection*, *16*(6), 541–554. https://doi.org/10.1111/j.1469-0691.2010.03226.

Hidron, A. I., Edwards, J. R., Patel, J., Horan, T. C., Sievert, D. M., Pollock, D. A., Fridkin, S. K., & National Healthcare Safety Network Team and Participating National Healthcare Safety Network Facilities. (2008). Antimicrobial-resistant pathogens associated with healthcare-associated infections: annual summary of data reported to the national healthcare safety network at the centers for disease control and prevention, 2006–2007. *Infection Control & Hospital Epidemiology*, *29*(11), 996–1011. https://doi.org/10.1086/591861.

Holm, L., & Sander, C. (1995). DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. *Trends in Biochemical Sciences*, *20*(9), 345–347. https://doi.org/10.1016/s0968-0004(00)89071-4

Huang, J., Ma, J., Shang, K., Hu, X., Liang, Y., Li, D., Wu, Z., Dai, L., Chen, L., & Wang, L. (2016). Evolution and diversity of the antimicrobial resistance associated mobilome in

Streptococcus suis: A Probable mobile genetic elements reservoir for other Streptococci. *Frontiers in Cellular and Infection Microbiology*, *6*. https://doi.org/10.3389/fcimb.2016.00118.

Iannelli, F., Giunti, L., & Pozzi, G. (1998). Direct sequencing of long polymerase chain reaction fragments. *Molecular Biotechnology*, *10*(2), 183–185. https://doi.org/10.1007/BF02760864.

Islam, M. R., Kim, H., Kang, S.-W., Kim, J.-S., Jeong, Y.-M., Hwang, H.-J., Lee, S.-Y., Woo, J.-C., & Kim, S.-G. (2007). Functional characterization of a gene encoding a dual domain for uridine kinase and uracil phosphoribosyltransferase in Arabidopsis thaliana. *Plant Molecular Biology*, *63*(4), 465–477. https://doi.org/10.1007/s11103-006-9101-3.

Jackson, J., Chen, C., & Buising, K. (2013). Aminoglycosides: How should we use them in the 21st century? *Current Opinion in Infectious Diseases*, *26*(6), 516–525. https://doi.org/10.1097/QCO.0000000000000012.

Krause, K. M., Serio, A. W., Kane, T. R., & Connolly, L. E. (2016). Aminoglycosides: An Overview. *Cold Spring Harbor Perspectives in Medicine*, *6*(6), a027029. https://doi.org/10.1101/cshperspect.a027029.

Lam, M. M. C., Seemann, T., Bulach, D. M., Gladman, S. L., Chen, H., Haring, V., Moore, R. J., Ballard, S., Grayson, M. L., Johnson, P. D. R., Howden, B. P., & Stinear, T. P. (2012). Comparative analysis of the first complete *Enterococcus faecium* genome. *Journal of Bacteriology*, *194*(9), 2334–2341. https://doi.org/10.1128/JB.00259-12.

Lambowitz, A. M., & Zimmerly, S. (2011). Group II Introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology*, *3*(8), a003616–a003616. https://doi.org/10.1101/cshperspect.a003616.

Lebreton, F., Willems, R. J. L., & Gilmore, M. S. (2014). *Enterococcus* diversity, origins in nature, and gut colonization. In M. S. Gilmore, D. B. Clewell, Y. Ike, & N. Shankar (Eds.), *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection*. Massachusetts Eye and Ear Infirmary. http://www.ncbi.nlm.nih.gov/books/NBK190427/.

Lei, C.-W., Kang, Z.-Z., Wu, S.-K., Chen, Y.-P., Kong, L.-H., & Wang, H.-N. (2019). Detection of the phenicol–oxazolidinone–tetracycline resistance gene *poxtA* in *Enterococcus faecium* and *Enterococcus faecalis* of food-producing animal origin in China. *Journal of Antimicrobial Chemotherapy*, *74*(8), 2459–2461. https://doi.org/10.1093/jac/dkz198

Leinweber, H., Alotaibi, S. M. I., Overballe-Petersen, S., Hansen, F., Hasman, H., Bortolaia, V., Hammerum, A. M., & Ingmer, H. (2018). Vancomycin resistance in *Enterococcus faecium* isolated from Danish chicken meat is located on a pVEF4-like plasmid persisting in poultry for 18 years. *International Journal of Antimicrobial Agents*, *52*(2), 283–286. https://doi.org/10.1016/j.ijantimicag.2018.03.019

Lima, C. D., Wang, J. C., & Mondragón, A. (1994). Three-dimensional structure of the 67K N-terminal fragment of *E. coli* DNA topoisomerase I. *Nature*, *367*(6459), 138–146. https://doi.org/10.1038/367138a0.

Lin, Y.-T., Tseng, S.-P., Hung, W.-W., Chang, C.-C., Chen, Y.-H., Jao, Y.-T., Chen, Y.-H., Teng, L.-J., & Hung, W.-C. (2020). A possible role of Insertion Sequence IS*1216V* in dissemination of multidrug-resistant elements MES_{PM1} and $MES_{6272\text{-}2}$ between *Enterococcus* and ST59 *Staphylococcus aureus. Microorganisms*, *8*(12), 1905. https://doi.org/10.3390/microorganisms8121905.

Meinhart, A., & Alonso, J. C. (2003). *Crystal structure of the plasmid maintenance system ε/ζ: Functional Mechanism of Toxin ζ inactivation by ε2ζ2 complex formation. Proc Natl Acad Sci USA*. 2003 Feb 18; 100(4): 1661–1666. doi: 10.1073/pnas.0434325100.

Mercuro, N. J., Davis, S. L., Zervos, M. J., & Herc, E. S. (2018). Combatting resistant enterococcal infections: A pharmacotherapy review. *Expert Opinion on Pharmacotherapy*, *19*(9), 979–992. https://doi.org/10.1080/14656566.2018.1479397.

Morroni, G., Brenciani, A., Antonelli, A., Maria D'Andrea, M., Di Pilato, V., Fioriti, S., Mingoia, M., Vignaroli, C., Cirioni, O., Biavasco, F., Varaldo, P. E., Rossolini, G. M., & Giovanetti, E. (2018). Characterization of a Multiresistance Plasmid Carrying the *optrA* and *cfr* resistance genes from an *Enterococcus faecium* clinical isolate. *Frontiers in Microbiology*, *9*, 2189. https://doi.org/10.3389/fmicb.2018.02189.

Murayama, K., Orth, P., de la Hoz, A. B., Alonso, J. C., & Saenger, W. (2001). Crystal structure of ω transcriptional repressor encoded by *Streptococcus pyogenes* plasmid pSM19035 at 1.5 A resolution 1 1Edited by R. Huber. *Journal of Molecular Biology*, *314*(4), 789–796. https://doi.org/10.1006/jmbi.2001.5157.

Palmer, K. L., Kos, V. N., & Gilmore, M. S. (2010). Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. *Current Opinion in Microbiology*, *13*(5), 632–639. https://doi.org/10.1016/j.mib.2010.08.004.

Paulsen, I. T., Banerjei, L., Myers, G. S. A., Nelson, K. E., Seshadri, R., Read, T. D., Fouts, D. E., Eisen, J. A., Gill, S. R., Heidelberg, J. F., Tettelin, H., Dodson, R. J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R. T., Durkin, S., Kolonay, J., … Fraser, C. M. (2003). Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*, *299*(5615), 2071–2074. https://doi.org/10.1126/science.1080613.

Pöntinen, A. K., Top, J., Arredondo-Alonso, S., Tonkin-Hill, G., Freitas, A. R., Novais, C., Gladstone, R. A., Pesonen, M., Meneses, R., Pesonen, H., Lees, J. A., Jamrozy, D., Bentley, S. D., Lanza, V. F., Torres, C., Peixe, L., Coque, T. M., Parkhill, J., Schürch, A. C., … Corander, J. (2021). Apparent nosocomial adaptation of *Enterococcus faecalis* predates the modern hospital era. *Nature Communications*, *12*(1), 1523. https://doi.org/10.1038/s41467-021-21749-5.

Pratto, F., Cicek, A., Weihofen, W. A., Lurz, R., Saenger, W., & Alonso, J. C. (2008). *Streptococcus pyogenes* pSM19035 requires dynamic assembly of ATP-bound *ParA* and

ParB on *parS* DNA during plasmid segregation. *Nucleic Acids Research*, *36*(11), 3676–3689. https://doi.org/10.1093/nar/gkn170.

Qin, S., Wang, Y., Zhang, Q., Chen, X., Shen, Z., Deng, F., Wu, C., & Shen, J. (2012). Identification of a novel genomic island conferring resistance to multiple aminoglycoside antibiotics in *Campylobacter coli*. *Antimicrobial Agents and Chemotherapy*, 56(10), 5332–5339. https://doi.org/10.1128/AAC.00809-12

Qin, X., Galloway-Peña, J. R., Sillanpaa, J., Roh, J. H., Nallapareddy, S. R., Chowdhury, S., Bourgogne, A., Choudhury, T., Muzny, D. M., Buhay, C. J., Ding, Y., Dugan-Rocha, S., Liu, W., Kovar, C., Sodergren, E., Highlander, S., Petrosino, J. F., Worley, K. C., Gibbs, R. A., … Murray, B. E. (2012). Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes. *BMC Microbiology*, *12*(1), 135. https://doi.org/10.1186/1471-2180-12-135.

Ramirez, M. S., & Tolmasky, M. E. (2010). Aminoglycoside modifying enzymes. *Drug Resistance Updates: Reviews and Commentaries in Antimicrobial and Anticancer Chemotherapy*, *13*(6), 151–171. https://doi.org/10.1016/j.drup.2010.08.003.

