Faculty of Veterinary Medicine
Department of Veterinary Biosciences
University of Helsinki, Finland

Genomic insights about the *Lactobacillus* genus

Ravi Kant

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Veterinary Medicine of the University of Helsinki, for public examination in the Walter Auditorium, EE-Building, Agnes Sjöbergin katu 2, Helsinki, on June 26, 2018, at 12 noon.

Helsinki 2018

Supervisors: **Docent Ingemar von Ossowski**
Department of Veterinary Biosciences
University of Helsinki
Helsinki, Finland

Professor Emeritus Airi Palva
Department of Veterinary Biosciences
University of Helsinki
Helsinki, Finland

Pre-examiners: **Docent Jaana Mättö**
Laboratory Services
Finnish Red Cross Blood Service
Helsinki, Finland

Docent David Fewer
Department of Food and Environmental Sciences
University of Helsinki
Helsinki, Finland

Opponent: **Professor Per Saris**
Department of Food and Environmental Sciences
University of Helsinki
Helsinki, Finland

Custos: **Professor Olli Vapalahti**
Department of Veterinary Biosciences
University of Helsinki
Helsinki, Finland

ISBN 978-951-51-4332-7 (paperback)
ISBN 978-951-51-4333-4 (PDF)
http://ethesis.helsinki.fi

Unigrafia, Helsinki University Print

Helsinki 2018

Contents

List of original publications

This thesis is based on the three original publications listed below. Each article is indicated in the text by the Roman numerals I, II, and III. Complete articles are appendices to this thesis.

I. **Ravi Kant**, Jochen Blom, Airi Palva, Roland J. Siezen, and Willem M. de Vos (2011) Comparative genomics of *Lactobacillus*. Microbial Biotechnology 4(3):323-332.

II. **Ravi Kant**, Johanna Rintahaka, Xia Yu, Pia Sigvart-Mattila, Lars Paulin, Jukka-Pekka Mecklin, Maria Saarela, Airi Palva, and Ingemar von Ossowski (2014) A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. PLoS ONE 9(7):e102762.

III. **Ravi Kant**, Airi Palva, and Ingemar von Ossowski (2017) An *in silico* pan-genomic probe for the molecular traits behind *Lactobacillus ruminis* gut autochthony. PLoS ONE 12(4):e0175541.

Abbreviations

ABC	ATP-binding cassette
BLAST	basic local alignment search tool
CDS	coding DNA sequence
COG	clusters of orthologous group
CRISPR	clustered regularly interspaced short palindromic repeat
DE	description lines
ECM	extracellular matrix
EMP	Embden-Meyerhof-Parnas
Fbp	fibronectin-binding protein
FBP	fructose-1,6-bisphosphate
FucP	fucose permease
GI	genomic island
GPH	glycoside-pentoside-hexuronide
HT	high-throughput
IS	insertion sequence
LAB	lactic acid bacteria
LCG	*Lactobacillus* core genome
LGT	lateral gene transfer
MabA	modulator of adhesion and biofilm
MCF	mucus-binding factor
MCP	methyl-accepting chemotaxis protein
NGS	next-generation sequencing
OHS	cation symporter, oligosaccharide:H^+ symporter
OLC	overlap-layout-consensus
ORF	open reading frame
PCR	polymerase chain reaction
PK	phosphoketolase
PTS	phosphotransferase system
RBH	reciprocal best hit
SDP	sortase-dependent protein
SMRT	single-molecule real-time
SNP	single-nucleotide polymorphism
SOLiD	sequencing by oligonucleotide ligation and detection
TCDB	transporter classification database
WGS	whole-genome shotgun

Abstract

This thesis details an *in silico* exploration of the genetic potential of the Gram-positive genus *Lactobacillus* through the establishment and analysis of a pan-genome dataset. Lactobacilli are an intensively researched and studied group of bacteria, which is in part owed to their exploitation for various man-made food and industrial purposes and their advocated beneficial use as gut probiotics. Bacterial pan-genomics is an outgrowth of the comparative genomics field and by comparing the genomes from many strains of the same species the data obtained serves to catalogue the entire genetic repertoire available to a particular species. In effect, the pan-genome represents the complete assortment of putative genes in a species (or genus) of bacteria. The research presented in this thesis is comprised of three distinct studies (I, II, and III), each of which used the pan-genome approach to give a theoretical account of *Lactobacillus* genetics.

In study I, a pan-genome was assembled at the genus level using 20 fully sequenced genomes from 14 different *Lactobacillus* species. Here, complete genome sequences were selected for creating a broader framework that would allow more comprehensive genomic comparisons. Varying aspects were apparent among the genome sequences used, including sizes that ranged from ~1.8 to ~3.3 Mbps and a G+C content of between ~33% and ~51%. The assembled pan-genome was sized at 14,000 protein-encoding genes, and out of which a small 383-gene *Lactobacillus* core genome (LCG) was derived. The genetic content of the LCG was used for reconstructing a molecular phylogeny, which then permitted a taxonomic grouping of the 20 genomes into three main clades. Additional classifications of the LCG involved identifying core group and signature group genes, as well as so-called ORFan genes that were further sorted as either LCG-specific or group-specific.

In study II, a pan-genome of the *Lactobacillus rhamnosus* species was constructed from 13 different genomes. *L. rhamnosus* is a highly adaptable bacterium that thrives in a variety of hosts and environments. Presumably, there is little doubt that numerous genetic peculiarities are the source of the niche-related phenotypes that enable the inherent ecological adaptability of *L. rhamnosus* strains. For this, the genetic content of the assembled pan-genome was examined for those geno-phenotypic variations occurring at the cell-surface level and whether these correlate to a particular habitat preference of various *L. rhamnosus* strains. The *L. rhamnosus* pan-genome itself had an estimated size of 4,893 protein-encoding genes, which was further partitioned into the 2,095-gene core and 2,798-gene accessory genomes. Pan-genomic comparisons were benchmarked against the gut-adapted *L. rhamnosus* GG strain and focused primarily on seven functionally characterized surface-exposed proteins. Most notably, the operonic genes for the mucoadhesive SpaCBA pilus were part of the

accessory genome and can be regarded as a genomic novelty in *L. rhamnosus*. Nonetheless, for those *L. rhamnosus* strains with a functional SpaCBA piliation trait, this would improve niche-specific fitness and presumably prolong transient (allochthonous) colonization of mucosal epithelial surfaces in the gut or elsewhere in the body.

In study III, a pan-genomic appraisal was performed on the *Lactobacillus ruminis* species by compiling the genomes of nine different strains obtained from human, bovine, porcine, and equine digestive tracts. *L. ruminis* is a piliated and flagellated strict anaerobe and one of the few indigenous (autochthonous) lactobacilli in the gut. The pan-genome was utilized to pinpoint the molecular basis for the intractable colonization behavior of *L. ruminis*, where the focus was on those geno-phenotypes associated with cellular surface morphology and anaerobic fermentation and respiration. The size of the *L. ruminis* pan-genome was predicted to contain 4,301 protein-related genes, while the number of genes in the core and accessory genomes was 1,234 and 3,067, respectively. As inferred from the pan-genomic data, the presence of certain surface proteins and a substitute anaerobic energy-yielding metabolism might represent the adaptive phenotypes that help make *L. ruminis* a gut-autochthonic species.

1. Literature review

1.1 Determination of bacterial genome sequences

The sequence determination of a genome offers the unique opportunity to decode the phenotypic, physiological, and ecological properties of any type of microorganism. Thus, the field of microbial genomics underwent a seismic advance twenty-three years ago when the genome of *Haemophilus influenzae* was first to be sequenced (Fleischmann et al., 1995). This became possible as DNA sequencing technologies went through a revolutionary modernization during the 1990s, and a direct consequence of consortium-based projects to sequence the genomes of model microbes such as *Escherichia coli* and *Bacillus subtilis* (Burland et al., 1993; Glaser et al., 1993). However, a "big bang" of sorts occurred in 1995 when Craig Venter and his team performed the first shotgun sequencing of complete bacterial genomes (Fleischmann et al., 1995), which led to the development of the whole-genome shotgun (WGS) sequencing approach. It was some ten years later when the next revolution in DNA sequencing technology took place, with the introduction of the high-throughput (HT) or next-generation sequencing (NGS) method (Margulies et al., 2005; Shendure et al., 2005; Mardis, 2008). These and related sequencing platforms began arriving on the commercial market during the 2000s decade, and by being coupled with innovative bioinformatic tools and approaches at the sequencing facilities of universities and public health care institutes, it was soon accompanied by an ever-expanding growth of deposited bacterial genome sequences in public databases (**Figure 1**). More recently, a further revolution in DNA sequencing was the advancement in the long-read technologies. Here, the first long-read technology that gained widespread popularity and use was the single-molecule real-time (SMRT) sequencing method developed by Pacific Biosciences (Eid et al., 2009). This particular approach provides high-quality assemblies (Bashir et al., 2012; Chin et al., 2013) on its own or in combination with short-read sequencing, and will likely offer the possibility of a new era of more fully completed genome sequences.

Throughout the years, the expansion and improvement of DNA sequencing technologies has fueled an explosive increase in the number of sequenced bacterial genomes, thus providing new biological information and clues to a better understanding of the molecular biology for a variety of different bacteria. For instance, comparative analyses that compile the genomes of different strains from the same species into what is called a "pan-genome" have revealed the gene content within an entire species is much more than that of a single strain (Tettelin et al., 2008; Vernikos et al., 2015). Moreover, this sort of study also helps in understanding one of the dominant genetics forces behind

bacterial evolution, namely the concept of lateral gene transfer between microorganisms (Tettelin et al., 2008; Vernikos et al., 2015). Still further, genomic comparisons of different bacterial genera and species have helped to reveal the evolutionary origins of virulence and niche specification (e.g., Dettman et al., 2013). In addition, with the steady advance in sequencing technologies, this has allowed light to be shed on the genetics of microbial interactions, e.g., via the comparative metagenomic and metatranscriptomic analyses of bacterial communities (Qin et al., 2010; Forde and O'Toole, 2013; Jorth et al., 2013; Oh et al., 2014).

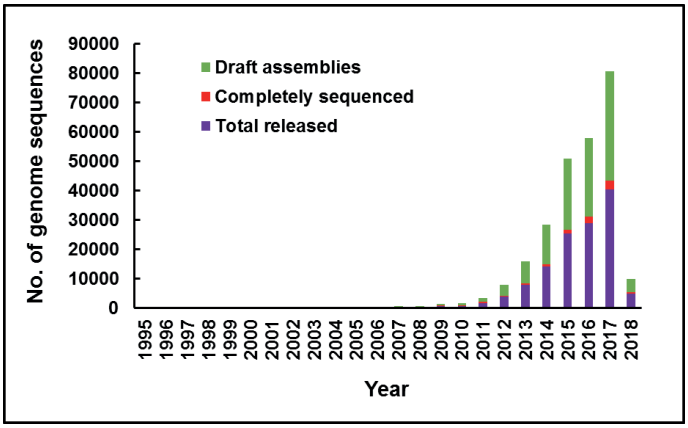

Figure 1. Deposition of bacterial genome sequences in the NCBI database (1995-2018; May 2).

1.2 DNA sequencing platforms

British biochemist Fred Sanger revolutionized the field of molecular biology in the late 1970s by his development of the chain-termination approach for sequencing DNA molecules (Sanger et al., 1977). Amongst bacteria, the first genomes sequenced had involved creating and mapping a library of large-insert DNA clones, and afterward generating small-insert libraries from each of these clones, which ultimately enabled the sequencing of the DNA inserts (see **Figure 2**). Stemming from a slow but steady advance in DNA sequencing technologies, but particularly upon the arrival of the WGS approach, the rapid elucidation of entire bacterial genome sequences has become commonplace and routine (see section 1.1). Although DNA sequencing had mostly been based on the Sanger method for upwards of three decades (Sanger et al., 1977), this technique began to lose its popularity during the past ten years with the advent of NGS, for which the technology has proven to be much more efficient and affordable (Margulies et al., 2005; Metzker, 2005; Shendure et al., 2005; Schuster, 2008). Once further developing the technological capacity to generate long read lengths (MacLean et

al., 2009; van Dijk et al., 2014), NGS was soon favored for sequencing bacterial genomes, though the Sanger-based methodology continues to be used for small-scale sequencing, especially for closing sequence gaps in whole-genome sequencing projects.