Ricci, S., De Giorgi, S., Lazzeri, E., Luddi, A., Rossi, S., Piomboni, P., De Leo, V., & Pozzi, G. (2018). Impact of asymptomatic genital tract infections on in vitro Fertilization (IVF) outcome. *PLOS ONE*, *13*(11), e0207684. https://doi.org/10.1371/journal.pone.0207684.

Rouch, D. A., Byrne, M. E., Kong, Y. C., & Skurray, R. A. (1987). The *aacA-aphD* Gentamicin and Kanamycin Resistance Determinant of Tn*4001* from *Staphylococcus aureus*: Expression and Nucleotide Sequence Analysis. *Microbiology*, *133*(11), 3039–3052. https://doi.org/10.1099/00221287-133-11-3039.

Santoro, F., Oggioni, M. R., Pozzi, G., & Iannelli, F. (2010). Nucleotide sequence and functional analysis of the tet (M)-carrying conjugative transposon Tn*5251* of *Streptococcus pneumoniae*: Tn*5251* of *Streptococcus pneumoniae*. *FEMS Microbiology Letters*, no-no. https://doi.org/10.1111/j.1574-6968.2010.02002.

Schwarz, F. V., Perreten, V., & Teuber, M. (2001). Sequence of the 50-kb conjugative multiresistance plasmid p*RE25* from *Enterococcus faecalis* RE25. *Plasmid*, *46*(3), 170–187. https://doi.org/10.1006/plas.2001.1544.

Shaw, J. H., & Clewell, D. B. (1985). Complete nucleotide sequence of macrolide-lincosamide-streptogramin B-resistance transposon Tn*917* in *Streptococcus faecalis*. *Journal of Bacteriology*, *164*(2), 782–796. https://doi.org/10.1128/jb.164.2.782-796.1985.

Singleton, M. R., Scaife, S., & Wigley, D. B. (2001). Structural analysis of DNA replication fork reversal by RecG. *Cell*, *107*(1), 79–89. https://doi.org/10.1016/S0092-8674(01)00501-3

Sletvold, H., Johnsen, P. J., Simonsen, G. S., Aasnæs, B., Sundsfjord, A., & Nielsen, K. M. (2007). Comparative DNA analysis of two *vanA* plasmids from *Enterococcus faecium* strains isolated from poultry and a poultry farmer in Norway. *Antimicrobial Agents and Chemotherapy*, *51*(2), 736–739. https://doi.org/10.1128/AAC.00557-06

Sletvold, H., Johnsen, P. J., Wikmark, O.-G., Simonsen, G. S., Sundsfjord, A., & Nielsen, K. M. (2010). Tn*1546* is part of a larger plasmid-encoded genetic unit horizontally disseminated among clonal *Enterococcus faecium* lineages. *Journal of Antimicrobial Chemotherapy*, *65*(9), 1894–1906. https://doi.org/10.1093/jac/dkq219.

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., & Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, *44*(14), 6614–6624. https://doi.org/10.1093/nar/gkw569.

Tendolkar, P. M., Baghdayan, A. S., & Shankar, N. (2006). Putative surface proteins encoded within a novel transferable locus confer a high-biofilm phenotype to *Enterococcus faecalis*. *Journal of Bacteriology*, *188*(6), 2063–2072. https://doi.org/10.1128/JB.188.6.2063-2072.2006.

Tyson, G. H., Sabo, J. L., Hoffmann, M., Hsu, C.-H., Mukherjee, S., Hernandez, J., Tillman, G., Wasilenko, J. L., Haro, J., Simmons, M., Wilson Egbe, W., White, P. L., Dessai, U., &

Mcdermott, P. F. (2018). Novel linezolid resistance plasmids in *Enterococcus* from food animals in the USA. *Journal of Antimicrobial Chemotherapy*. https://doi.org/10.1093/jac/dky369.

Weaver KE. (2019). Enterococcal genetics. *Microbiol Spectr,* 7(2). https://doi: 10.1128/microbiolspec.GPP3-0055-2018. PMID: 30848235.

Werner, G., Hildebrandt, B., & Witte, W. (2001). Aminoglycoside-streptothricin resistance gene cluster *aadE–sat4–aphA-3* disseminated among multiresistant isolates of *Enterococcus faecium*. *Antimicrobial Agents and Chemotherapy*, *45*(11), 3267–3269. https://doi.org/10.1128/AAC.45.11.3267-3269.2001

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology*, *13*(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Yang, W., & Steitz, T. A. (1995). Crystal Structure of the Site-Specific Recombinase Resolvase Complexed with a 34 bp Cleavage Site. *Cell*, 28;82(2):193-207. https://doi.org/10.1016/0092-8674(95)90307-0. PMID: 7628011.

Zhao, F., Tong, Q., Fu, Y., Ruan, Z., Shi, K., Ji, J., Yu, Y., & Xie, X. (2020). Molecular characteristics of PaLoc and acquired antimicrobial resistance in epidemic *Clostridioides difficile* isolates revealed by whole-genome sequencing. *Journal of Global Antimicrobial Resistance*, *23*, 194–196. https://doi.org/10.1016/j.jgar.2020.09.016.

Zheng, R., & Blanchard, J. S. (2000). Kinetic and Mechanistic Analysis of the *E. coli panE* - Encoded Ketopantoate Reductase. *Biochemistry*, *39*(13), 3708–3717. https://doi.org/10.1021/bi992676g.

Zhu, W., Murray, P. R., Huskins, W. C., Jernigan, J. A., McDonald, L. C., Clark, N. C., Anderson, K. F., McDougal, L. K., Hageman, J. C., Olsen-Rasmussen, M., Frace, M., Alangaden, G. J., Chenoweth, C., Zervos, M. J., Robinson-Dunn, B., Schreckenberger, P. C., Reller, L. B., Rudrik, J. T., & Patel, J. B. (2010). Dissemination of an *Enterococcus* Inc18-Like *vanA*

Plasmid Associated with Vancomycin-Resistant *Staphylococcus aureus*. *Antimicrobial Agents and Chemotherapy*, *54*(10), 4314–4320. https://doi.org/10.1128/AAC.00185-10.

Zúñiga, M., Pardo, I., & Ferrer, S. (2003). Conjugative plasmid pIP501 undergoes specific deletions after transfer from *Lactococcus lactis* to *Oenococcus oeni*. *Archives of Microbiology*, *180*(5), 367–373. https://doi.org/10.1007/s00203-003-0599-3.

FIGURE LEGEND

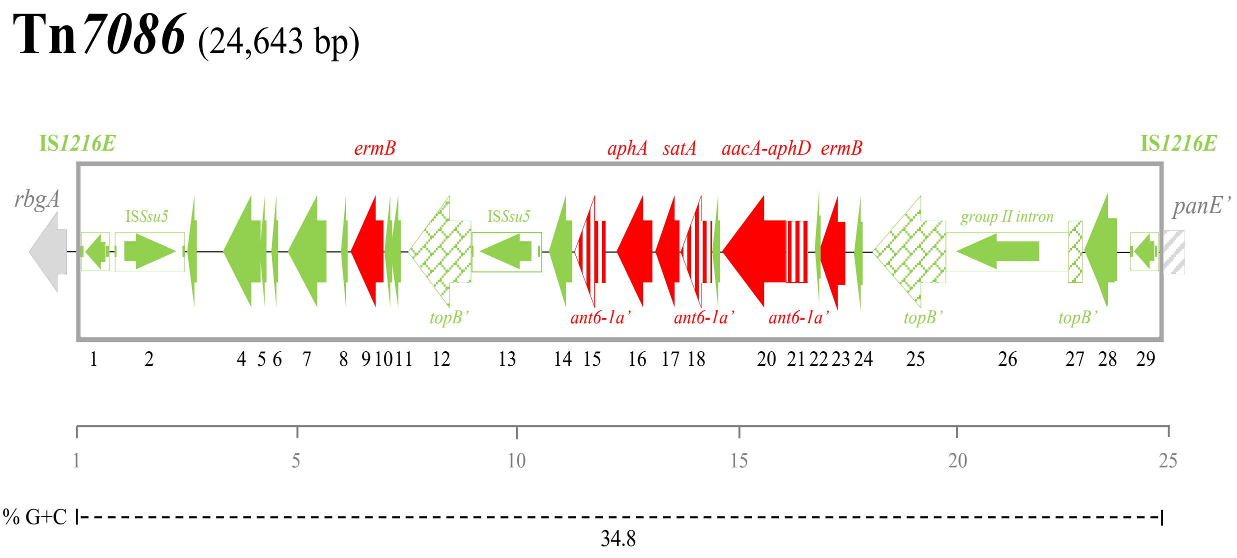


***Figure 1.* Structure of Tn*7086* of *E. faecalis* strain 4638.** Tn*7086* was found integrated 142 bp upstream the 5' end of the *rbgA* gene, into the *panE* gene. Tn*7086* is 24,643 bp-long and contains 29 ORFs. ORFs and their direction of transcription are represented by arrows, while annotated ORFs are indicated by sequential numbers. ISs are reported as boxed arrows and inverted repeat within ISs are indicated by solid rectangles, whereas group II intron is represented only as a boxed arrow. All genes are depicted in green except for antimicrobial resistance genes which are highlighted in red. Chromosomal genes flanking the Tn*7086* insertion site are represented as grey arrows. Arrows and boxes with a pattern fill indicate truncated genes. The GC content of the element is indicated by dotted bars. Scale, kilobases.