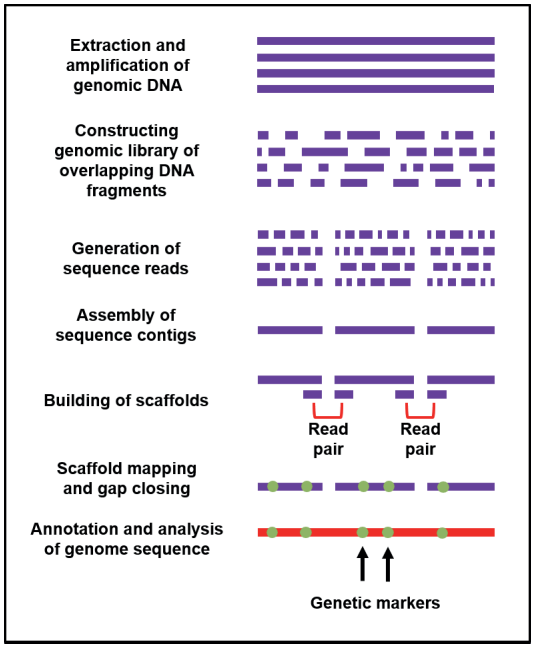

Figure 2. Illustrative outline of the basic workflow needed for a bacterial genome-sequencing project via the whole-genome shotgun approach (adapted from Adams, 2008).

WGS sequencing has been a popular approach for revealing the genetic makeup of various bacterial genera and species (see **Figure 2**). Briefly, this technique involves cultivating a microbe of interest from a single colony, followed by the extraction of its chromosomal DNA and then using this genetic material to prepare a library. The amount of DNA required for sequencing can vary depending on the method used (Loman et al., 2012). For the next step, the genomic DNA is broken into randomly overlapping fragments and used as templates for polymerase chain reaction (PCR) amplification, with the PCR products then sequenced from one or both ends and mapped as either single-end or paired-end reads. As a former approach, cloned plasmids with different fragments of genomic DNA were used and then sequenced with the Sanger method. The final step of the WGS process involves detecting the overlaps between the reads and generating a set of non-overlapping segments called "contigs". Frequently, a fully sequenced genome cannot be achieved from a single shotgun library since not all sequence reads will overlap, and thus additional PCR amplification or cloning of DNA regions followed by sequencing is used to fill in any gaps between the contigs for generating a complete bacterial genome

sequence. Earlier on, these sequence gaps were closed for many of the genomes being sequenced, but this extra step is sometimes at a higher labor cost (see **Table 1**) and could represent a financial bottleneck for present-day large-scale genome sequencing projects. However, with the emergence of the SMRT technologies that enabled read lengths greater than 10 kb (Eid et al., 2009), this provided an inexpensive and convenient solution for sealing gaps in genome sequences, and now reflects the increased number of completely sequenced genomes made available during the past few years (see **Figure 1**).

DNA sequencing platform	Model year	DNA amplification	Sequencing mechanism	Total read length	Time/ run	Sequence data output/run	Accuracy	Estimated cost/Mb (USD)
454 GS FLX	2007	Emulsion PCR	Pyrosequencing	700 bp	24 hours	0.7 Gb	~100%	$10
PacBio RS	2011	Single molecule (no amplification)	Fluorophore-linked nucleotides	1500 bp	2 hours	100 Mb	~87%	$2
Hi Seq 2000	2010	Bridge amplification	Reversible terminator-based method	150 bp	3-11 days	600 Gb	~98%	$0.07
SOLiDv4	2010	Emulsion PCR	Ligation and two-base coding	35-50 bp	7-14 days	120 Gb	~98%	$0.13
Ion Torrent PGM	2011	Emulsion PCR	Ion semiconductor	200 bp	2 hours	20-1000 Mb	~98%	$1
Sanger 3730xl	2002	PCR	Dideoxy chain termination	400-900 bp	20-180 mins.	1.9-84 Kb	~100%	$2400

Table 1. Basic attributes of DNA sequencing platforms (adapted from Liu et al., 2012 and Quail et al., 2012).

As summarized in **Table 1**, there are half a dozen DNA sequencing platforms that are routinely used for determining a bacterial genome sequence. Genome sequencing by the Sanger method (Sanger et al., 1977), which relies on random termination of DNA synthesis, was first developed and its protocol begins with heat denaturation of the genomic DNA fragments into single strands, followed by the addition of a short complementary synthetic oligonucleotide primer strand (Sanger et al., 1977). The DNA primer is designed so that its 3' end is upstream of the DNA segment to be sequenced and is normally radio-labelled to permit the detection of the final product on the sequencing gel via autoradiography. The DNA solution with the annealed primer and template is aliquoted into four separate reaction mixes containing each type of nucleotide (i.e., ATP, CTP, GTP, and TTP), a DNA polymerase enzyme, and sparing quantities of a single dideoxynucleotide (i.e., ddATP, ddCTP, ddGTP, or ddTTP). As a key component of the sequencing method, each type of dideoxynucleotide lacks a 3'-hydroxyl group, which prevents the formation of a phosphodiester bond with another nucleotide, and thus curtailing any additional DNA polymerase-catalyzed DNA synthesis. With the random inclusion of a dideoxynucleotide into the synthesized DNA, this will generate DNA strands of various lengths, and once denatured can be sized electrophoretically on polyacrylamide gels. After making an autoradiograph of the sequencing gel, the band pattern of the four reaction mixes is read from bottom to top, alternating from one reaction mix to another,

to give the complementary DNA sequence of the template strand in the 5' to 3' direction (**Figure 3**). Added innovations have seen radioactivity replaced with the use of fluorescent chemical compounds (called fluorophores) emitting colored light, where each nucleotide has its own color and the sequencing reaction is performed in one mix and run as a single lane on the sequencing gel (Guo et al., 2008) (**Figure 3**). This has allowed for DNA sequencing to become automated via computerized machines and instruments, and thereby helping to substantially increase the output of sequenced data.

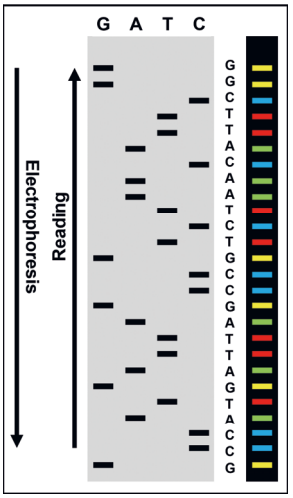

Figure 3. Sanger sequencing method showing the electrophoretic separation of DNA fragments via the use of radioactive labelling (left) or colored fluorophores (right).

With new advances, the Sanger method eventually became outdated for sequencing bacterial genomes and was replaced by several NGS platforms, as these can handle a massive amount of DNA material, with both higher quality and lower costs. Here, the novel aspect of the NGS systems is the enormous capacity of their template cluster feature, wherein there are numerous copies of template DNA, e.g., these being amplified by emulsion PCR (Dressman et al., 2003). The clusters of DNA template can then be sequenced using various approaches, e.g., by the ion semiconductor method (Rothberg et al., 2011), the ligation technique (Shendure et al., 2005), or pyrosequencing (Margulies et al., 2005). Alternatively, the SMRT sequencing platform (Eid et al., 2009) operates differently by using "zero-mode waveguides", where each sequencing well to which single molecules of the coupled DNA polymerase and DNA template are attached is illuminated from only the bottom, and "fluorophore-linked nucleotides", whose incorporation into the growing DNA strand by DNA polymerase can be monitored in real time. However, a trade-off with the various available DNA sequencing systems is between the amount of data produced and their cost and read length. For instance, whereas the Illumina and SOLiD

(Sequencing by Oligonucleotide Ligation and Detection) platforms (**Table 1**) are less expensive and both can generate a large amount of sequenced data, the read lengths are shorter. Thus, genomic sequencing using such systems are best suited for projects with a high number of bacterial genomes or when genomes need to be sequenced again. In contrast, for those sequencing platforms giving longer read lengths it becomes more costly and less sequenced data is generated (**Table 1**). In terms of cost-effectiveness, these approaches are more suitable for sequencing projects with fewer bacterial genomes.

Sequencing fidelity represents an important parameter when determining a bacterial genome sequence, and here the different sequencing platforms tend to vary in the number of mistakes that might arise during a particular run (Mardis, 2008; Liu et al., 2012; Quail et al., 2012). For instance, the pyrosequencing and Ion Torrent technologies use the intensity of signals as the means to distinguish the various nucleotides, and with such an approach the accurate detection of long stretches of the same base, e.g., 4 or more and called homopolymers, becomes challenging and is a common source of sequence error (Voelkerding et al., 2009). On the other hand, Illumina sequencing is routinely plagued by noise-related errors (Sheikh and Erlich, 2012). Noise from "fading" is attributed to a waning intensity of the fluorescence signal that drops below the detection threshold with every repeated cycle of sequencing. Noise also arises from "phasing" when either no nucleotide or an extra one is added during a sequencing round, at which point the signal detection becomes muddled from an amalgam of different DNA strands. Another unwanted signal is due to excitation crosstalk between the various nucleotide fluorophores. So far, however, the most error-free DNA sequences can be achieved with the SOLiD system (Liu et al., 2012), in which the sequencing process is by ligation with DNA ligase rather than by DNA polymerase-driven synthesis. The precision of this sequencing platform results from the fact that all bases of the DNA template are read twice, and thus during the data decoding process, any "authentic" errors, e.g., from single base changes, must be detected two times as well. Yet, drawbacks with such a system include short read lengths (75 bp) and a long run time (one week) (Garrido-Cardenas et al., 2017). Moreover, ligation-based DNA sequencing methods are known to have difficulty with discerning palindromic sequences (Huang et al., 2012). Anecdotally, the analysis and assembly of SOLiD sequenced data tends to be more demanding and any use with various other genomic tools and approaches is somewhat problematic.

1.3 Genome sequence data preprocessing and assembly

Since NGS is able to generate a vast amount of genome sequence data, the handling and storage of such large datasets typically requires the use of

computer clusters, i.e., a number of interconnected computers working together. Strong computing power is also needed in preprocessing the fluorescent light intensities of the sequencing output to a user-friendly format that yields a readable nucleotide sequence. This part of the sequence data processing pipeline is known as "base-calling" (Ledergerber and Dessimoz, 2011), which through the use of computer programs will automatically predict individual nucleotides based on the intensity of the light signals. For this, manufacturer-embedded software tools are ordinarily used, although other base-caller algorithms under GNU general public license are gaining popularity due to their increased accuracy. During the base-calling step, the fluorescence signals are converted to what are dubbed as "nucleotide calls", each of which is first scored for quality and then fine-tuned for any irregularities that might arise from the particular type of sequencing platform being used (Sheikh and Erlich, 2012). Read quality improvements of the sequencing data can also be made by pruning away unwanted sequences, such as adapters, oligoprimers, low-quality segments, and any other related artifacts. However, although increasing the coverage (sequencing many times over) is a means to improve the correctness of sequencing, the use of an accurate base-caller will lessen the need for undue coverage, and thus ultimately lower the expense of genome sequencing (Ledergerber and Dessimoz, 2011).

Once the preprocessing stage is completed, the genome sequence data is put through the assembly process (see **Figure 2**) (Flicek and Birney, 2009; Pop, 2009; Miller et al., 2010). This involves identifying any overlapping regions, and then based on these nucleotide matches joining the sequenced reads together into longer sets of continuous sequence, commonly known as contigs. Popular algorithms used for contig formation include the overlap-layout-consensus (OLC) and de Bruijn graph methods (Flicek and Birney, 2009; Pop, 2009). The contigs are next sorted in order and connected into what are called "scaffolds" by including gaps to indicate the location of any missing reads of DNA sequence. Scaffolding is typically helped by having the sequence reads of paired DNA fragments, these becoming available when the chromosome is sequenced from both ends. Owing to the sometimes prohibitive cost and tedium of manually closing gaps, an acceptable norm was established in which many bacterial genomes are published or made available as draft assemblies. However, this is beginning to be less frequent as new advances in NGS are producing much longer sequence reads with fewer gaps, and thus more bacterial genomes are sequenced to completion.

1.4 Structural and functional annotation of bacterial genomes

After a bacterial genome has been sequenced, either to completion or in

draft assembly form, the corresponding nucleotide sequence must be annotated or, simply put, explained as to what it all means or represents. This predictive process is commonly known as genome annotation and includes two basic aspects: (1) identifying the presence of the various different genetic elements (i.e., structural annotation) and (2) assigning biological information to each genetic element (i.e., functional annotation) (Beckloff et al., 2012). As the structural and functional annotation of bacterial genomes represents a painstaking and tedious undertaking, each is achieved computationally by software programs. Here, the accompanying analyses involve the use of automated *in silico* prediction tools and algorithms, although manual "hands-on" annotation called curation is often employed and integrated into the process for deciphering a genomic sequence (Beckloff et al., 2012). Moreover, genome annotations are sometimes further enriched biologically, e.g., from the proteomic and mass spectrometric data analysis of gene products (Gupta et al., 2007).