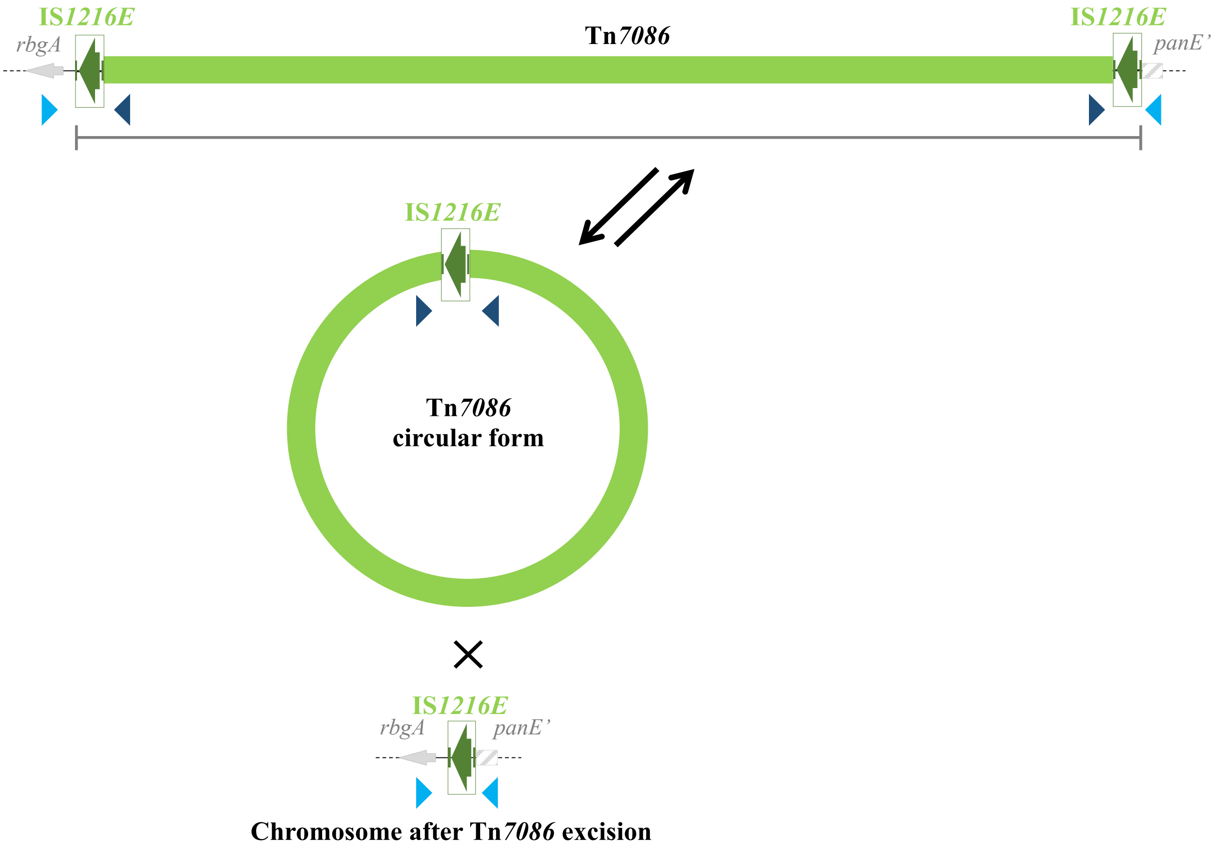


Figure 2. **Mechanism of excision/integration of Tn*7086* in *E. faecalis* strain 4638.** Tn*7086* is able to excise from the bacterial chromosome producing circular forms. Excision of Tn*7086* occurs by recombination between the flanking IS*1216E* copies (boxed dark green arrows) and leaves a single IS*1216E* copy in the chromosome, whereas the ends of the transposon in the circular forms are joined by the other IS*1216E* copy. Arrowheads represent PCR primers used for detection of circular forms (dark blue) and reconstituted chromosomal site of integration (light blue). The structure of Tn*7086* is described in Figure 1.

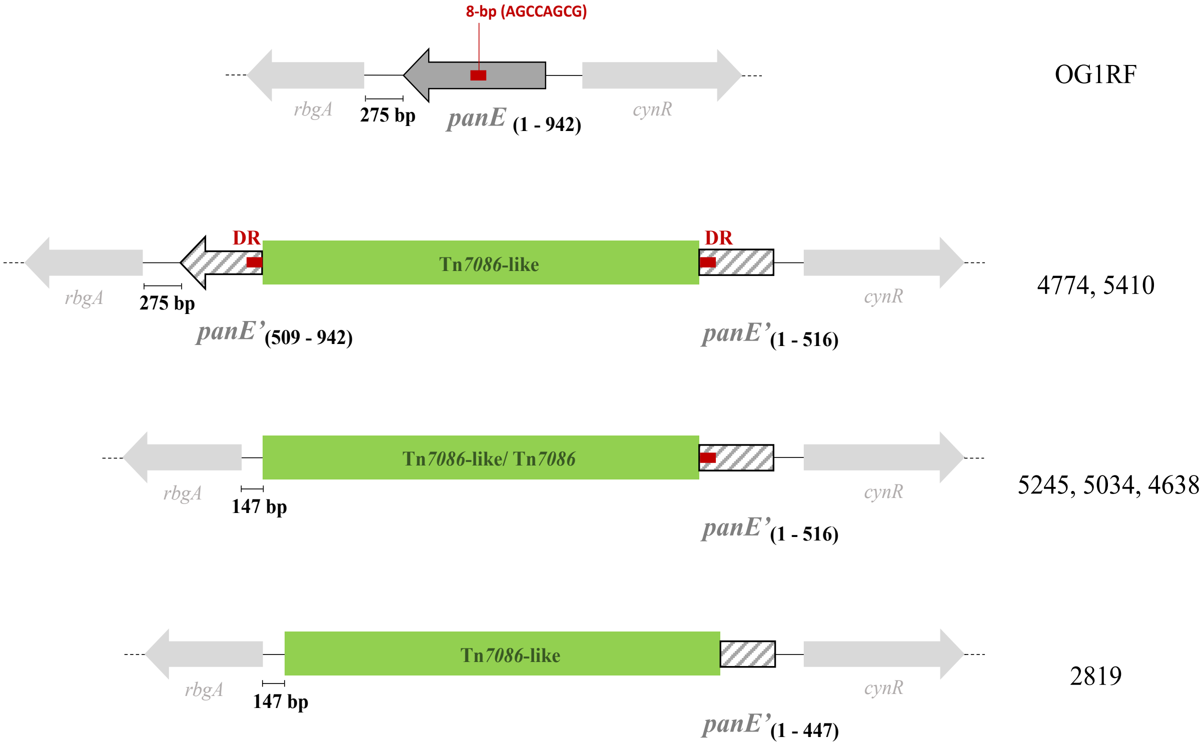


Figure 3. **Chromosomal integration sites of Tn*7086*-like composite transposons in infertility-associated *E. faecalis* CC16/ST480 strains.** All Tn*7086*-like elements (green blocks) were found all integrated in the chromosomal gene *panE* localized upstream of the *rbgA* and downstream of the *cynR* genes (grey arrows). Upon integration: (i) a duplication of an 8-bp target site sequence (AGCCAGCG, red square) of *panE* occurred in strains 4774 and 5410; (ii) a 555-bp deletion involving the 3'-end of *panE* and 128 bp of the downstream intergenic region occurred in strains 5034 and 5245 as described for Tn*7086* in strain 4638; (iii) a 624-bp deletion involving a larger portion of the 3'-end of *panE* inclusive of the 8-bp target site and the same 128 nucleotides downstream *panE* was found in strain 2819. Disrupted *panE* genes (*panE*') are depicted as striped arrows/ boxes. The strains are compared to OG1RF reference strain which displays an undisrupted *panE* gene. The figure is not in scale.