Structural annotation involves gene prediction, and therein the detection of open reading frames (ORFs) encompassing a gene structure with coding regions for proteins (i.e., coding DNA sequence, or CDS) and associated regulatory sites. For bacteria, anywhere upwards to 90% of the genome can be comprised of protein-encoding genes, and while a straightforward approach for their detection would be to screen for various DNA segments (e.g., 100 bp or more in length) bordered by an initiation and termination codon, this tends to yield too many false-positive genes (Koonin and Galperin, 2003). Instead, the inference of genes is helped by different statistical models and algorithms that discriminate between protein and non-protein encoding regions on the basis of characteristic likenesses with other gene products in public databases (Koonin and Galperin, 2003) or the existence of upstream sequence motifs, such as ribosomal binding sites or transcriptional consensus elements (Delcher et al., 2007; Hyatt et al., 2010). Frequently, this approach is accompanied by homology searches against protein databases via, e.g., BLAST (basic local alignment search tool) (Altschul et al., 1997). Yet during the structural annotation phase, there are other encoded parts of the genome that can also be revealed, such as the ribosomal and transfer RNAs (i.e., rRNA and tRNA, respectively), genomic islands (GI), various different mobile genetic elements (e.g., transposons, plasmids, and prophages), and CRISPR (clustered regularly interspaced short palindromic repeat) sequences, and for which a set of tailored computational-based prediction approaches are normally used (Lagesen et al., 2007; Lowe and Eddy, 1997; Laslett and Canback, 2004; Langille et al., 2010; Wagner et al., 2007; Varani et al., 2011; Zhou and Xu, 2010; Lima-Mendez et al., 2008; Zhou et al., 2011; Edgar, 2007).

The genome-wide predictive aspect of the functional annotation method is rather akin to that of structural annotation, with each being a highly complex

and intricate task, and both achieved by utilizing computational tools and databases (Rost et al., 2003; Schnoes et al., 2009). Yet, whereas the structural annotation process is providing the genetic framework for the molecular organization of a bacterial genome, the functional annotation phase that follows is more descriptive and assigns names to gene products (proteins) based on *in silico* predicted biochemical roles and properties (Koonin and Galperin, 2003). Here, functional annotation of bacterial genomes offers explanatory "description lines" (DE) about the various characteristics of a gene product, including its expression, regulation, and interaction (Friedberg, 2006). However, since the format for such DE terms is not uniform and frequently delivered in vague and inconsistent wording, and then along with the definitions for protein function being often biased by subjectivity, this lessens their use for comparing different genome sequences (Klimke et al., 2011). Notwithstanding, over the years, efforts to standardize the functional gene assignment process has led to the development of several usable annotation tools and databases, e.g., such as COG (clusters of orthologous group) (https://www.ncbi.nlm.nih.gov/COG/; Tatusov et al., 1997), UniProt (http://www.uniprot.org/, Chen et al., 2017), SEED (http://www.theseed.org/wiki/Main_Page; Overbeek et al., 2005), TIGRFAM (http://www.jcvi.org/cgi-bin/tigrfams/index.cgi; Haft et al., 2003), BRENDA (http://www.brenda-enzymes.org/; Placzek et al., 2017), and TCDB (transporter classification database) (http://www.tcdb.org/; Saier et al., 2016).

Typically for the functional annotation process, the biological descriptions ascribed to sequence data are via the shared resemblance in primary structure (amino acid sequence) between a putative gene product and other proteins deposited in public databases (Friedberg, 2006; Lee et al., 2007). Here, the implicit assumption is made that homologous proteins with a similar primary structure will likely also have the same function, albeit this is not always the case. Nonetheless, in this method of annotating gene functions, database searching for similar protein sequences is automated and performed using heuristic search algorithms (e.g., BLAST) (Altschul et al., 1997), wherein the name and functional description from the best scoring hit(s) are chosen and used for defining the query gene (protein). However, as this "single hit" approach does not always give the most accurate description or a reliable evolutionary context, other methods that pool the biological information from numerous search hits are used and integrated within the annotation process (e.g., Martin et al., 2004; Hawkins et al., 2006; Wass and Sternberg, 2008). Moreover, prediction accuracy can be further enhanced where a preference is given to the annotational information extracted from orthologous proteins and not those deemed paralogs, thus making a functional distinction between genes evolved from a common ancestor and genes related by genetic duplication (Friedberg, 2006). Searching for motif and domain similarity in protein databases like, e.g., InterPro (https://www.ebi.ac.uk/interpro/; Finn et al., 2017), Pfam

(http://pfam.xfam.org/; Finn et al., 2016), and CDD (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; Marchler-Bauer et al., 2015), represents another predictive approach for improving the correctness of functional annotations, and can be particularly helpful for those sequences having low homology-based hits. In addition, there are other tailored databases and tools that allow sequence similarity searches for different categories of proteins, e.g., such as those for virulence (Zhou et al., 2007), antibiotic resistance (Liu and Pop, 2009), and transcriptional DNA-binding (Wilson et al., 2008) factors, proteolytic (Rawlings et al., 2014) and carbohydrate-active (Lombard et al., 2014) enzymes, as well as bacteriocin (van Heel et al., 2013) and secretion signaling (Petersen et al., 2011) peptides. While not being an exhaustive list and summary of all the methods used for annotating gene function in bacterial genomes, many of these are part of an automated functional assignment pipeline, but whereas others necessitate individual predictions and can be obtained by running scripts, querying online servers, or executing BLAST-type homology searches of databases.

1.5 Computational pipelines for bacterial genome annotation

As mentioned briefly in the section above, the annotation of bacterial genomes can occur through a number of computational pipelines, e.g., such as RAST (http://blog.theseed.org/servers/presentations/t1/annotation-with-rast.html; Aziz et al., 2008), NCBI-PGAAP (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/; Tatusova et al., 2016), JGI-IMG (https://img.jgi.doe.gov/; Markowitz et al., 2012), ERGO (https://www.igenbio.com/ergo/; Overbeek et al., 2003), and JCVI-PAP (https://sourceforge.net/projects/jcviprok/; Davidsen et al., 2010), with these being available as online servers and fully automated or when downloaded and run locally. Most annotation pipelines are user-friendly, as they undergo constant improvement and updating, and as well, are often part of a collection of integrated tools and methods, thus permitting a more efficient utilization and analysis of genome sequence data. Typically, these annotation pipelines are able to pinpoint and identify the structural aspects of the bacterial genome (e.g., CDSs and RNAs), and ultimately assign genetic functions by BLAST-type searching of nucleotide and protein databases for biological information about predicted roles. As a means to limit any likelihood of false positive or negative errors, searches within the annotation pipeline will normally incorporate more than one algorithm for each prediction and later combine the final output. After the structural and functional annotation of protein-coding genes (or otherwise), and depending on the pipeline, these can also be run through various other computational methods and algorithms to, e.g., predict secretomes (Bendtsen et al., 2005a), signal peptides (Bendtsen et al., 2005b; Petersen et al., 2011), and

CRISPR (Barrangou and Horvath, 2012), reconstruct metabolic pathways (Aziz et al., 2008), and establish subcellular localization (Markowitz et al., 2012). While many annotation pipelines have in common similar computational approaches, their output data can often vary and be different for the same bacterial genome (Bakke et al., 2009). Additionally, there are some annotation servers that offer the possibility of modifying and editing the prediction output or also including extra annotation data generated from other pipeline sources.

1.6 Bacterial comparative and pan-genomics

The genetic information brought on from the sequencing and annotation of a bacterial genome is in itself considerably useful for studying and understanding the biological potential of an individual type of microbe. However, further to this is obtaining a much broader view of a particular bacterial species and its strains by making comparisons between their genomes, and in such a way that any genetic relatedness or divergence can potentially be uncovered and scrutinized. Characteristically, the two main approaches to this sort of genetic analysis are (1) comparative genomics and (2) pan-genomics, and each of these will be explained further below.

Comparative genomics represents the basic means by which all newly sequenced bacterial genomes are examined and analyzed, and, as its name implies, involves comparing the whole (or in part) genomes of different species and strains of bacteria (Edwards and Holt, 2013). With a view to determine how the genetics between bacteria types are similar or different, this kind of comparative analysis examines a wide variety of genomic aspects, such as nucleotide sequence, GC content, genes and their synteny (order), GIs, transcriptional and translational regulatory elements, and phyletic pattern. As the underlying basis of comparative genomics, any such genetic attributes in common with different bacteria are expected to retain some ancestral similarity at the DNA level (Edwards and Holt, 2013). Taken from this, phenotypic and molecular inferences can be made about bacteria and their evolutionary relationships, niche and habitat adaptations, etiology and pathogenesis, and ecological interactions. Bacterial comparative genomics normally involves a pairwise or multiple sequence alignment of genomes, and to visualize such comparisons there are a number of available software programs in use (e.g., Mauve, BRIG, and ACT) (Edwards and Holt, 2013). Relatedly, this alignment approach is particularly helpful for arranging the proper order of a draft genome, whereby the sequence reads are aligned and mapped against a completely sequenced reference genome (Batzoglou, 2005). What usually follows next in a bacterial genome comparison is determining which of the orthologous genes are shared and conserved among the different genomes (i.e., orthologous groups).

Typically, this would be achieved using amino acid sequences in a grouping strategy based on, e.g., reciprocal best hit (RBH) BLAST scores (Tatusov et al., 1996). Notably by identifying the orthologs in common, this can verify the functional annotation of a given gene or set of genes (Friedberg, 2006; Lee et al., 2007). Further, as an outcome of their phylogenetic reconstruction, various orthologous "housekeeping" genes can also be useful in revealing the ancestral history or evolutionary lineage of a bacterial species. However, aside from the comparative analysis of genomes being based on sequence alignment techniques, the pairwise similarity between genomes can be visualized graphically by using the dot-matrix approach, e.g., such as through the DOTTER (Sonnhammer and Durbin, 1995) or Gepard (Krumsiek et al., 2007) software applications.

First coming to conceptual prominence in 2005 (Tettelin et al., 2005), bacterial pan-genomics is an offshoot of comparative genomics that compares the genomes of several strains from the same species as the means for determining the overall genetic content that a given species has at its disposal (for review, see Tettelin et al., 2008; Ussery et al., 2009; Guimaraes et al., 2015; Computational Pan-Genomics Consortium, 2018). Here, the pan-genome was conceived to represent the entire collection of potential genes in a species, which in fact can sometimes be double the number for that of a single genome. Not surprisingly, the pan-genome concept can also be extended to bacteria at the genus level (Ussery et al., 2009). Ultimately though, in the case of a species pan-genome, by defining the complete repertoire of all genes this lets one conceptualize an understanding about the molecular and phenotypic interactions between bacteria in their occupied ecological niche or adapted environmental habitat. The genetic makeup of the pan-genome is classified into two integral parts: (1) a core genome and (2) an accessory or dispensable genome (Ussery et al., 2009; Guimaraes et al., 2015). The core genome refers to the set of genes that are conserved in every genome of the pan-genome. For a species pan-genome, the core genes are expected to be present in all strains of a species and thus, in addition to defining the basic genetic nature of a species, these would be considered essential for life by encoding the basic housekeeping and regulatory functions for cellular viability (Guimaraes et al., 2015). The accessory genome is then what remains of the pan-genome, and it is seen to represent the diversity of a bacterial species (Guimaraes et al., 2015). Genetic content included here are the genes found in two or more but not all strains and those genes that are only specific to one strain (called unique). Despite often deemed dispensable for the survival of a bacterial species, some accessory genes can be helpful in other ways, e.g., such as for better tailoring the adaptation of a strain to a particular ecological lifestyle or specific environment (Ussery et al., 2009; Guimaraes et al., 2015). Nonetheless, while identifying the core and accessory genes of a pan-genome will certainly expand the genetic perspective of a bacterial species and its strains, the pan-genome per se is not a natural entity

and must be viewed with some circumspect as a conjectural pooling of genes (Ussery et al., 2009). It should be mentioned that while the phyletic reconstruction of the evolutionary relationships between bacterial species or strains is an important part of any genomic comparison, these are commonly done using housekeeping and 16S rRNA genes. However, tree-building based on genome-wide data (i.e., core genome) is often considered as delivering a more reliable estimate of ancestral lineage and history (Rokas et al., 2003; Blom et al., 2009).

With the number of published pan-genome studies having steadily risen (see Vernikos et al., 2015), software programs have been designed and adopted for wading through the huge amount of data that is normally produced as output (see Guimaraes et al., 2015; Xiao et al., 2015). Here, a host of analysis tools and methods are readily available for various types of characterizations, such as, e.g., plotting pan-genome and core genome curves, constructing phylogenomic trees, identifying and analyzing single-nucleotide polymorphisms (SNPs) and homologous gene clusters, and as well, annotating, curating, and visualizing pan-genomic data. Although the use of Venn diagrams to depict the shared genes among genomes in a pan-genome is customary and helpful for visually illustrating the genetic similarity or variation between strains of a species, plotting the predicted size of the pan-genome (or core genome) is considered a gold standard of pan-genomic analysis and almost invariably such a development curve is always included. For this, measurement of the developing pan-genome size as more sequenced genomes are added is through a conventional Heaps' law plot, with the x-axis representing the increased number of genomes and the y-axis representing the total number of genes. Typically, the curve trajectory for size estimates is derived using the regression models and algorithms developed originally by Tettelin and coworkers (Tettelin et al., 2005; Tettelin et al., 2008), in which, e.g., fitting of pan-genome data can be according to a power law regression model and that for core genome data via exponential regression (**Figure 4**).