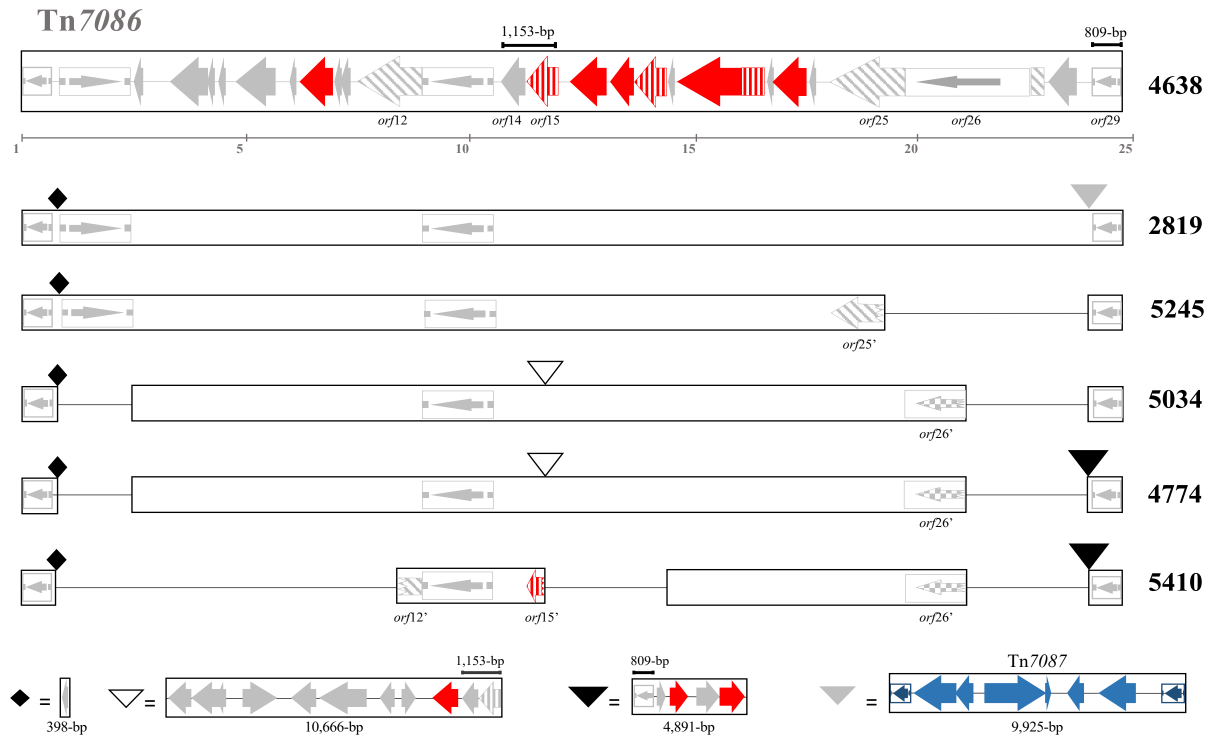
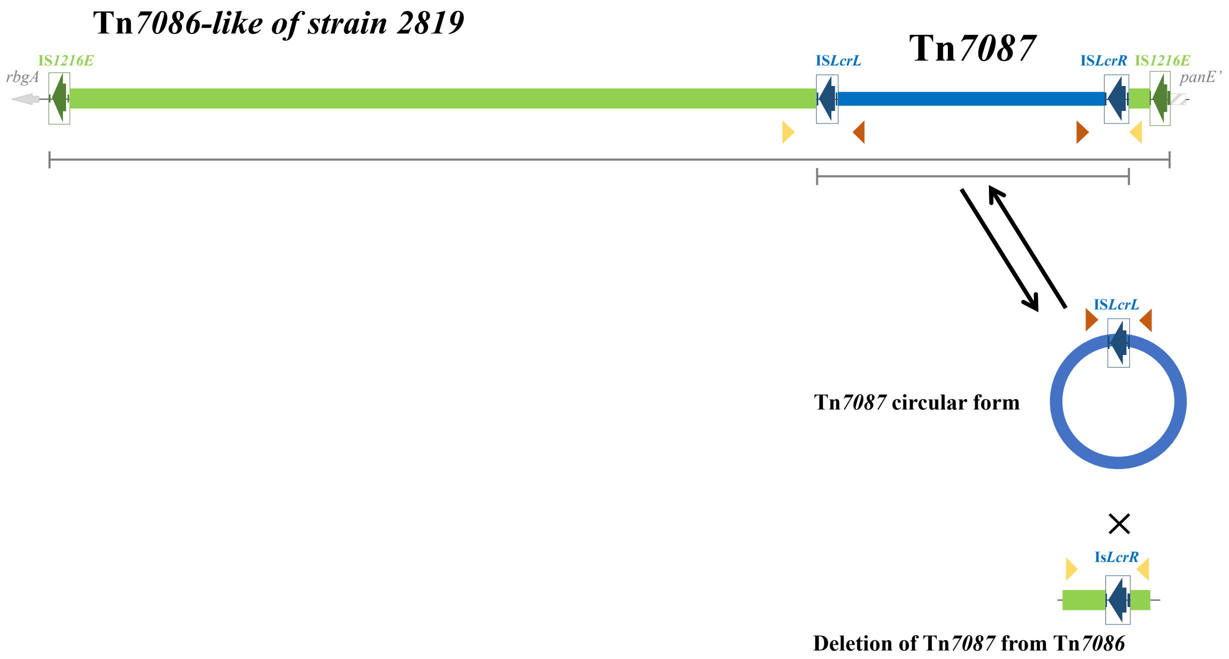


***Figure 4.* Structure of Tn*7086*-like composite transposons in infertility-associated *E. faecalis* CC16/ST480 strains.** The 24.6-kb-long Tn*7086* DNA sequence of *E. faecalis* strain 4638 (query) is compared to the homologous DNA sequences of Tn*7086*-like transposons found in the other infertility-associated strains (on the right). Tn*7086*-like transposons are depicted in a sequential order based on decreasing sequence homology percentage compared to Tn*7086*. All transposons are flanked by two identical IS*1216E* copies, whereas their DNA sequences vary in length from 15.4 kb (strain 5410) up to 35.3 kb (strain 2819). Homologous sequences are drawn as white blocks (scale in kb). ISs are reported as boxed arrows and their IR indicated by solid rectangles, whereas serrated edge arrows and serrated box with pattern fill indicate truncated *orf* in comparison with those found in Tn*7086*. Red arrows highlight antimicrobial resistance genes. For clearer alignment, homologous sequence blocks were represented as devoid of additional genetic elements, indicated at the bottom of the figure as a solid black diamond (398 bp-long) and as solid black (4.8 kb-long insert), solid grey (9.9 kb-long insert) and empty (10.6 kb-long insert) triangles. The 10.6-kb and the 4.8-kb inserted DNA sequences produced a 1,153-bp duplication of *orf*14-16 and an 809-bp duplication of *orf*29, respectively (indicated on the Tn*7086* DNA sequence of strain 4638). The 9.9-kb insert of strain 2819 is an independent composite transposon named Tn*7087*.



***Supplementary Figure S1.* Mechanism of excision/integration of Tn*7087* from Tn*7086* in *E. faecalis* strain 2819.** Tn*7087* is a 9,925 bp-long element integrated into the Tn*7086*-like composite transposon of *E. faecalis* strain 2819. Tn*7087* excises from Tn*7086* and produces circular forms. Excision of Tn*7087* occurs by recombination between the flanking IS*Lcr* copies (boxed dark blue arrows), leaving IS*Lcr*L in the Tn*7086*, whereas the ends of Tn*7087* in the circular forms are joined by IS*Lcr*R copy. Arrowheads represent PCR primers used for detection of circular forms (brown) and reconstituted chromosomal site of integration (yellow).

CHAPTER 6

Complete genome sequence of *Lactobacillus crispatus* type strain ATCC 33820

Lucia Teodori*, Lorenzo Colombini*, Anna Maria Cuppone, Elisa Lazzeri, David Pinzauti, Francesco Santoro, Francesco Iannelli, Gianni Pozzi

Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy

*Lucia Teodori and Lorenzo Colombini contributed equally to this work. Author order was determined by flipping a coin.

Microbiology Resource Announcement 2021

AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology® Resource Announcements

GENOME SEQUENCES

Complete Genome Sequence of *Lactobacillus crispatus* Type Strain ATCC 33820

Lucia Teodori,[a] Lorenzo Colombini,[a] Anna Maria Cuppone,[a] Elisa Lazzeri,[a] David Pinzauti,[a] Francesco Santoro,[a] Francesco Iannelli,[a] Gianni Pozzi[a]

[a]Department of Medical Biotechnologies, University of Siena, Siena, Italy

Lucia Teodori and Lorenzo Colombini contributed equally to this work. Author order was determined by flipping a coin.

ABSTRACT The complete genome sequence of *Lactobacillus crispatus* type strain ATCC 33820 was obtained by combining Nanopore and Illumina sequencing technologies. The genome consists of a 2.2-Mb circular chromosome with 2,194 open reading frames and an average GC content of 37.0%.

Lactobacillus crispatus is the most frequently isolated species among the vaginal lactobacilli of the human microbiota of healthy women; its presence is associated with reduced risk of preterm delivery, viral sexually transmitted infections, and bacterial vaginosis (1). To date (June 2021), only eight *L. crispatus* complete genomes are available in the NCBI database (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/1815/). Here, we contribute to the genomic characterization of this species by publicly releasing the genome of strain ATCC 33820, the type strain of *Lactobacillus crispatus* (Fig. 1). The strain was purchased from the American Type Culture Collection in October 2020 and grown in 250 ml of DeMan-Rogosa-Sharpe (MRS) broth at 37°C to an optical density at 590 nm (OD_{590}) of 1.9. Bacterial cells were harvested by centrifugation (5,000 × *g* for 30 min at 4°C), and the cell pellet was dry-vortexed and incubated for 1 h at 37°C in protoplasting buffer (20% raffinose, 50 mM Tris-HCl [pH 8.0], 5 mM EDTA) containing 4 mg/ml lysozyme. Protoplasts were centrifuged (5,000 × *g* for 5 min), resuspended in 15 ml of deionized H_2O with 100 μg/ml proteinase K (Merck KGaA, Darmstadt, Germany), and incubated for 30 min at 37°C to obtain osmotic lysis, with 0.5% SDS added after 15 min. Then, 0.55 M NaCl was added, and the mixture was incubated for 10 min at room temperature. High-molecular-weight DNA was purified by three extractions with 1 volume of Sevag (chloroform-isoamyl alcohol, 24:1 [vol:vol]), precipitated in 0.6 volume of cold isopropanol, and spooled on a glass rod. DNA was resuspended in saline-sodium citrate (SSC)/10 buffer and then adjusted to 600 μl SSC 1×. The DNA solution was homogenized using a rotator mixer and stored at 4°C. DNA sequencing was performed with both Oxford Nanopore GridION and Illumina NovaSeq 6000 instruments. The Nanopore sequencing library was prepared using the Nanopore sequencing kit SQK-LSK 109 (Oxford Nanopore Technologies, Oxford, UK), and the sample was sequenced using an R9.4 flow cell (FLO-MIN106). Real-time high-accuracy base calling (quality cutoff, >Q7) of Nanopore reads was performed using Guppy v4.0.11 (https://github.com/nanoporetech/pyguppyclient), and base-called reads were analyzed with NanoPlot v1.18.2 (2). Illumina sequencing was performed at MicrobesNG (University of Birmingham, UK) using a Nextera XT library preparation kit (Illumina Inc., San Diego, CA, USA), followed by paired-end sequencing. Illumina reads were trimmed using Trimmomatic v0.30 (3) and analyzed with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Nanopore and Illumina sequencing generated 136,000 long reads (630,559,194 bp; N_{50}, 8.7 kb) and 762,936 read pairs (2 × 250 bp), respectively. Nanopore reads were filtered using Filtlong v0.2.0 with the parameter --target_bases to retain a total of 230 Mbp (https://github.com/rrwick/Filtlong) (N_{50}, 19,822 bp) and assembled using Unicycler v0.4.7 (4). The resulting circular contig was polished using

Citation Teodori L, Colombini L, Cuppone AM, Lazzeri E, Pinzauti D, Santoro F, Iannelli F, Pozzi G. 2021. Complete genome sequence of *Lactobacillus crispatus* type strain ATCC 33820. Microbiol Resour Announc 10:e00634-21. https://doi.org/10.1128/MRA.00634-21.