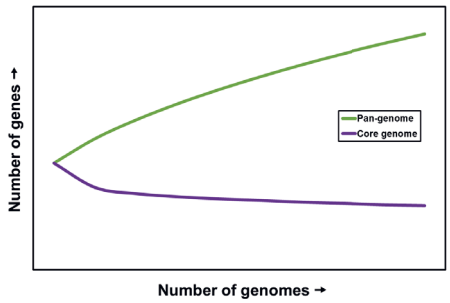

Figure 4. Typical development plot of new genes for a pan-genome and core genome fitted by power law and exponential regression, respectively.

As the power law regression model is contingent on only two variables, i.e., a proportionality constant or intercept and a decay exponent called α, this in turn allows for a descriptive statistical interpretation of the pan-genome data (Tettelin et al., 2008). In this case, a Heaps' law fitting of the α-parameter can gauge the level of openness for a pan-genome, and from which certain inferences can be made about the genetic flexibility of a bacterial species and its adaptiveness to a particular ecological niche or habitat. For example, with a calculated $\alpha < 1$, this suggests the pan-genome is open and that the gene pool for a species is not yet fully characterized (Tettelin et al., 2008). From an ecological perspective, this would support the adaptability of a species to various habitats and/or a changing environment, and perhaps having the proclivity for undergoing LGT. Conversely, when the development plot is calculated as $\alpha > 1$, the pan-genome is considered closed, such that there begins to be a near constancy in the gene pool size of a pan-genome, and thus reflecting the diminished adaptive prowess of a particular species (Tettelin et al., 2008). In some rare instances, if $\alpha = 1$, the pan-genome size would continue to increase, albeit only gradually, and also still retain unlimited genetic potential (Tettelin et al., 2008). By comparison, when exponential regression is applied to characterizing the core genome data of a pan-genome, the steepness and leveling off of the descending curve in the development plot will generally reflect when a stable number of core genes is realized for a bacterial species (Tettelin et al., 2008). As an inference, this might indicate what possible role the genetic traits of the accessory genome can offer different strains as a competitive or adaptive advantage.

1.7 The genus *Lactobacillus*

The Gram-positive genus *Lactobacillus*, which consists primarily of rod-shaped and non-spore-forming microbes, is a prominent member of the so-called lactic acid bacteria (LAB). According to present-day tallies there are just over 200 recognized *Lactobacillus* species (http://www.bacterio.net/lactobacillus.html), and these are able to survive in both aerobic and anaerobic conditions. Lactobacilli tend to grow optimally at mesophilic temperatures (30-40 °C) and under slightly acidic conditions (pH 5.5-6.2), but they also exhibit wide ranges of growth temperature and pH (i.e., 2-53 °C and 3-8, respectively) (Salvetti et al., 2012). While the vast majority of lactobacilli are non-motile, there appear to be at least a dozen species with a motility phenotype (Salvetti et al., 2012). Most *Lactobacillus* species are nutritionally fastidious and largely saccharolytic, and on the latter point they are commonly found to inhabit a wide variety of carbohydrate-rich environments, such as many food products, sewage wastes and effluents, soil, various plant vegetation, and the mucus-lined cavities and orifices of humans and animals (e.g.,

digestive tract, mouth, vagina, and respiratory airways) (Felis and Dellaglio, 2007).

One of the earliest phylogenetic classifications of the genus *Lactobacillus* using 16S rRNA gene sequence was based on 55 species (Collins et al., 1991). At that time, the phylogeny inferred from the sequence data had revealed that the genus *Lactobacillus* is comprised of three divergent taxonomic clusters, i.e., the *L. delbrueckii* group, the *L. casei-Pediococcus* group, and the *Leuconostoc* group. Since then, many additional species have been identified over the years and with the constant increments the group classification of lactobacilli was rearranged according to an expanded number of phylogenetic clades, e.g., 12 in 2007 (Felis and Dellaglio, 2007), 15 in 2012 (Salvetti et al., 2012), 18 in 2014 (Pot et al., 2014), and, most recently, 24 in 2017 (Duar et al., 2017).

True to their namesake, lactobacilli use lactic acid fermentation through one of two different metabolic pathways as the primary means for producing cellular energy (Kandler, 1983). For instance, those species falling under the definition of "obligate homolactic" are able to generate energy through the Embden-Meyerhof-Parnas (EMP) pathway, whereby one molecule of glucose is converted to two molecules of lactate. Alternatively, other species defined as either "facultative heterolactic" or "obligate heterolactic" will use the phosphoketolase (PK) pathway for metabolizing glucose and subsequently producing the end-products of lactate, ethanol, and CO_2. In terms of energy yield, the EMP pathway generates two molecules of ATP per metabolized glucose molecule, twice as much as produced by the PK pathway. Among the early classification schemes, this metabolic behavior served as the basis for organizing the *Lactobacillus* species and included the following three group categories: (1) obligately homofermentative (e.g., *L. delbrueckii*, *L. ruminis*, *L. salivarius*, and *L. acidophilus*), (2) facultatively heterofermentative (e.g., *L. plantarum*, *L. sakei*, *L. rhamnosus*, and *L. casei*), and (3) obligately heterofermentative (e.g., *L. reuteri*, *L. buchneri*, *L. fermentum*, and *L. brevis*). Though, in an effort to better reflect the energy metabolism traits of the different species within the overall molecular phylogeny of the *Lactobacillus* genus, a new two-group classification of "homofermentative" versus "heterofermentative" was instead proposed recently (Zheng et al., 2015). However, it is worth mentioning that whereas the majority of lactobacilli will generate their energy requirements fermentatively via substrate-level phosphorylation, there are certain species that possess the genetic means for an anaerobic respiratory metabolism (Brooijmans et al., 2009; Zotta et al., 2016).

Helped in part by a characteristic metabolic resourcefulness that promotes adaptability to various habitats and environments, lactobacilli are exploited for use in a range of man-made applications, some with a long history and others somewhat recent (Giraffa et al., 2010). For instance, a number of

Lactobacillus species are well-used as starter cultures or co-cultures for the production of fermented foods (e.g., cheeses, yogurts, sausages, fish, and vegetables) and beverages (e.g., wine and beer), sourdough bread, and feed silage (Giraffa et al., 2010). Other prospective uses of lactobacilli are via their natural antimicrobial bacteriocins for preserving and protecting foods (Leroy and De Vuyst, 2004), or via their exopolysaccharides for improving the firmness, texture, and taste of certain low-fat food products (Leroy and De Vuyst, 2004). From a scientific, clinical, and commercial perspective, the lactobacilli have also drawn much interest in the study of their therapeutic use as probiotics for maintaining good intestinal health and remedying certain gut-related ailments and problems (Di Cerbo et al., 2015). This probiosis also covers certain female infections, as a few *Lactobacillus* species are considered to be helpful in the treatment of bacterial vaginosis (Di Cerbo et al., 2015). In the context of both the gut and vagina, lactobacilli display good adhesion potential for colonizing host tissues, and thus this is seen as a key competitive feature for the displacement and removal of harmful pathogens (Yadav et al., 2015). Moreover, probiotic lactobacilli also demonstrate promising prospects for eliciting beneficial responses from the host immune system, which conceivably might help to maintain a state of physiologic well-being in the gut or elsewhere in the body (Hevia et al., 2015).

1.8 *Lactobacillus* genomics

Owing to the commercial and societal importance of lactobacilli, but also to a scholarly interest in their ecological history and origins, the recent years has seen a steady increase in the number of sequenced genomes for species belonging to the *Lactobacillus* genus. What follows in this section is a general overview of *Lactobacillus* genomics and afterwards, in the next two sections, a more focused description on the genomics of the *L. rhamnosus* and *L. ruminis* species.

Lactobacillus genomics really first took off in 2003 with the genome sequencing of the *L. plantarum* WCFS1 strain (Kleerebezem et al., 2003). As it were, this was merely the starting point and by 2018 (May 2) the number of sequenced *Lactobacillus* genomes in the NCBI database had swollen to more than 1600, out of which approximately 15% are complete sequences, with the rest being draft assemblies (**Figure 5**). Based on the breadth of related publications this has spawned (and continues to do so), there is little doubt that by determining these genome sequences, the data obtained has helped advance a greater scientific understanding of the lactobacilli group of bacteria and their various biological and ecological nuances. Given that the distribution of phenotypic characteristics among the various *Lactobacillus* species is rather

varied, this is also reflected at a genetic level, as there is a high level of diversity in their genomes, e.g., as observed with size, GC content, and number of CDSs (Salvetti and O'Toole, 2017).

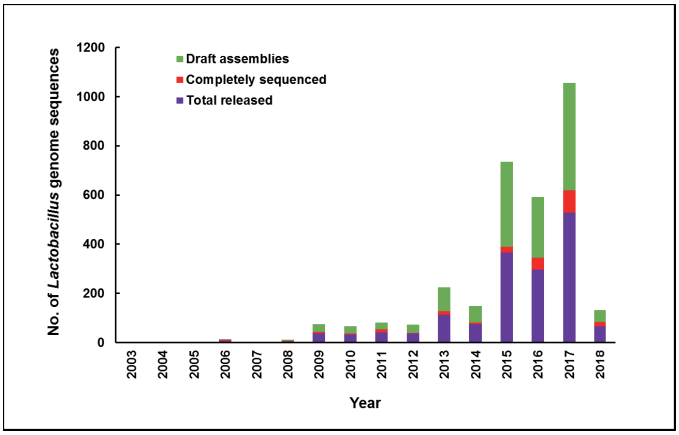

Figure 5. Deposition of *Lactobacillus* genome sequences in the NCBI database (2003-2018; May 2).

Given that early on the sequencing of genomes was costly and slow, the first genomic studies on lactobacilli mainly dealt with the characterization of a single representative species, such as, e.g., *L. plantarum*, *L. johnsonii,* and *L. acidophilus* (Kleerebezem et al., 2003; Pridmore et al., 2004; Altermann et al., 2005). Consequently, the first comparative analyses of *Lactobacillus* genomes were at an interspecies level (Boekhorst et al., 2004; Pridmore et al., 2004). What generally emerged from these types of comparisons is that the extent of any genomic similarity or differences will depend on the taxonomic closeness of the *Lactobacillus* species being analyzed. For instance, while there is an absence of synteny in the genomes from distantly related *Lactobacillus* species, the opposite is detected for taxonomically similar lactobacilli, where the genomes are overall well conserved (e.g., Berger et al., 2007; Ventura et al., 2008).

Among some of the approaches used to compare *Lactobacillus* genomes, these include: (1) mapping short genomic sequence reads against a designated reference genome (Douillard et al., 2013a), (2) using the comparative genomic hybridization technique with a reference genome (Siezen et al., 2010), and (3) comparing outright the sequences of whole genomes from several different strains (e.g., Broadbent et al., 2012; Smokvina et al., 2013; Spinler et al., 2014). On the whole, though the comparisons using a reference genome tended to provide only a partial assessment of the genomic disparity between strains, this method is still useful for revealing which of the genetic attributes are core-conserved among different genomes. On the other hand, the whole-genome comparisons provide much more insight and can identify a broader collection of

similarities and differences, either when done with several different *Lactobacillus* species (e.g., Canchaya et al., 2006; Claesson et al., 2008; Sun et al., 2015) or with just one particular species (e.g., Nelson et al., 2010; Broadbent et al., 2012; Smokvina et al., 2013; El Kafsi et al., 2014; Spinler et al., 2014).

As would be expected, several pan-genomic appraisals of the *Lactobacillus* genus have been undertaken and, depending on availability and quality, a varying number of species whole-genome sequences were used for such studies (e.g., Canchaya et al., 2006; Claesson et al., 2008; Lukjancenko et al., 2012; Sun et al., 2015). In one of the largest samplings, a study encompassing 213 *Lactobacillus* strains and associated genera revealed the corresponding pan-genome consists of 44,668 gene families, but with only 73 genes representing the core genome and of these a disproportionate share encoding for proteins that are responsible for cell growth and replication (Sun et al., 2015). Somewhat mirroring this, a pan-genomic study with a considerably reduced sampling size (i.e., 12 complete genomes from 11 different *Lactobacillus* species), but with a very strict criteria for orthologue detection, had reported obtaining a relatively small core genome of only 141 genes. Again, most of these core genes were assigned to housekeeping functions, predominantly with roles in nucleotide transport and metabolism, cell-wall biosynthesis, and post-translational modification (Claesson et al., 2008). Still, some other pan-genome studies have revealed a higher number of core genes, i.e., 363 (Lukjancenko et al., 2012) and 593 (Canchaya et al., 2006). Significantly, these comparative analyses of the lactobacilli genomes have shown that amongst the predicted core genes only a small proportion are considered specific to the genus *Lactobacillus*. However, in overall terms, many of these pan-genomic investigations have revealed important genetic and functional details about the lactobacilli bacteria and their ecological lifestyle, and, in that way, given greater knowledge regarding the evolution, adaptability, diversity, and industrial use of various *Lactobacillus* species.

1.9 *Lactobacillus rhamnosus* genomics

L. rhamnosus is taxonomically close to *L. casei* and *L. paracasei*, and together these three species form what is called the "casei group" of lactobacilli (Salvetti et al., 2012). As a group, these species are considered homogeneous, e.g., since all are facultative heterofermentative and their GC content is around 45-47% (Salvetti et al., 2012). Ecologically, *L. rhamnosus* is a pervasive species, with various strains adapting to several habitats in the body, such as the digestive and respiratory tracts, mouth, vaginal lining, and lactating mammary glands, but also on occasion transiently colonizing blood and infected tissue (Ahrné et al., 2005; Martin et al., 2007; Vancanneyt et al., 2006). Moreover, *L. rhamnosus* is also associated with fermented cheeses and yogurts (Bernardeau et al., 2008) and includes a spoilage role

in beer (Haakensen et al., 2009). As is much the case with other *Lactobacillus* species, certain strains of *L. rhamnosus* are observed to have health-benefiting properties, and thus they have come to be promoted heavily for probiotic use in fermented dairy products or as dietary supplements. Yet it is also the case that some other *L. rhamnosus* strains are used industrially as adjunct starter cultures.

Much of the impetus for unraveling the genomics of the *L. rhamnosus* species originates from a commercial interest in the molecular mechanisms behind the probiosis of *L. rhamnosus* GG (ATCC 53103), a human gut-adapted strain known for having many advocated health benefits and the worldwide marketing moniker of LGG® (for review, see Pace et., 2015). Thus, for the first detailed study of *L. rhamnosus* genomics, this involved a comparative analysis of the genomes from *L. rhamnosus* GG and *L. rhamnosus* LC705, a dairy starter culture strain (Kankainen et al., 2009). From the genomic comparison of these two strains, it appeared that their genomes are closely similar, e.g., as in size (~3 Mbp), number of encoded genes (2,944 in GG and 2,992 in LC705), and GC-content (47% each). Moreover, both genomes display a comparable number of rRNA operons, tRNA genes, and prophage clusters. However, there are also some apparent differences in the two genomes, as the number of transposases vary (69 in GG and 29 in LC705), and as did the occurrence of plasmids (one in LC705) and CRISPR loci (one in GG). Further, while the synteny is well conserved between the two genomes, it is noticeably interspersed by DNA sequence that differs from the overall genome (as in nucleotide makeup, codon usage, and dinucleotide occurrence) and this was taken to represent genomic islands, five and four in GG and LC705, respectively. Also, a comparative analysis of the 3000 or so predicted proteins encoded by the genomes had shown that on average there is a high level of amino acid identity (98%). A further examination of the gene-encoded products revealed the number of strain-specific proteins is slightly higher for the genome of LC705 than that of GG (383 vs. 331). With respect to those predicted proteins with no counterpart in the genomes of other *Lactobacillus* species, these amounted to 143 in GG and 176 in LC705, with a good proportion (17 and 12%, respectively) being assigned to carbohydrate metabolism and transport functions.

However, one of the most exciting outcomes from the genomic comparison of the *L. rhamnosus* GG and LC705 strains was the revelation that both genomes encode the genes for sortase-dependent piliation (Kankainen et al., 2009). Up until then, these long and limb-like surface protrusions were only known to be present amongst Gram-positive pathogens (e.g., certain species of *Streptococcus*, *Corynebacterium*, and *Enterococcus*) and thus quickly regarded as a key virulence factor of such harmful bacteria (for review, see Danne and Dramsi, 2012; Proft and Baker, 2009). Structurally, the sortase-dependent pilus has a distinctive composition and architecture, being made up of two or three types of protein subunits (called pilins), each with its individual location and function. For pilus assembly, the pilin subunits are covalently coupled together via the transpeptidase action of the pilus-specific C-type sortase enzyme, with the polymerized form eventually attached to the

cell wall by the housekeeping A-type sortase. In the genome, the genes for the sortase-dependent pilus are always grouped together in an island or operon and will encode for both the pilin proteins (found at the pilus tip and/or base and comprising the pilus backbone) and the C-type sortase enzyme.

Based on the genomic comparison of the two *L. rhamnosus* strains, each of them contains the genes for the so-called *spaFED* pilus operon (i.e., *spaF-spaE-spaD-srtC2*), which encodes the tip SpaF, basal SpaE, and backbone SpaD pilin subunits, and along with the SrtC2 C-type sortase (Kankainen et al., 2009). Yet it is only the genome of the GG strain that was found to have the genes for an additional pilus operon, known as *spaCBA* (i.e., *spaC-spaB-spaA-srtC1*), and like the *spaFED* operon that also encodes for tip, basal, and backbone pilins (called SpaC, SpaB, and SpaA, respectively) and a C-type sortase (SrtC1) (Kankainen et al., 2009). In all cases, it was shown that the predicted primary structure for each of the SpaCBA and SpaFED pilin subunits displays the distinguishing canonical sequence motifs and domains that are found in a conventional Gram-positive pilin-protein. However, further experimentation established that of the two pilus operons it was only the *spaCBA* loci that are constitutively active, and thus which leads to the native production of fully assembled SpaCBA pili on the surface of GG cells (Kankainen et al., 2009). This finding confirmed the results of a prior study that observed pilus-like formations at the cell poles of an extracellular polysaccharide-lacking mutant of the GG strain (Lebeer et al., 2009). On the other hand, those genes associated with the *spaFED* operon appeared to be inactive in the GG and LC705 strains of *L. rhamnosus*, or at least under the testing conditions, but otherwise were expressible in a recombinant form using *Lactococcus lactis* as an alternative host (Rintahaka et al., 2014). Additional characterization of the SpaCBA pilus revealed that it can adhere to human intestinal mucus, with the SpaC tip pilin being the main binding determinant (Kankainen et al., 2009). This finding clearly explained why the GG strain is a comparatively strong and effective binder of mucus, but as well, why this transient or allochthonous strain seems to have a somewhat prolonged stay in the human gut. As a wider outcome, the use of comparative genomics for revealing this piliated strain of *L. rhamnosus* brought in an alternative way of thinking about sortase-dependent piliation, meaning that it no longer just represents a virulence factor, but instead can be seen as also a niche-adaptation factor. Consequently, as the *spaCBA*-encoded pilus was thought to represent a new and potentially important mechanism behind the intestinal microecology and probiosis of the *L. rhamnosus* GG strain, the ensuing years have led to many studies aimed at characterizing its molecular and biological function (for review, see von Ossowski, 2017).

Continuing with *L. rhamnosus* genomics, what soon followed were some additional studies that offered a further comparative examination of the GG strain genome, but as well, other representative genomes from this and related species. For instance, a comparative analysis study between the genomes of *L. rhamnosus* GG and *L. casei* BL23 (along with for each two additional genomes of strains isolated from

probiotic products) had revealed their sizes are comparable at ~3 Mbp and none are accompanied by plasmids (Douillard et al., 2013b). As for the latter point, it should be noted that the BL23 strain is derived from *L. casei* ATCC 393 (a dairy isolate) after it was cured of its endogenous plasmid pLZ15 (Mazé et al., 2010). In the one-to-one comparison of genomes from the GG and BL23 strains, the conserved synteny is high between the two, with the only observed perturbations being primarily from genomic islands containing genes for transposases, prophages, and carbohydrate transport and metabolism (Douillard et al., 2013b). This "mobility" aspect would seem to further highlight lateral gene transfer as a major evolutionary force and thus a potentially significant source of genetic diversity among these bacteria (Douillard et al., 2013b). As far as any mutual or species-specific genes between the two strains, these numbers summed up to 2,180 (GG and BL23), 836 (GG), and 835 (BL23), but of some interest was the shared presence of genes encoding the *spaCBA* pilus operon (Douillard et al., 2013b). However, in this regard, a marked difference was found with the *spaCBA* genes in the GG strain, as these occur within a region containing the sequences for transposable elements, raising the possibility that this pilus operon was acquired through the lateral transfer of genes. Further, it was also found that only the *spaCBA* pilus operon of the GG strain is preceded upstream by a potential regulatory region, and whose origins in the genome might have been as an iso-IS30 element (Douillard et al., 2013b). Thus, as the presence of this putative controlling element likely represents the reason why *L. rhamnosus* GG exhibits constitutive production of SpaCBA pili, its absence from the genome of *L. casei* BL23 probably explains why this particular strain has an inactive *spaCBA* operon and is non-piliated. Other notable differences observed between the genomes of these two strains lie with the genes for carbohydrate metabolism (Douillard et al., 2013b). For example, while both strains have the genetic machinery to transport and metabolize maltose, the continuity of the maltose gene cluster is interrupted by an additional ORF in the genome of *L. rhamnosus* GG, and this in turn explains the inability of this strain to use maltose as an energy source. In contrast, the *L. casei* BL23 strain has an intact and undisrupted set of maltose genes, and thus is able to subsist on maltose (Douillard et al., 2013b). Another example involves the ability to utilize the hexose sugar fucose, which is part of the glycan structure of the mucin proteins that make up the epithelial mucus lining in the intestine, or then elsewhere within the body. While the gut-adapted *L. rhamnosus* GG strain can metabolize fucose (Becerra et al., 2015), it is not the case for *L. casei* BL23. Predictably, this particular difference between the two strains is reflected at the genomic level, with a cluster of fucose-related genes being found present in the *L. rhamnosus* GG genome, but which are missing for *L. casei* BL23 (Douillard et al., 2013b).

In a more expansive study of *L. rhamnosus* genomics, a comparative analysis was performed on the genomes of 100 strains that originated from numerous habitats and sources (Douillard et al., 2013a). Interestingly, this genomic comparison of the *L. rhamnosus* species is sometimes mistakenly referred to as a pan-genome study (e.g., Espino et al., 2014; Cavanagh et al., 2015; Chun et al., 2017; Duar et al., 2017), but in

fact it involved mapping the sequence reads of the various strains (99 in total) onto the *L. rhamnosus* GG reference genome. Additionally, a few of the strains had their genomic sequence reads mapped against the reference genome of *L. rhamnosus* LC705. Yet, because this was a study of mapped sequence reads, a complete analysis of the genetic diversity in the *L. rhamnosus* species is limited to that found in the GG strain. Nonetheless, an orthologous "core set" of 2,419 genes (covering ~80% of the GG genome) was defined for the large sampling of *L. rhamnosus* strains, with the number of core genes staying relatively constant irrespective of how many genomes (20 or more) were used in the calculations (Douillard et al., 2013a). However, what significantly emerged from this comparison of *L. rhamnosus* genomes was the presence of two recognizable groupings drawn along the lines of geno-phenotypic traits and properties (Douillard et al., 2013a). For example, among those strains belonging to the group "A" category, it was inferred from their genomes that they have a genetically adaptive predisposition for nutrient-rich environments. This was well exemplified by the carbohydrate metabolism of these strains, as members of this group are able to metabolize lactose, a disaccharide sugar commonly found in milk (Douillard et al., 2013a). Since many of the strains in group A are derived from cheese products, the ability to utilize a milk-carbohydrate would be consistent with an adaptation to a dairy ecological niche. Moreover, it was also found that among these dairy-derived strains only a few possess active *spaCBA* pilus genes (13%), which suggests that mucus-binding pili are not needed for providing an ecological advantage or fitness benefit to *L. rhamnosus* cells in the prevailing environment (Douillard et al., 2013a). Other *L. rhamnosus* strains in group A are isolated from the mouth and vagina, but oddly enough none had the presence of a *spaCBA* pilus operon in their genomes, yet despite residing in a mucus-lined environment.

On the other hand, those *L. rhamnosus* strains (GG included) falling under the group "B" categorization maintain a genetic bias for adaptation to the human body, such as in the gut, or then transiently in blood and infected tissue specimens (Douillard et al., 2013a). Specifically for the intestinal strains, phenotypic characteristics like mucus-binding piliation, mucin-fucose utilization, and bile resistance were held in common, and thus these would offer the competitive and adaptive edge for both surviving and colonizing within the gut environment (Douillard et al., 2013a). At a genomic level, while more than half (~56%) of gut-adapted *L. rhamnosus* strains contain an intact *spaCBA* pilus operon, many still did not (Douillard et al., 2013a). Moreover, whereas intact loci for fucose transport and metabolism are a common attribute of the gut isolates, they are less prevalent among the dairy strains of group A and thus many cannot metabolize fucose (Douillard et al., 2013a). This appears logical since as an energy source, fucose is less plentiful in milk (if at all present) and encoding the related genes would provide no advantageous fitness benefit to cells.