Editor David A. Baltrus, University of Arizona

Copyright © 2021 Teodori et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Francesco Santoro, santorof@unisi.it.

Received 25 June 2021
Accepted 22 July 2021
Published 12 August 2021

Downloaded from https://journals.asm.org/journal/mra on 06 September 2021 by 5.171.89.151.

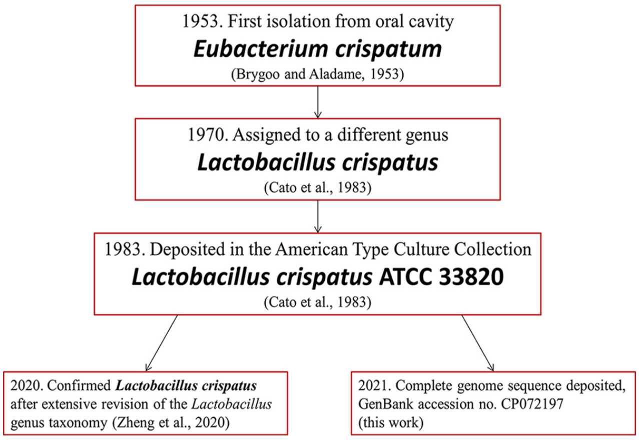

1953. First isolation from oral cavity
Eubacterium crispatum
(Brygoo and Aladame, 1953)
1970. Assigned to a different genus
Lactobacillus crispatus
(Cato et al., 1983)
1983. Deposited in the American Type Culture Collection
Lactobacillus crispatus ATCC 33820
(Cato et al., 1983)
2020. Confirmed Lactobacillus crispatus after extensive revision of the Lactobacillus genus taxonomy (Zheng et al., 2020)
2021. Complete genome sequence deposited, GenBank accession no. CP072197 (this work)

FIG 1 History of *Lactobacillus crispatus* type strain ATCC 33820. *L. crispatus* type strain ATCC 33820 was isolated at the Institut Pasteur in 1953 by E. R. Brygoo and N. Aladame from an oral sample of a European individual in Saigon and was considered a new species of the genus *Eubacterium* (Collection of the Institut Pasteur, Paris, strain II) (7). Later, it was deposited in the Virginia Polytechnic Institute and State University as VPI 3199 and identified as *Lactobacillus* (8). Further characterization upon American Type Culture Collection deposition indicated that ATCC 33820 DNA was 100% homologous to the previously defined *L. acidophilus* group A2 (8). Over the years, the *L. crispatus* type strain has been distributed among different collections and also designated DSM 20584 = CCUG 30722 = CIP 102990 = CIPP II = JCM 1185 = LMG 9479. Recently, Zheng and colleagues (9) reclassified the genus *Lactobacillus* into 25 genera through a polyphasic approach; however, the nomenclature of *Lactobacillus crispatus* remained unchanged. Strain ATCC 33820 was acquired by our laboratory in October 2020. Arrows indicate sequential steps in the history of the *L. crispatus* type strain. Red boxes contain the year, followed by a brief description of the event, the strain name (in bold), and the reference (in parentheses).

Medaka v0.7.1 (https://github.com/nanoporetech/medaka) with all Nanopore reads, followed by two polishing rounds with Pilon v1.22 using the Illumina reads (5). Assembly quality was evaluated using Ideel (https://github.com/mw55309/ideel). Annotation was performed with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v5.1 (6). Default parameters were used for all software unless otherwise specified. The genome of *L. crispatus* ATCC 33820 consists of a single circular chromosome (2,239,089 bp) with an overall GC content of 37.0%. The assembly contains 2,194 open reading frames, 78.8% with putative biological function, 64 tRNA genes, 3 rRNA operons, and 3 structural RNAs.

Data availability. Sample information and sequence and genomic assembly/annotation are accessible under the NCBI BioProject, BioSample, and whole-genome sequence accession numbers PRJNA716945, SAMN18472633, and CP072197, respectively. Raw Nanopore and Illumina sequencing reads are accessible under Sequence Read Archive accession numbers SRR14509463 and SRR14509462, respectively.

ACKNOWLEDGMENTS

This work was funded by the Italian Ministry of Education, University and Research (MIUR-Italy), under grant number 20177J5Y3P ("Progetti di Ricerca di Rilevante Interesse Nazionale–Bando 2017"). The GridION x-5 was acquired through the "Excellence Departments" grant from MIUR to the Department of Medical Biotechnologies, University of Siena.

Illumina genome sequencing was provided by MicrobesNG.

REFERENCES

1. Petrova MI, Lievens E, Malik S, Imholz N, Lebeer S. 2015. Lactobacillus species as biomarkers and agents that can promote various aspects of vaginal health. Front Physiol 6:81. https://doi.org/10.3389/fphys.2015.00081.
2. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
3. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
4. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

Downloaded from https://journals.asm.org/journal/mra on 06 September 2021 by 5.171.89.151.

5. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.
6. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
7. Brygoo ER, Aladame N. 1953. Study of a new strictly anaerobic species of the genus Eubacterium: Eubacterium crispatum n. sp. Ann Inst Pasteur (Paris) 84:640–641.
8. Cato EP, Moore WEC, Johnson JL. 1983. Synonymy of strains of "Lactobacillus acidophilus" group A2 (Johnson et al. 1980) with the type strain of Lactobacillus crispatus (Brygoo and Aladame 1953) Moore and Holdeman 1970. Int J Syst Evol Microbiol 33:426–428.
9. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, O'Toole PW, Pot B, Vandamme P, Walter J, Watanabe K, Wuyts S, Felis GE, Gänzle MG, Lebeer S. 2020. A taxonomic note on the genus Lactobacillus: description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae. Int J Syst Evol Microbiol 70:2782–2858. https://doi.org/10.1099/ijsem.0.004107.

Downloaded from https://journals.asm.org/journal/mra on 06 September 2021 by 5.171.89.151.

CHAPTER 7

Complete genome sequence of *Streptococcus pneumoniae* strain Rx1, a Hex mismatch repair-deficient standard transformation recipient

Anna Maria Cuppone*, Lorenzo Colombini*, Valeria Fox, David Pinzauti,
Francesco Santoro, Gianni Pozzi, Francesco Iannelli

Laboratory of Molecular Microbiology and Biotechnology, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy

*Anna Maria Cuppone and Lorenzo Colombini contributed equally to this work. The order of names was determined according to contributions to the research project of which this work is part.

Microbiology Resource Announcement 2021

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Complete Genome Sequence of *Streptococcus pneumoniae* Strain Rx1, a Hex Mismatch Repair-Deficient Standard Transformation Recipient

Anna Maria Cuppone,[a] Lorenzo Colombini,[a] Valeria Fox,[a] David Pinzauti,[a] Francesco Santoro,[a] Gianni Pozzi,[a] Francesco Iannelli[a]

[a]Department of Medical Biotechnologies, University of Siena, Siena, Italy

Anna Maria Cuppone and Lorenzo Colombini contributed equally to this work. The order of names was determined according to contributions to the research project of which this work is part.

ABSTRACT The complete genome sequence of *Streptococcus pneumoniae* strain Rx1, a Hex mismatch repair-deficient standard transformation recipient, was obtained by combining Nanopore and Illumina sequencing technologies. The genome consists of a 2.03-Mb circular chromosome, with 2,054 open reading frames and a GC content of 39.72%.

Streptococcus pneumoniae is a human pathogen and the most important model organism for studying bacterial genetics and genomics. Widely used laboratory strains include type 2 Avery's strain D39 and its derivatives Rx1 and R6, which are standard transformation recipients (1, 2). We characterized the complete genome sequence of Rx1, a highly transformable and Hex mismatch repair system-deficient strain. To track the genomic changes that gave rise to Rx1, we also sequenced the genome of its unencapsulated parental strain R36A (Table 1). Strains, which were obtained from the Guild laboratory collection (3), were grown in tryptic soy broth at 37°C for 4 h until they reached an optical density at 590 nm (OD_{590}) of 0.8. Pneumococcal cells were harvested by centrifugation (5,000 × *g* for 30 min at 4°C), and the cell pellet was dry vortex-mixed and lysed in 0.1% deoxycholate-0.008% SDS. High-molecular-weight DNA was purified three times with 1 volume of chloroform-isoamyl alcohol (24:1 [vol/vol]), precipitated in 0.6 volumes of ice-cold isopropanol, and spooled on a glass rod. DNA was resuspended in 10× saline-sodium citrate (SSC) buffer (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) and then adjusted to 1× SSC and maintained at 4°C. The DNA solution was homogenized using a rotary mixer. Oxford Nanopore Technologies MinION and Illumina HiSeq 2500 instruments were used for DNA sequencing. DNA was not sheared; size selection was obtained with 0.8 volumes of AMPure XP beads (Beckman Coulter). The Nanopore sequencing library was prepared using the SQK-LSK108 kit (Oxford Nanopore Technologies) following the manufacturer's instructions, and the sample was sequenced using an R9.4 flow cell (FLO-MIN106). Postsequencing high-accuracy base calling and adapter trimming of raw Nanopore reads were performed using Guppy v4.0.11 with configuration dna_r9.4.1_450bps_hac, and base-called reads were analyzed with NanoPlot v1.18.2 (4). Illumina sequencing was performed at MicrobesNG (University of Birmingham) using the Nextera XT library preparation kit (Illumina Inc.), followed by paired-end sequencing. Illumina reads were trimmed using Trimmomatic v0.30 (5) and analyzed with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Nanopore and Illumina sequencing generated 3,892 long reads (26,780,859 bp [N_{50}, 18.3 kbp]) and 86,582 read pairs (2 × 250 bp), respectively, for Rx1, whereas 4,771 long reads (27,433,219 bp [N_{50}, 16.9 kbp]) and 278,462 read pairs were obtained for R36A. Sequence coverage was 31.6× for Rx1 and 67.0× for R36A. A hybrid assembly of Nanopore and Illumina reads was obtained using Unicycler v0.4.712 (6). Assembly completeness and quality were assessed using Bandage v.0.8.1 (7) and Ideel (https://github.com/mw55309/ideel), respectively.