In a recent pan-genomic study post-dating the work in Study II, 40 strains of *L. rhamnosus* (primarily from human-related and dairy sources) were used for in-

depth genomic comparisons that focused mainly on characterizing the variable genetic makeup of this species (Ceapa et al., 2016). The predicted size of the *L. rhamnosus* pan-genome was estimated at 4,711 genes, and from this amount there are 2,164 genes comprising the core genome, with remainder making up the accessory genome. Among the core genes, these are less in number than what was observed in the genomic mapping study of *L. rhamnosus* (Douillard et al., 2013a), although expectedly they encode for the basic housekeeping functions typically needed for maintaining cell viability. However, in the case of the 2,547 accessory genes, these were found to be often associated with genetic rearrangement and lateral gene transfer, e.g., transposons, phages, and plasmids. Moreover, these loci also encoded a range of cellular functions, such as those involved with bacteriocin and pilus production, extracellular polysaccharide biosynthesis, carbohydrate transport and metabolism, CRISPR-Cas (CRISPR-associated) systems, and a variety of membrane transporter proteins (Ceapa et al., 2016). Based on the variability of these genetic traits in the different strains, the *L. rhamnosus* variome content, along with the capacity for gene movement, would be in keeping with the environmental adaptability of this species and its occupancy of diverse ecological niches (Ceapa et al., 2016).

As mentioned beforehand, in addition to benignly inhabiting various regions of the human body, certain strains of *L. rhamnosus* are found to be associated with infected tissue, e.g., by being present at the early phase of infection in the dental pulps of carious teeth (Nadkarni et al., 2014). In a comparative analysis of genomes from such strains, one published study tried to pinpoint whether the invasion of tooth pulp tissue by *L. rhamnosus* is dependent on a uniquely different genotype (Nadkarni et al., 2014). For this, a genomic comparison between two dental pulp isolates of *L. rhamnosus* (i.e., LRHMDP2 and LRHMDP3), along with *L. rhamnosus* GG as a reference strain, had revealed several genetic anomalies that could conceivably be taken as the invasive biomarkers for bacterial tooth infection (Nadkarni et al., 2014). Regarding the LRHMDP2 and LRHMDP3 strains, both their genomes were found to encode for a cell surface morphology that differs from *L. rhamnosus* GG, and this was promoted as the possible mechanism that allows *L. rhamnosus* to invade dental pulps (Nadkarni et al., 2014). Included among the key genomic differences were the presence of genes for a modified exopolysaccharide layer and MabA-like protein, as well as a collagen-binding protein domain with a unique repeat sequence, but the absence of the genes for the *spaCBA* pilus operon (Nadkarni et al., 2014).

The analysis of *L. rhamnosus* genomes was also part of broader investigations into the genomics of the casei group lactobacilli. For instance, in once such study, a pan-genomic comparison was performed with four strains from the casei group, i.e., *L. rhamnosus* ATCC 53103 (GG), *L. casei* ATCC 393, *L. paracasei* JCM 8130, and *L. paracasei* ATCC 334, and this revealed that among the 4,315 genes of the predicted pan-genome there are 1,793 shared genes (Toh et al., 2013). When an additional six strains (*L. paracasei* BDII, *L. paracasei* BL23, *L. paracasei* LC2W, *L. paracasei* Zhang, *L. rhamnosus* LC 705, and *L. rhamnosus* ATCC 8530) were

included in the genomic assessment, 94% of the core genes remained in common (Toh et al., 2013). Here, it was proposed that this number of genes (1,682) approximates the size of the core genome for the casei group, and most probably these would have shared the same ancestral source. Moreover, it was also found that among the casei group genomes there is pervasive synteny, which suggests this particular clade of lactobacilli has maintained a high level of genomic stability (Toh et al., 2013). However, the presence of a varied number of genomic islands (gene clusters) was interspersed throughout these different genomes, some of which were found in the same chromosomal location (Toh et al., 2013). Among the various gene clusters were those for carbohydrate metabolism, and in the case of the cheese-isolated *L. paracasei* ATCC 334 strain many of the genes for utilizing carbohydrates were missing, possibly from evolutionary decay due to habitation in dairy environs. Another shared gene cluster of the casei group lactobacilli was for SpaCBA piliation (first seen with the *L. rhamnosus* species), though it seemed that for certain *L. paracasei* strains the gene for the SpaC tip pilin had been subjected to truncation, and thus was only partially intact (Toh et al., 2013).

Somewhat along the same lines, another quite recent study undertook a comparative analysis of 184 genomes as a more thorough approach to confirm the taxonomic positioning of the lactobacilli within the casei group (Wuyts et al., 2017). Genomes used for the genomic comparison were divided among 92 *L. rhamnosus*, 37 *L. casei*, 38 *L. paracasei*, 15 unclassified *Lactobacillus* species, and two *L. zeae*. By considering the genetic nuances in GC content, molecular phylogeny, and pairwise genomic relatedness, the genomes could be grouped into three genetically distinct clades (Wuyts et al., 2017). As inferred from a core-gene phylogenetic reconstruction, clade A holds the genomes from most of the *L. casei* strains, all *L. paracasei*, and one unclassified *Lactobacillus* species, whereas for clade B there are the genomes from six *L. casei*, two unclassified *Lactobacillus* species, and both *L. zeae*. Clade C is the largest in size, containing the genomes from all of the *L. rhamnosus* strains and the 12 remaining unclassified *Lactobacillus* species. Interestingly, among the three clades, only the strains of *L. rhamnosus* had formed a monophyletic taxon, with the implication they descend from a common ancestor. Overall, it was suggested that based on the occurrence of type strains, the clades A, B, and C should be designated representative of the species *L. paracasei*, *L. casei*, and *L. rhamnosus*, respectively (Wuyts et al., 2017). In addition, for this large set of genomes the size of the total gene pool was estimated as 521,567 genes, averaging out at about 2,828 genes per each genome (Wuyts et al., 2017). This gene pool was further categorized into orthogroups (i.e., a collection of genes derived from one gene in the latest shared ancestor of all species being examined) and amounted to 5,915, out of which 1,814 and 4,101 were designated as core and accessory, respectively (Wuyts et al., 2017). By comparison, the size of this core orthogroup is just slightly larger than of the core genome in the small sampling pan-genome study of the casei group lactobacilli (Toh et al., 2013). Incidentally, the number of core orthogroups for the clade (C) encompassing a majority *L. rhamnosus* strains was 2,133, which is quite close to the

core genome size of 2,164 genes calculated in the pan-genome study of *L. rhamnosus* (Ceapa et al., 2016), but falling short of the 2,419 cores genes predicted in the *L. rhamnosus* mapping study (Douillard et al., 2013a).

1.10 *Lactobacillus ruminis* genomics

At a taxonomic level, *L. ruminis* is found in the same phylogenetic clade containing 24 other species (i.e., *L. acidipiscis*, *L. agilis*, *L. animalis*, *L. apodemi*, *L. aquaticus*, *L. aviarius*, *L. cacaonum*, *L. capillatus*, *L. ceti*, *L. equi*, *L. ghanensis*, *L. hayakitensis*, *L. hordei*, *L. mali*, *L. murinus*, *L. nagelii*, *L. oeni*, *L. pobuzihi*, *L. ruminis*, *L. saerimneri*, *L. salivarius*, *L. sucicola*, *L. satsumensis*, *L. uvarum*, and *L. vini*), and collectively these are called the "salivarius group" (Salvetti et al., 2012). This group encompasses both homolactic and heterolactic fermenters, with *L. ruminis* falling into the former category as it possesses an obligately homofermentative metabolism (Salvetti et al., 2012). As another mark of diversity, the GC content of the salivarius group lactobacilli ranges widely between 32-44% (Salvetti et al., 2012). Conspicuously, this grouping of lactobacilli is noted for containing the greatest number of the *Lactobacillus* species exhibiting the motility trait (Salvetti et al., 2012), which in fact also includes *L. ruminis* (Forde et al., 2011; Yu et al., 2017).

In regard to *L. ruminis* specifically, this species exists as a strict anaerobe and has a certain notoriety for being among the few autochthonic lactobacilli residing in the digestive tract of humans and animals, meaning that it is an indigenous member of the gut microbiota (Reuter, 2001; O'Callaghan and O'Toole, 2013). For the latter property, *L. ruminis* has seemingly evolved a genetic makeup that allows it to subsist on the luminal nutrients of the intestine, this exemplified by its capacity to metabolize carbohydrous material (O'Donnell et al., 2011; O'Donnell et al., 2015). *L. ruminis* has also acquired the genotype for the piliation trait (Forde et al., 2011), as its genome encodes a sortase-dependent pilus operon (i.e., *lrpC-lrpB-lrpA-srtC*), which has been proven to be active and expressed as fully assembled pilus structures consisting of the LrpC tip, LrpB basal, and LrpA backbone pilins (Yu et al., 2015; Yu et al., 2017). These pili (called LrpCBA) have been characterized with the ability to bind extracellular matrix (ECM) proteins (Yu et al., 2015), and this can be construed as a key adaptive trait of *L. ruminis* for its localization in the anoxic and deeply folded epithelium of the intestine (for review, see von Ossowski, 2017). Presumably, many of these genetic traits would have combined phenotypically to give *L. ruminis* its gut-autochthonal behavior.

Despite being identified way back in 1961 (Lerche and Reuter, 1961), and as well having a range of remarkable properties, the *L. ruminis* species was largely neglected as a topic for genomics research until just a half-dozen years ago. Because *L. ruminis* was recently judged to be an immuno-stimulative (Taweechotipatr et al., 2009) and pathogen-displacing (Yun et al., 2005) bacterium, and then along with its

unique traits for motility, piliation, and autochthonous growth, this species suddenly became ecologically interesting to study, but as well, conceivably health-benefiting as a possible probiotic. Spurred on by this, the first genomic comparison of *L. ruminis* was published in 2011, a study whose aims were to expand the number of characterized genomes in the salivarius group as well as provide the necessary genome data for a phenotypic examination of the lactobacillar motility trait (Forde et al., 2011). Here, the fully sequenced genome from a bovine isolate of *L. ruminis* (ATCC 27782) was compared to the draft genome assembly of the human ATCC 25644 strain. This comparative analysis revealed that the two genomes share a highly conserved synteny, which is only interrupted by a large inverted genomic segment near the replication terminus (Forde et al., 2011). Then again, the genomic comparison of *L. ruminis* ATCC 27782 with the *L. salivarius* UCC118 strain showed that even though their genetic content supports a phylogenetic closeness, the synteny between the two genomes is much less substantial, further exemplifying the relatively high diversity within the *Lactobacillus* genus (Forde et al., 2011). Nonetheless, the similarity between the genomes of the two *L. ruminis* strains extends to their size (~2 Mbp), GC content (~44%), and rRNA operons (six each), although some differences are apparent, e.g., as in the number of protein-encoding genes (1,901 in ATCC 27782 and 2,251 in ATCC 25644) and tRNAs (67 in ATCC 27782 and 49 in ATCC 25644) (Forde et al., 2011). Other similarities between the two genomes involve the genetics for various cellular attributes and functions. Some notable examples include the genes for flagellar biogenesis and chemotaxis, sortase-dependent piliation, CRISPR-Cas proteins, and bacteriocin production (Forde et al., 2011). Further, both *L. ruminis* genomes contain the genes for the proteins involved in several different pathways for utilizing carbohydrates, such as fructose, glucose, galactose, mannose, starch, and sucrose (Forde et al., 2011). Thus, at the genome level, the *L. ruminis* species has acquired a set of appropriate geno-phenotypic traits for adapting to the intestinal microcosm and its competitive menagerie of bacteria.

While the above mentioned genomic comparison of *L. ruminis* involved the genomes of human and bovine isolates, a recent study published in 2015 went a step further to perform a comparative analysis that included the genomes from porcine and equine strains (O'Donnell et al., 2015). In total, this comparison involved six *L. ruminis* genomes that were derived from two human strains (ATCC 25644 and S23), two bovine strains (ATCC 27782 and ATCC 27780), one porcine strain (DPC 6830), and one equine strain (DPC 6832) (O'Donnell et al., 2015). Based on the genome comparison of these strains, the major genomic features (e.g., size, GC-content, and CDSs) were comparable to those observed previously (Forde et al., 2011), although with some slight variances. Moreover, while a graphical comparison of the six genomes showed they mainly share similar sequence (99%), there was also the intermingling of gaps and variable regions that mostly stemmed from genes encoding CRISPR, restriction-modification, phage-associated, and hypothetical proteins (O'Donnell et al., 2015). Since part of the aim of this study was to assess whether the similarities in the *L. ruminis* strains correlate with host-gut sources, a phylogenetic

reconstruction was performed using a dataset of 907 core orthologous genes identified from the six genomes (O'Donnell et al., 2015). From the reconstructed phylogeny, the four genomes from the human and bovine *L. ruminis* strains were grouped into individual clades according to their isolation host, although together they would form a larger clade that was separate from the paired up genomes of the porcine and equine isolates (O'Donnell et al., 2015). In terms of *L. ruminis* genomics, this closer phylogenetic relatedness between the human and bovine strains is suggestive of a recent and shared ancestral origin. On the other hand, though the genomes from the porcine and equine strains are phylogenetically similar to each other, the genetic lineage of these two isolates appears more distant from the other four *L. ruminis* strains (O'Donnell et al., 2015).