Citation Cuppone AM, Colombini L, Fox V, Pinzauti D, Santoro F, Pozzi G, Iannelli F. 2021. Complete genome sequence of *Streptococcus pneumoniae* strain Rx1, a Hex mismatch repair-deficient standard transformation recipient. Microbiol Resour Announc 10:e00799-21. https://doi.org/10.1128/MRA.00799-21.

Editor Steven R. Gill, University of Rochester School of Medicine and Dentistry

Copyright © 2021 Cuppone et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Francesco Iannelli, francesco.iannelli@unisi.it.

Received 6 August 2021
Accepted 11 September 2021
Published 14 October 2021

Downloaded from https://journals.asm.org/journal/mra on 14 October 2021 by 82.60.166.172.

TABLE 1 Genealogy of the *S. pneumoniae* Rx1 strain

Strain	Description[a]	Relevant properties[b]	GenBank accession no. (year)[a]
D39	Avery's strain, clinical isolate (1916); type 2, virulent (3, 19–23)	pDP1$^+$, Hex$^+$, DpnI$^+$, *comC1-comD1*, *pspC3.1*	CP000410.1 (2007) (24)
R36	D39 passaged 36 times in anti-type 2 serum (1944); rough, avirulent (3, 21, 22)	pDP1$^+$, Hex$^+$, DpnI$^+$, *comC1-comD1*, *pspC3.1*	Not available
R36A	Highly transformable R36 colony morphology variant (1944); rough, avirulent (3, 20, 23, 25)	pDP1$^-$, Hex$^+$, DpnI$^+$, *comC1-comD1*, *pspC3.1*	CP079922 (2021) (this study)
R6	Highly transformable R36A single-colony isolate (1962); rough, avirulent (3, 26, 27)	pDP1$^-$, Hex$^+$, DpnI$^+$, *comC1-comD1*, *pspC3.1*	AE007317.1 (2001) (16)
A66	Avery's strain, clinical isolate (1949); type 3, virulent (23, 25)	Hex$^+$, DpnI, *comC2-comD2*, *pspC11.4*	LN847353.1 (2015) (28)
SIII-N	R36A transformed with A66 DNA (1949); type 3, virulent (20, 23, 25, 29)	*comC1-comD1*, *pspC3.1*	Not available
Rx	Spontaneous rough derivative of R36A (1959); reduced type 3 capsule production, avirulent (3, 17, 23, 30)	pDP1$^-$, Hex$^-$ (HexB$^-$), *comC1-comD1*, *pspC3.1*	Not available
Rx1	Highly transformable derivative of Rx (1959); reduced type 3 capsule production (Ugd mutant), avirulent (3, 31)	pDP1$^-$, Hex$^-$ (HexB$^-$), DpnI$^-$ (DpnC$^-$), *comC1-comD1*, *pspC3.1'*	CP079923 (2021) (this study)

[a] The year of the first strain description (except for the D39 isolation year) or of the sequence release is reported in parentheses.

[b] pDP1 is a 3,161-bp cryptic plasmid (32). Hex is the DNA mismatch repair system encoded by *hexA* and *hexB* (33). DpnI is a restriction system composed of the DpnI/DpnC endonuclease and DpnD (34). *comC-comD* competence genes encode the competence-stimulating peptide (CSP) and its ComD receptor (35–38). *pspC* encodes the virulence surface protein PspC (39, 40).

Annotation was performed with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v5.1 (8). Default parameters were used for all tools unless otherwise specified. The Rx1 genome consists of a 2,030,186-bp single circular chromosome containing 2,054 open reading frames (ORFs), of which 1,813 have a predicted function. The 2,039,955-bp circular chromosome of R36A contains 2,059 ORFs, of which 1,834 have a putative function. Both

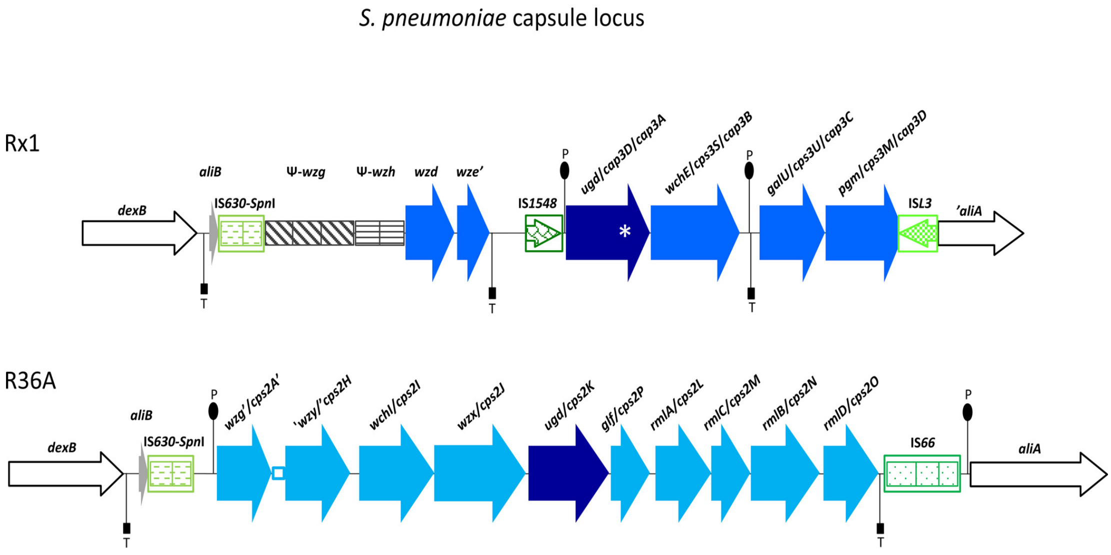


FIG 1 *S. pneumoniae* capsule locus. Rx1 harbors a type 3 capsule locus acquired by A66 DNA through a double crossover between IS*630-SpnI* and *aliA*. At the 3′ end, recombination produced the insertion of an IS*L3* transposase and a 950-bp deletion of the *aliA* 5′ end, as in the A66 capsule locus. IS*1548* identifies (i) a 5′ fragment, common to all serotypes (14), that contains *wzg* and *wzh* pseudogenes and *wzd* and *wze* genes and is not involved in type 3 capsular synthesis (15) and (ii) a 3′ fragment containing *ugd/cap3D/cap3A* UDP-glucose dehydrogenase gene, *wchE/cps3S/cap3B* synthase gene, *galU/cps3U/cap3C*, and *pgm/cps3M/cap3D* genes involved in UDP-glucose biosynthesis (15–17). The nucleotide change g.317,495C>T in *ugd/cps3A/cps3D* (indicated with an asterisk) causes p.R320C in the UDP-glucose dehydrogenase UDP-binding domain. The type 2 capsule locus of R36A harbors a 7,505-bp deletion involving the 3′ end of *wzg/cps2A*, seven genes (namely, *wzh/cps2B*, *wzd/cps2C*, *wze/cps2D*, *wchA/cps2E*, *wchF/cps2T*, *wchG/cps2F*, and *wchH/cps2G*), and the 5′ end of *wzy/cps2H* (18). The deletion event left an inverted 25-bp fragment (indicated with an open box) belonging to the lost *wzg/cps2A* 3′ end.

Downloaded from https://journals.asm.org/journal/mra on 14 October 2021 by 82.60.166.172.

genomes have (i) a GC content of 39.72%, (ii) 58 tRNA genes, 3 rRNA operons, and 3 structural RNAs, (iii) a 36.6-kb pneumococcal pathogenicity island 1 (PPI1) (9), (iv) prophage remnants, and (v) remnants of the integrative and conjugative element Tn*5253* (10–12). Rx1 and R36A capsule loci are schematized in Fig. 1. Rx1 harbors type I restriction-modification system SpnD39III variant C, while R36A harbors variant D (13). In Rx1, g.168,614C>A, g.1,979,527G>A, and g. 1,629,603delA nucleotide changes introduce premature termination codons in *hexB*, *pspc3.1*, and *dpnC*, respectively.

Data availability. The complete genome sequences of R36A and Rx1 are available under GenBank accession no. CP079922 and CP079923, respectively. The sequencing project is available under NCBI BioProject accession no. PRJNA748391. Nanopore and Illumina sequencing reads are available under Sequence Read Archive (SRA) accession no. SRR15216323 and SRR15216322, respectively, for R36A and SRA accession no. SRR15216380 and SRR15216379, respectively, for Rx1.