As already introduced before, the *L. ruminis* species is autochthonously specialized to the human and animal gut via a number of genetic influences, one of which is an adaptive ability to utilize the carbohydrous luminal contents of the intestine (Forde et al., 2011). To expand on this further, one study undertook to examine the growth pattern of nine different *L. ruminis* strains on a wide variety of simple and complex carbohydrate substrates, and then try to establish a correlation with the metabolic genes and pathways annotated from the genomes of the bovine ATCC 27782 and human ATCC 25644 isolates (O'Donnell et al., 2011). From the metabolism profiles of 50 carbohydrates, the fermentative capacities amongst the nine strains had varied widely, but all strains retained some capacity to utilize prebiotic substrates (defined as host-indigestible polysaccharides that help stimulate the growth of gut-friendly bacteria or probiotics). At the genomic level, the ATCC 27782 and ATCC 25644 strains each encoded for 16 carbohydrate metabolism pathways (both complete and partial), which allowed for glycolysis, pentose-glucuronate interconversion, and utilization of fructose, mannose, sucrose, and starch (O'Donnell et al., 2011). Moreover, at least 10 and 14 carbohydrate transporters were predicted in the two genomes (ATCC 27782 and ATCC 25644, respectively), and these belonged to various transporter families, i.e., ATP-binding cassette (ABC), glycoside-pentoside-hexuronide (GPH) cation symporter, oligosaccharide:H^+ symporter (OHS), and phosphotransferase system (PTS) (O'Donnell et al., 2011). In addition, putative operons for utilizing prebiotics were identified in both genomes (three in ATCC 27782 and six in ATCC 25644) (O'Donnell et al., 2011). For instance, among those predicted only for the ATCC 25644 strain was a fructooligosaccharide utilization operon. As glycosyl hydrolases play an important role in metabolizing (and synthesizing) prebiotic carbohydrates these were also identified in both genomes (14 in ATCC 27782 and 20 in ATCC 25644), and including β-fructofuranosidase, a key enzyme required for the metabolism of fructooligosaccharides (O'Donnell et al., 2011). From this geno-phenotypic assessment, it would appear that *L. ruminis* has the capacity to exploit various prebiotics as an energy source. Since *L. ruminis* is regarded as a potential probiotic candidate (Taweechotipatr et al., 2009; Yun et al., 2005; Yu et al., 2017), this autochthonic species could be coaxed into greater numbers

by such substrates and, as a conceivable outcome, this might further reinforce any health benefits to a given host already harboring this gut bacterium.

Investigations into *L. ruminis* genomics were shifted beyond divulging genetic potential and towards revealing actual biochemical responses when one study used a transcriptomics approach (RNA-seq) to examine how this species is able to survive the competition for nutritional resources among various members of the gut microbiota (Lawley et al., 2013). As background, an earlier study showed that *L. ruminis* lacked the ability to utilize plant β -glucans (i.e., glucose polymer polysaccharides typically found in the cell-wall endosperm of cereal crops, such as wheat, barley, oat, and rye), though it was also observed that certain strains could grow on the tetrasaccharide components of hydrolyzed β-glucans (Snart et al., 2006). Based on this, it was reasoned that a natural source of this fermentable substrate is available as the released byproduct of other gut bacteria (Lawley et al., 2013). Among the tetrasaccharide utilizing *L. ruminis* strains analyzed transcriptomically in the study (Lawley et al., 2013), the human-derived L5 strain was shown to have elevated gene expression for the cellobiose utilization and chemotactic motility operons. From a phenotypic perspective, it was observed that *L. ruminis* cells were mainly flagellated when growing on tetrasaccharide and their movement was in the direction of this substrate (Lawley et al., 2013). As well, it was further revealed that the L5 strain could utilize the tetrasaccharide byproducts from the β-glucan degrading *Coprococcus* species, another intestinal commensal (Lawley et al., 2013). Taken together, it was concluded that *L. ruminis* manages to genetically adapt itself to the competitive surroundings of the gut environment, as in this case, by metabolizing the tetrasaccharide remnants made available when other bacteria degrade plant β-glucans (Lawley et al., 2013). Moreover, even though the amount of produced tetrasaccharide is likely limiting, this is offset by the active expression of motility and chemotaxis genes in *L. ruminis*, as this would allow the sensory movement for tracking down the location of this particular nutrient (Lawley et al., 2013). In all likelihood, such adaptive traits can help bolster the autochthonic character of the gut-dwelling *L. ruminis* species.

2. Aims of the study

The overall objective of this doctoral thesis was to study the genetic makeup of the *Lactobacillus* genus by using a pan-genomic analysis approach. Ultimately, the work for this study was divided into three separate sections (Study I, II, and III). Study I aimed to use the genomes from a number of *Lactobacillus* species and provide an overview of their genomic organization that then allows classification into different subgroups. Study II focused on the *Lactobacillus rhamnosus* species, wherein the aim was to reveal the genomic basis for a number of functionally relevant surface-associated proteins. Study III

examined the genetics behind the gut autochthony of the *Lactobacillus ruminis* species, with an emphasis on those genes and proteins associated with the cellular surface morphology and the anaerobic fermentative and respiratory processes.

3. Materials and methods

Publicly available scripts, databases, web servers, and stand-alone software platforms (some also complemented by in-house scripts) were used to perform a variety of genomic and bioinformatic analyses in studies I, II, and III (see **Table 2**). Detailed descriptions of these tools and other related methods are found in the three articles presented in this thesis (see appendices I, II, and III).

Tools and resources	Study	Reference
Databases		
CDD	I, III	Marchler-Bauer et al. (2015)
COG	I, III	Tatusov et al.(1997)
GenBank	I, II, III	Benson et al. (2008)
NCBI RefSeq	I, II, III	Pruitt et al. (2005)
Web servers		
BLASTP	I, II, III	Altschul et al. (1997)
SignalP v3.0	I	Emanuelsson et al. (2007)
SignalP 4.1	II, III	Petersen et al. (2011)
SecretomeP 2.0	III	Bendtsen et al. (2005a)
TatP 1.0	III	Bendtsen et al. (2005b)
TMHMM 2.0	III	Krogh et al. (2001)
PRED-LIPO	III	Bagos et al. (2008)
CW-PRED	III	Fimereli et al. (2012)
Stand-alone software platforms		
EDGAR	I, II, III	Blom et al. (2009)
LocateP	I	Zhou et al. (2008)
MUSCLE	I, II, III	Edgar (2004)
GBLOCKS	I, II, III	Talavera & Castresana (2007)
PHYLIP	I, II, III	Felsenstein (2005)
RPS-BLAST	I, III	Marchler-Bauer et al. (2002)
HMMER	I	Eddy (1998)
R (programming language)	II, III	none available
Sequencing, assembly, and related		
Score Ratio Value	I, II, III	Lerat et al. (2003)
454 GS FLX System	II, III	Margulies et al. (2005)
GS Assembler	II, III	none available

Table 2. Bioinformatic tools and resources used in the thesis study and their references.

4. Results and discussion

What follows in the next sections is a summary of the results and discussion from the three published articles supporting this thesis. A more detailed description of the results and discussion can be found in each of the articles, which are included as the appendices I, II, and III.

4.1 Study I: Pan-genomics of lactobacilli

Research done for study I was published in Microbial Biotechnology as the peer-reviewed article entitled "Comparative genomics of *Lactobacillus*" (see **Appendix I**) and part of a special 2011 issue on lactic acid bacteria. As mentioned in the literature review section, *Lactobacillus* genomics had emerged itself as an interesting field of study for research during 2003, and thus it was just about eight years later that study I sought to exploit the increased number and database-availability of completed *Lactobacillus* genome sequences for comparative analysis and investigation. Specifically, a pan-genome was assembled using 20 fully sequenced *Lactobacillus* genomes encompassing 14 various types of species (i.e., *L. acidophilus*, *L. brevis*, *L. casei*, *L. crispatus*, *L. delbrueckii*, *L. fermentum*, *L. gasseri*, *L. helveticus*, *L. johnsonii*, *L. plantarum*, *L. reuteri*, *L. rhamnosus*, *L. sakei*, and *L. salivarius*), out of which a core-genome molecular phylogeny was generated and used to group the genomes accordingly.

Among the *Lactobacillus* genomes analyzed, some widely varying aspects were noted about their sequences (see **Table 1; Appendix I**), such as sizes ranging from ~1.8 to ~3.3 Mbps and a G+C content of between ~33% and ~51%. For the latter, the range in G+C content was roughly two-fold greater than that typically found within a distinct genus of bacteria (Fujisawa et al., 1992), and thus indicative that these 14 *Lactobacillus* species might not reflect a uniform taxonomic category. Of some further interest, the number of protein-encoding genes predicted for the 20 genomes showed considerable variation (i.e., 1721 to 3100 genes) and suggests that during their evolution they experienced significant gene loss or gain, which would be in line with the findings of another study, though that was based on a more narrow selection of *Lactobacillus* genomes (Makarova et al., 2006). To assess what fraction of the protein-encoding genes contain sequence for secretion signals, and thus representing the *Lactobacillus* secretome (see **Table 1; Appendix I**), predictions were made using both the SignalP (Emanuelsson et al., 2007) and LocateP (Zhou et al., 2008) programs, even though the output from the latter is reputed to be more accurate and detailed for Gram-positive bacteria. Nonetheless, by comparison, the SignalP

predictions of the 20 genomes provided the larger-sized secretome, yet for both programs the proportion of recognized genes encoding secretion signal sequences is subject to considerable variation across the different species. For instance, among the SignalP predicted genes, the largest proportion (>30%) was associated with the genome of the two *L. rhamnosus* strains (GG and LC705). Moreover, despite sharing similar sized genomes, the *L. johnsonii* NCC533 and *L. acidophilus* NCFM probiotic strains possess a larger SignalP secretome than dairy-derived *L. helveticus* or *L. delbrueckii*, which incidentally both have the least number of proteins exhibiting a C-terminal LPXTG-like sorting motif (see **Table 1; Appendix I**).

Concerning the *Lactobacillus* pan-genome, its size was estimated at about 14,000 protein-encoding genes. Further details on the annotated identities of the pan-genome genes are available online as Supporting Information at http://onlinelibrary.wiley.com/doi/10.1111/j.1751-7915.2010.00215.x/full (see **Table S1**). Deduced from this large gene pool, the proportion representing the core genome was a mere 383 genes (see **Table S2**), which left the vast majority of the *Lactobacillus* genetic repertoire part of the accessory genome. Nonetheless, the shared set of orthologous genes in the 20 genomes, designated as the *Lactobacillus* core genome (LCG), was more than 2.5-fold larger than that of another genomic comparative study involving only 12 *Lactobacillus* genomes. While in this study there were just 141 core genes identified, this lower number was attributed to gene selections based on stricter benchmarks and the use of a COG classification scheme (Claesson et al., 2008).

Interestingly, about a quarter of the core genes (100 in total) were arranged as operonic islands or clusters, and though this indicates a shared function, it also suggests that among the genes their organization and control is conserved. Further still, this genomic aspect would seem to indicate the *Lactobacillus* species shares the same evolutionary lineage, which itself might extend further to other descendants since several of the core genes have counterparts in related Gram-positive bacteria. Among the more notable core-gene clusters identified, these included functional gene annotations for ribosomal proteins, proton-translocating ATPases, several housekeeping proteins, lipotechoic acid d-alanylation protein, and as well the proteins for three two-component regulatory systems. Another identifiable genomic feature was the side-by-side presence of genes for carbon catabolite control protein and a prolidase, thus suggesting the metabolism of carbohydrates and nitrogen compounds shares a similar regulation mechanism. Intriguingly, upwards of 20% of the LCG consists of genes (80 in total) with hypothetical annotations. However, one of the more interesting single genes with an annotated putative function was for the fibronectin-binding protein FbpA, and although this protein lacked secretion signals and anchoring domains at its N- and C-terminus,

respectively, and thus might only be loosely bound to the outer cellular surface, there remains the possibility that it has a role in lactobacillar biofilm formation. Additional functional assignments of the LCG were done using COG classifications (see **Figure 1; Appendix I**), and here it was surprising that, despite the predictions for a large *Lactobacillus* secretome size (**Table 1; Appendix I**), COG predictions indicated that only 5% of the core genes encoded secretable proteins, and thus most of the secreted proteins would likely be strain-specific.