ACKNOWLEDGMENTS

This work was supported by the Italian Ministry of Education, University, and Research (MIUR-Italy), under grant 20177J5Y3P (Progetti di Ricerca di Rilevante Interesse Nazionale, Bando 2017). Illumina genome sequencing was provided by MicrobesNG (http://www.microbesng.uk).

REFERENCES

1. Pearce BJ, Iannelli F, Pozzi G. 2002. Construction of new unencapsulated (rough) strains of *Streptococcus pneumoniae*. Res Microbiol 153:243–247. https://doi.org/10.1016/s0923-2508(02)01312-8.
2. Santoro F, Iannelli F, Pozzi G. 2019. Genomics and genetics of *Streptococcus pneumoniae*. Microbiol Spectr 7:GPP3-0025-2018. https://doi.org/10.1128/microbiolspec.GPP3-0025-2018.
3. Smith MD, Guild WR. 1979. A plasmid in *Streptococcus pneumoniae*. J Bacteriol 137:735–739. https://doi.org/10.1128/jb.137.2.735-739.1979.
4. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
5. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
6. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
7. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics 31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383.
8. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.
9. Brown JS, Gilliland SM, Spratt BG, Holden DW. 2004. A locus contained within a variable region of pneumococcal pathogenicity island 1 contributes to virulence in mice. Infect Immun 72:1587–1593. https://doi.org/10.1128/IAI.72.3.1587-1593.2004.
10. Santoro F, Oggioni MR, Pozzi G, Iannelli F. 2010. Nucleotide sequence and functional analysis of the *tet*(M)-carrying conjugative transposon Tn*5251* of *Streptococcus pneumoniae*. FEMS Microbiol Lett 308:150–158. https://doi.org/10.1111/j.1574-6968.2010.02002.x.
11. Iannelli F, Santoro F, Oggioni MR, Pozzi G. 2014. Nucleotide sequence analysis of integrative conjugative element Tn*5253* of *Streptococcus pneumoniae*. Antimicrob Agents Chemother 58:1235–1239. https://doi.org/10.1128/AAC.01764-13.
12. Santoro F, Romeo A, Pozzi G, Iannelli F. 2018. Excision and circularization of integrative conjugative element Tn*5253* of *Streptococcus pneumoniae*. Front Microbiol 9:1779. https://doi.org/10.3389/fmicb.2018.01779.
13. Manso AS, Chai MH, Atack JM, Furi L, De Ste Croix M, Haigh R, Trappetti C, Ogunniyi AD, Shewell LK, Boitano M, Clark TA, Korlach J, Blades M, Mirkes E, Gorban AN, Paton JC, Jennings MP, Oggioni MR. 2014. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. Nat Commun 5:5055. https://doi.org/10.1038/ncomms6055.
14. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, Donohoe K, Harris D, Murphy L, Quail MA, Samuel G, Skovsted IC, Kaltoft MS, Barrell B, Reeves PR, Parkhill J, Spratt BG. 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet 2:e31. https://doi.org/10.1371/journal.pgen.0020031.
15. Arrecubieta C, Garcia E, López R. 1995. Sequence and transcriptional analysis of a DNA region involved in the production of capsular polysaccharide in *Streptococcus pneumoniae* type 3. Gene 167:1–7. https://doi.org/10.1016/0378-1119(95)00657-5.
16. Hoskins J, Alborn WE, Arnold J, Blaszczak LC, Burgett S, DeHoff BS, Estrem ST, Fritz L, Fu DJ, Fuller W, Geringer C, Gilmour R, Glass JS, Khoja H, Kraft AR, Lagace RE, LeBlanc DJ, Lee LN, Lefkowitz EJ, Lu J, Matsushima P, McAhren SM, McHenney M, McLeaster K, Mundy CW, Nicas TI, Norris FH, O'Gara M, Peery RB, Robertson GT, Rockey P, Sun PM, Winkler ME, Yang Y, Young-Bellido M, Zhao G, Zook CA, Baltz RH, Jaskunas SR, Rosteck PR, Skatrud PL, Glass JI. 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol 183:5709–5717. https://doi.org/10.1128/JB.183.19.5709-5717.2001.
17. Prudhomme M, Martin B, Mejean V, Claverys JP. 1989. Nucleotide sequence of the *Streptococcus pneumoniae hexB* mismatch repair gene: homology of HexB to MutL of *Salmonella typhimurium* and to PMS1 of *Saccharomyces cerevisiae*. J Bacteriol 171:5332–5338. https://doi.org/10.1128/jb.171.10.5332-5338.1989.
18. Iannelli F, Pearce BJ, Pozzi G. 1999. The type 2 capsule locus of *Streptococcus pneumoniae*. J Bacteriol 181:2652–2654. https://doi.org/10.1128/JB.181.8.2652-2654.1999.
19. Griffith F. 1928. The significance of pneumococcal types. J Hyg (Lond) 27:113–159. https://doi.org/10.1017/s0022172400031879.
20. Avery OT, Macleod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79:137–158. https://doi.org/10.1084/jem.79.2.137.
21. MacLeod CM, Krauss MR. 1947. Stepwise intratype transformation of pneumococcus from R to S by way of a variant intermediate in capsular polysaccharide production. J Exp Med 86:439–452. https://doi.org/10.1084/jem.86.6.439.
22. Austrian R. 1953. Morphologic variation in pneumococcus. I. An analysis of the bases for morphologic variation in pneumococcus and description of a hitherto undefined morphologic variant. J Exp Med 98:21–34. https://doi.org/10.1084/jem.98.1.21.
23. Ravin AW. 1959. Reciprocal capsular transformations of pneumococci. J Bacteriol 77:296–309. https://doi.org/10.1128/jb.77.3.296-309.1959.
24. Lanie JA, Ng W-L, Kazmierczak KM, Andrzejewski TM, Davidsen TM, Wayne KJ, Tettelin H, Glass JI, Winkler ME. 2007. Genome sequence of Avery's

Downloaded from https://journals.asm.org/journal/mra on 14 October 2021 by 82.60.166.172.

virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6. J Bacteriol 189:38–51. https://doi.org/10.1128/JB.01148-06.

25. Taylor HE. 1949. Additive effects of certain transforming agents from some variants of pneumococcus. J Exp Med 89:399–424. https://doi.org/10.1084/jem.89.4.399.
26. Ottolenghi E, Hotchkiss RD. 1962. Release of genetic transforming agent from pneumococcal cultures during growth and disintegration. J Exp Med 116:491–519. https://doi.org/10.1084/jem.116.4.491.
27. Tomasz A, Hotchkiss RD. 1964. Regulation of the transformability of pneumococcal cultures by macromolecular cell products. Proc Natl Acad Sci U S A 51:480–487. https://doi.org/10.1073/pnas.51.3.480.
28. Hahn C, Harrison EM, Parkhill J, Holmes MA, Paterson GK. 2015. Draft genome sequence of the *Streptococcus pneumoniae* Avery strain A66. Genome Announc 3:e00697-15. https://doi.org/10.1128/genomeA.00697-15.
29. Austrian R, Bernheimer HP, Smith EE, Mills GT. 1959. Simultaneous production of two capsular polysaccharides by pneumococcus. II. The genetic and biochemical bases of binary capsulation. J Exp Med 110:585–602. https://doi.org/10.1084/jem.110.4.585.
30. Dillard JP, Vandersea MW, Yother J. 1995. Characterization of the cassette containing genes for type 3 capsular polysaccharide biosynthesis in *Streptococcus pneumoniae*. J Exp Med 181:973–983. https://doi.org/10.1084/jem.181.3.973.
31. Guild WR, Shoemaker NB. 1974. Intracellular competition for a mismatch recognition system and marker-specific rescue of transforming DNA from inactivation by ultraviolet irradiation. Mol Gen Genet 128:291–300. https://doi.org/10.1007/BF00268517.
32. Oggioni MR, Iannelli F, Pozzi G. 1999. Characterization of cryptic plasmids pDP1 and pSMB1 of *Streptococcus pneumoniae*. Plasmid 41:70–72. https://doi.org/10.1006/plas.1998.1364.
33. Claverys JP, Lacks SA. 1986. Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol Rev 50:133–165. https://doi.org/10.1128/mr.50.2.133-165.1986.
34. Lacks SA, Mannarelli BM, Springhorn SS, Greenberg B. 1986. Genetic basis of the complementary DpnI and DpnII restriction systems of *S. pneumoniae*: an intercellular cassette mechanism. Cell 46:993–1000. https://doi.org/10.1016/0092-8674(86)90698-7.
35. Havarstein LS, Coomaraswamy G, Morrison DA. 1995. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. Proc Natl Acad Sci U S A 92:11140–11144. https://doi.org/10.1073/pnas.92.24.11140.
36. Pozzi G, Masala L, Iannelli F, Manganelli R, Havarstein LS, Piccoli L, Simon D, Morrison DA. 1996. Competence for genetic transformation in encapsulated strains of *Streptococcus pneumoniae*: two allelic variants of the peptide pheromone. J Bacteriol 178:6087–6090. https://doi.org/10.1128/jb.178.20.6087-6090.1996.
37. Pestova EV, Håvarstein LS, Morrison DA. 1996. Regulation of competence for genetic transformation in *Streptococcus pneumoniae* by an auto-induced peptide pheromone and a two-component regulatory system. Mol Microbiol 21:853–862. https://doi.org/10.1046/j.1365-2958.1996.501417.x.
38. Iannelli F, Oggioni MR, Pozzi G. 2005. Sensor domain of histidine kinase ComD confers competence pherotype specificity in *Streptococcus pneumoniae*. FEMS Microbiol Lett 252:321–326. https://doi.org/10.1016/j.femsle.2005.09.008.
39. Iannelli F, Oggioni MR, Pozzi G. 2002. Allelic variation in the highly polymorphic locus *pspC* of *Streptococcus pneumoniae*. Gene 284:63–71. https://doi.org/10.1016/S0378-1119(01)00896-4.
40. Iannelli F, Chiavolini D, Ricci S, Oggioni MR, Pozzi G. 2004. Pneumococcal surface protein C contributes to sepsis caused by *Streptococcus pneumoniae* in mice. Infect Immun 72:3077–3080. https://doi.org/10.1128/IAI.72.5.3077-3080.2004.

Downloaded from https://journals.asm.org/journal/mra on 14 October 2021 by 82.60.166.172.

CHAPTER 8. General conclusions

Lactic acid bacteria are a heterogeneous group of microorganisms that includes strains of interest for both commercial and health purposes. The study of their genomes is important to better understand pathways and components responsible for environmental interactions and peculiar properties of single strains. In this thesis, I obtained and investigated the whole genome of lactic acid bacteria belonging to different species including *L. crispatus*, *E. faecalis* and *S. pneumoniae*, by analyzing the genomic features of both chromosome and mobilome associated to environmental response and antimicrobial resistance. The main focus was on the characterization of the *L. crispatus* M247 probiotic strain genome. It was demonstrated that the M247 mobilome includes an integrative and mobilizable element named Tn*7088*, carrying a class I bacteriocin biosynthetic gene cluster homologous to the listeriolysin S gene cluster of *Listeria monocytogenes*. Therefore, Tn*7088* may confer a niche adaptive advantage to its bacterial host. Then, the presence of genomic instability in the M247 strain was proved consisting of ISs mediated chromosomal rearrangements involving two DNA regions of 69.9-kb and 15.4-kb in length. These chromosomal rearrangements were probably implicated in the duplication of the 69.9-kb region which produced two long inverted repeats in the genome of a *L. crispatus* laboratory strain namely M247_Siena. Quantification analysis of chromosomal rearrangements, indicated that the newly generated 69.9-kb long inverted repeats of M247_Siena increased the intrinsic genomic instability of strain M247. In the second part of the thesis, the use of whole genome sequencing (for MLST) and antimicrobial susceptibility testing on a collection of infertility-associated *E. faecalis* showed that the enterococcal isolates that were resistant to high-level aminoglycosides had a clonal structure. The strains that clustered in the clonal complex/sequence type CC16/ST480 were further investigated by genomic comparison analysis and a novel composite transposon named Tn*7086* was identified. Characterization of Tn*7086* and of other Tn*7086*-like elements indicated that this new family of transposons shared integration site, excision/integration mechanism and also the genes conferring aminoglycosides resistance. Finally, the complete genome sequence of the type strain of *L.*

crispatus namely ATCC 33820 useful for a better understanding of the *L. crispatus* species characteristic traits, was obtained. The genome sequences of the *S. pneumoniae* Rx1 strain, a common laboratory strain devoid of the Mismatch Repair System, and of its parental strain R36A were also determined to track genome evolution.

APPENDIX. Scientific *Curriculum Vitae*

- **Education**
  - October 2018 - present: **PhD student** in the XXXIV cycle of doctoral program in Medical Biotechnologies by the Medical Biotechnologies Department of the University of Siena, Italy.

    Main areas of interest: bacterial genomics, whole genome sequencing, NGS, data analysis.
  - October 2016 - September 2018: **Master's Degree** in Medical Biotechnologies (courses held in English), Department of Medical Biotechnologies, University of Siena, Italy.

    Thesis title: "Complete genome sequences of the *Lactobacillus crispatus* probiotic strain M247 and its isogenic nonaggregating mutant Mu5". Mark: 110/110 cum laude and special mention.
  - October 2012 - July 2016: **Bachelor's Degree** in Biotechnologies, University of Pisa, Italy.

    Thesis title: "Effects of berberine on cell migration analysed in two different human cancer cell lines". Mark: 106/110.

- **Training courses**
  - 2020, May. "**Metagenomics applied to surveillance of pathogens and antimicrobial resistance**" on-line course Coursera platform, organized by Technical University of Denmark – DTU, Denmark.
  - 2020, April. "**Whole genome sequencing of bacterial genomes – tool and application**" on-line course Coursera platform, organized by Technical University of Denmark – DTU, Denmark.

- 2019, May. “**Introduction to Machine Learning algorithms**” ALMALE (2018DU0092), organized by the University of Siena, project “Tuscan Start-Up Academy 4.0”, Regione Toscana funds.

- **Languages**
  - **Italian**: native
  - **English**: very good knowledge of English language (written and spoken). 2018: B2 English qualification, Centro Linguistico Ateneo (CLA), University of Siena, Italy

- **List of publications**
  1. Complete Genome Sequence of *Lactobacillus crispatus* Type Strain ATCC 33820. Lucia Teodori a[†], **Lorenzo Colombini[†]**, Anna Maria Cuppone, Elisa Lazzeri, David Pinzauti, Francesco Santoro, Francesco Iannelli and Gianni Pozzi. *Microbiol Resour Announc*. 2021 August 12. doi: 10.1128/MRA.00634-21. (†These authors contributed equally to this work)
  2. Complete Genome Sequence of *Streptococcus pneumoniae* Strain Rx1, a Hex Mismatch Repair-Deficient Standard Transformation Recipient. Anna Maria Cuppone*, **Lorenzo Colombini***, Valeria Fox, David Pinzauti, Francesco Santoro, Gianni Pozzi, Francesco Iannelli. *Microbiol Resour Announc*. 2021 October 14. doi: 10.1128/MRA.00799-21. (*These authors contributed equally to this work)

- **Conferences**
  - 2021, 1-3. Nanopore Community Meeting, virtual conference organized by Oxford Nanopore Technologies, UK.
  - 2021, 17-18 June. London Calling 2021, virtual conference organized by Oxford Nanopore Technologies, UK.

- 2020, 1-3 December. Nanopore Community Meeting, virtual conference organized by Oxford Nanopore Technologies, UK
- 2020, 17-18 June. London Calling 2020, virtual conference organized by Oxford Nanopore Technologies, UK.
- 2019, 19-22 June. XXXIII SIMGBM Congress, Microbiology 2019, Florence, Italy. "Complete genome sequence of *Lactobacillus crispatus* M247 strain and its derivative Mu5 lacking the auto-aggregation phenotype". Lorenzo Colombini, Francesco Santoro, Anna Maria Cuppone, David Pinzauti, Gianni Pozzi, Francesco Iannelli. **Poster**.

- **Nucleotide sequences deposited in GenBank:**

1. *Streptococcus pneumoniae* strain R36A, complete genome. GenBank Accession no. CP079922, BioProject ID PRJNA748391. Cuppone,A.M., Colombini,L., Santoro,F., Pozzi,G and Iannelli,F. 2021
2. *Streptococcus pneumoniae* strain Rx1, complete genome. GenBank Accession no. CP079923, BioProject ID PRJNA748391. Cuppone,A.M., Colombini,L., Santoro,F., Pozzi,G. and Iannelli,F. 2021
3. *Lactobacillus crispatus* strain ATCC33820, complete genome. GenBank Accession no. CP072197.1, BioProject ID PRJNA716945. Colombini,L., Teodori,L., Cuppone,A.M., Lazzeri,E., Pinzauti,D., Santoro,F., Iannelli,F. and Pozzi,G. 2021
4. *Lactobacillus crispatus* strain M247, complete genome. GenBank Accession no. CP088015, BioProject ID PRJNA782912. Colombini,L., Santoro,F., Morelli,L., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.
5. *Lactobacillus crispatus* strain Mu5, complete genome. GenBank Accession no. CP054313, BioProject ID PRJNA634156. Colombini,L., Pinzauti,D.,

Cuppone,A.M., Lazzeri,E., Santoro,F., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.

6. *Lactobacillus crispatus* strain M247_Siena, complete genome. GenBank Accession no. CP046589, BioProject ID PRJNA594001. Colombini,L., Pinzauti,D., Cuppone,A.M., Lazzeri,E., Santoro,F., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.

- **Nucleotide sequences deposited in Sequence Read Archive:**
  1. *Streptococcus pneumoniae* strain R36A whole genome sequencing. BioProject ID PRJNA748391. Cuppone,A.M., Colombini,L., Santoro,F., Pozzi,G and Iannelli,F. 2021
  2. *Streptococcus pneumoniae* strain Rx1 whole genome sequencing. BioProject ID PRJNA748391. Cuppone,A.M., Colombini,L., Santoro,F., Pozzi,G. and Iannelli,F. 2021
  3. *Lactobacillus crispatus* strain ATCC33820 whole genome sequencing. BioProject ID PRJNA716945. Colombini,L., Teodori,L., Cuppone,A.M., Lazzeri,E., Pinzauti,D., Santoro,F., Iannelli,F. and Pozzi,G. 2021
  4. *Lactobacillus crispatus* strain M247 whole genome sequencing. BioProject ID PRJNA782912. Colombini,L., Santoro,F., Morelli,L., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.
  5. *Lactobacillus crispatus* strain Mu5 whole genome sequencing. BioProject ID PRJNA634156. Colombini,L., Pinzauti,D., Cuppone,A.M., Lazzeri,E., Santoro,F., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.
  6. *Lactobacillus crispatus* strain M247_Siena whole genome sequencing. BioProject ID PRJNA594001. Colombini,L., Pinzauti,D., Cuppone,A.M., Lazzeri,E., Santoro,F., Iannelli,F. and Pozzi,G. 2022. Not publicly available yet.