As a means to group similar *Lactobacillus* genomes, a molecular phylogeny was reconstructed using the 383 core genes (see **Figure 2; Appendix I**), and while there were some discrepancies with the more recognized 16S rRNA-based phylogenetic reconstructions, the confidence level was higher since the comparisons were whole-genome based. Nonetheless, the 20 different genomes could be gathered into three distinctive clades, which were designated according to the genomic presence of the most renowned strain of *Lactobacillus*, i.e., NCFM (i.e., *L. acidophilus* NCFM, *L. crispatus* ST1, *L. delbrueckii* ATCC 11842, *L. delbrueckii* BAA-365, *L. gasseri* ATCC 33323, *L. helveticus* DPC 4571, *L. johnsonii* FI9785, and *L. johnsonii* NCC533), WCFS (i.e., *L. brevis* ATCC 367, *L. fermentum* IFO 3956, *L. plantarum* JDM1, *L. plantarum* WCFS1, *L. reuteri* DSM 20016, *L. reuteri* JCM 1112, and *L. salivarius* UCC 118), and GG (i.e., *L. casei* ATCC 334, *L. casei* BL23, *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. sakei* 23K) (**Figure 2; Appendix I**). Of interest, while the larger NCFM clade appeared to have the best taxonomic structure, the WCFS and GG clades each included an outgroup genome from *L. salivarius* and *L. sakei*, respectively. Moreover, although a comparison between this reconstructed phylogeny and the COG distribution of the 20 *Lactobacillus* genomes (see **Table S3**) did not show any obvious correlations, the NCFM and GG clades contained the highest and lowest (respectively) average number of genes in the "translation, ribosomal structure and biogenesis" functional category, whereas it was the genes categorized as "transcription" and "replication, recombination and repair" that showed variation across all clades, though with a few exclusions. Further, among the lactobacilli with the largest genomes (i.e*., L. casei* BL23, *L. rhamnosus* GG and LC705, and *L. plantarum* WCSF1), these garnered the most genes assigned a carbohydrate utilization function, and much like what was already found previously for the *L. plantarum* WCFS1 genome (Kleerebezem et al., 2003).

Other LCG classifications included identifying (1) the core genes of each phylogenomic clade (i.e., core group genes) and (2) those genes found in each of the genomes of a clade, but missing from all other *Lactobacillus* genomes (i.e., signature group genes). Here, it was determined that there are 771 (NCFM), 636 (WCFS), and 991 (GG) core group genes (see **Table 2; Appendix I**, and **Tables S4-S6**) and 119 (NCFM), 14 (WCFS), and 88 (GG) signature group genes (see

Table 2; Appendix I, and **Tables S7-S9**). Further classifications involved the identification of so-called ORFans (i.e., genes present in the genome of a single species, but missing in all other genomes) that are either LCG-specific (i.e., genes present in LCG, but missing from all other genomes) or group-specific (i.e., genes in the core group genes of one clade, but missing from all other genomes). Here, there were 41 LCG-specific ORFans (see **Table 3; Appendix I**, and **Table S13**), whereas the number of group-specific ORFans was variable and tallied up to 56 (NCFM), 4 (WCFS), and 30 (GG) genes (see **Table 3; Appendix I**, and **Tables S10-S12**).

Among the LCG-specific ORFans, these predicted genes seemed to only encode for smaller proteins consisting of 75 residues or less, a surprising outcome that might have arisen from the method used in ORFans identification. Of these ORFans, 13 had functional annotations for hypothetical proteins, several of which were found in the genome as putative operonic gene clusters. On the other hand, 13 of the 28 annotated ORFans had predicted functions for ribosomal proteins, with a few of these also encoded as operons in the genome. Most of the group-specific ORFans of the NCFM clade (34 out of 56) were for genes whose annotations are for unknown proteins or then assigned unreliably, which made informed speculations about their functions less likely. Only one of the four group-specific ORFans of the WCFS clade had been annotated with a putative function, whereas the rest were for hypothetical proteins. However, for this lone ORFan gene, it was assigned a variety of annotations, though the one for a lactoylglutathione lyase (glyoxalase I) seems the most accurate. Since this enzyme is used for breaking down methylglyoxal (a toxic byproduct of glycolytic pathway triosephosphates), and in *Streptococcus mutans* it becomes upregulated in acidic conditions (Korithoski et al., 2007), its functionality would no doubt benefit the lactobacilli in the WCFS clade during acid adaptation. For group-specific ORFans of the GG clade, 50% of these genes were for hypothetical proteins and of the rest with putative functions several of these were for smaller-sized proteins, such as, e.g., a 4Fe-4S ferredoxin and those for bacteriocin immunity. Finally, attempts to pinpoint niche-specific genes in the 20 genomes were unsuccessful, though in many instances it is questionable whether the isolated source is the true niche of a *Lactobacillus* species or strain.

4.2 Study II: Pan-genomics of *L. rhamnosus*

Study II covers research published during 2014 and has been compiled into the PLoS ONE peer-reviewed article entitled “A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*” (see **Appendix II**). The literature review section of this thesis has already provided a good overview of the past and present status of *L. rhamnosus*

genomics, but now study II delves further by a pan-genome appraisal of this species, which at the time was the first of such investigations to be undertaken. However, unlike most published pan-genomic studies, where the genome analysis is a broadly based functional interpretation, study II instead takes on a more focused approach. Here, the genetics behind the pervasive adaptation of *L. rhamnosus* strains to various ecological habitats and environmental settings was explored *in silico*, wherein the emphasis was on proteinaceous cell-surface phenotypes and their variation as it relates to host-niche specialization. For characterizing the pan-genomic data, genomically well-characterized *L. rhamnosus* GG was chosen as a benchmark strain. Moreover, particular attention was placed on a number of membrane- and cell wall-associated proteins that have already been studied at a molecular and biochemical level in the *L. rhamnosus* GG strain or closely related species. The overall intent here was to identify which of the recognizably functional surface proteins could confer a selective or competitive advantage to different *L. rhamnosus* strains and thereby also help promote and sustain a particular type of host or environmental adaptability.

Pan-genome construction involved the use of 13 genome sequences from various *L. rhamnosus* strains. Habitats for these isolates are varied and include dairy (R0011, HN001, and LC705), infected pulp of carious teeth (LRHMDP2 and LRHMDP3), and the human respiratory (ATCC 8530) and intestinal (GG, ATCC 53103, ATCC 21052, LMS2-1, E800, PEL5, and PEL6) tracts. Three of the genomes were sequenced in-house (E800, PEL5, and PEL6), whereas the remainder were obtained from the NCBI RefSeq database. Among the compiled attributes of the genomes (for further details, see **Table 1; Appendix II**), four were fully completed sequences (LC705, GG, ATCC 8530, and ATCC 53103), with the rest being good-quality draft assemblies. As knowing the true source of an isolate is a key constraint when drawing conclusions from pan-genomic data, a molecular phylogeny was reconstructed with the gene sequences of the *L. rhamnosus* core genome as a way of determining whether a commonality exists between the genetic lineages and habitat origins of the various strains. From the phylogenomic tree reconstruction (see **Figure 1; Appendix II**), it appeared that some of the genomes had grouped into clades based on the shared origins of different *L. rhamnosus* strains, such as those from the human mouth (LRHMDP2 and LRHMDP3) and gut (GG, ATCC 53103, PEL5, and PEL6). Then again, this was not the case for the genomes from the dairy isolates (R0011, HN001, and LC705), as each of these formed an individual lineage with genomes from a human isolate. With an element of genomic relatedness missing amongst the dairy strains, this might cast doubt on their reported true origins, which could instead be from human sources. Yet further speculatively, as the genomes from the gut strains (LMS2-1, E800, and ATCC 21052) that paired up with those of the dairy isolates were not found to be part of the gut-origin clade (see above), these particular

strains of *L. rhamnosus* might have only transited the digestive tract and originally come from elsewhere, such as food, vegetation, or soil. Irrespective of this result, it remains reasonable that within the context of the pan-genome some level of inference can be made about the evolutionary and environmental relatedness among the different *L. rhamnosus* strains by analyzing the pooled gene content of the 13 genome sequences.

The pan-genome of *L. rhamnosus* has an estimated size of 4,893 protein-encoding genes, and this is further divided into a core and accessory genome of 2,095 and 2,798 genes, respectively (see **Figure 2; Appendix II**). As can be reasonably expected, this genetic repertoire numerically approximates the results of other and similar genomic studies on the *L. rhamnosus* species (Douillard et al., 2013a; Ceapa et al., 2016). The identities of the annotated genes in the *L. rhamnosus* pan-genome are listed in **Table S1**, which is found online as Supporting Information at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102762.
The entire gene pool in the pan-genome is slightly less than double (1.6-fold) the 2,977 genes averaged from the 13 genomes (**Table 1; Appendix II**). This is in contrast with the taxonomically similar *L. casei* species, whose pan-genome size exceeds the average number of genes of the strain-genomes by a factor of 3.2 (Broadbent et al., 2012). Thus, between these two *Lactobacillus* genomes, a greater degree of genetic instability is associated with *L. casei*, as also this species encodes a higher number of transposes than does *L. rhamnosus*. Despite that, the *L. rhamnosus* pan-genome indicates this species has sufficient genomic plasticity for the geno-phenotypic variation among various strains that allows their adaptation to different environments or habitats. This is further borne out by the Heaps' Law fitted development plot of the pan-genome (see **Figure 3; Appendix II**), as the α-parameter equates to 0.79, which being a value less than one indicates the gene content of the pan-genome is open and not yet saturated. However, when the pan-genome development curve is further examined, its trajectory shows a leveling-out trend at about 5,000 genes, with the possible progression to greater closure. Even so, any lessened expansion of the pan-genome should not pose limitations on further genetic variation in the *L. rhamnosus* species, as this can still result from other events such as single-nucleotide and insertion/deletion polymorphisms or the occurrence of extrachromosomal mobile genetic entities like plasmids and phages.

The 2,095-gene repertoire of the core genome is equivalent to 43% of the pan-genome (see **Figure 2; Appendix II**), which by proportion is much higher than for the *L. casei* pan-genome (1,715 out of 5,935 genes, 29%) (Broadbent et al., 2012), but identical to that found with the pan-genome of *L. paracasei* (1,800 out of 4,200 genes, 43%), the other close taxonomic counterpart of *L. rhamnosus* (Smokvina et al., 2013). Here, it seems that whereas the genetic variation in the

genomes of *L. rhamnosus* and *L. paracasei* are similar, it is comparatively much greater for *L. casei*. Expectedly, among the *L. rhamnosus* core genes, most encode the necessary housekeeping functions for the sustained viability of this species, and although the composition of this genetic pool somewhat matches that of *L. paracasei* and *L. casei*, some inherent species-specific differences are apparent. Of further interest, there are at least 75 core genes in *L. rhamnosus* that are not present in any other *Lactobacillus* species (based on genome sequences deposited in the NCBI RefSeq database as of June 2013), with most being annotated as hypothetical proteins, but many others for predicted proteins whose functions are in membrane transport, transcriptional regulation, and glycosyltransferase activities. As for the *L. rhamnosus* accessory genome, the pooled 2,798 genes can be further partitioned according to whether they are strain-specific, and thus unique to the genome of a particular strain (see **Figure 2; Appendix II**). When tallied up, there were 855 unique genes and these accounted for about 30% of the accessory genome. Additional partitioning of the accessory genome involved those genes that represent ORFan-like sequences, i.e., genes per individual genome that have no homologous counterparts in other *Lactobacillus* genomes (see **Figure 2; Appendix II**). Total numbers for these genes were 519, or 18% of all accessory genes. Among the annotated accessory genes, most had a "black-box" identity and were for hypothetical proteins or those proteins with no known function (**Table S1**). Nevertheless, some of these genes had identifiable phenotypes, such as mobile genetic elements, membrane transporters, transcriptional regulators, outer surface adhesins, and carbohydrate metabolism components. In overall terms, the gene profile of the accessory genome for strains of *L. rhamnosus* essentially mirrors that obtained for the taxonomic cousins *L. casei* and *L. paracasei* (Broadbent et al., 2012; Smokvina et al., 2013).

Given that the prime objective of this study was to probe the putative surface protein phenotypes of *L. rhamnosus*, the pan-genome survey of the 13 genomes was accompanied by an *in silico* pan-secretomic profiling of the predicted protein content. The focus here was on identifying classical secretory proteins and/or Sec-pathway exported proteins, and this was done via Gram-positive SignalP 4.1 predictions. For this, the estimated size of the *L. rhamnosus* pan-secretome includes 230 proteins, which, by comparison, represents just 4.7% of the overall pan-genome, and is further split into a core and accessory secretome of 103 and 127 proteins, respectively (see **Figure 4; Appendix II**). Numerical partitioning of the SignalP-predicted secreted proteins per each genome was also performed (see **Figure 4; Appendix II**). Most identified hits within the core secretome had annotations for hypothetical proteins, many types of ABC transporters, and a number of LPXTG-like and WxL domain surface proteins (see **Table S2**). Moreover, some of the predicted proteins found in the core secretome include those that were already characterized functionally in *L.*

rhamnosus GG (see sections that follow). To a lesser extent, the accessory secretome contained a similar set of annotated proteins, though a vast number (~75) were assigned hypothetical annotations (see **Table S2**). Of interest, the SpaCBA pilus-related proteins were some of the identifiable hits in the accessory secretome (see sections that follow).

Owing to the use of the SignalP program, the *L. rhamnosus* pan-secretome had excluded predictions for non-classically secreted adhesins, though one of which is a matter of interest and worthy of further mention. Amongst the pan-genome repertoire, an ORF (E800_2250) with an annotation specifying a collagen-binding protein was uniquely identified only in the *L. rhamnosus* E800 genome, which in effect also made it a strain-specific part of the accessory genome. By lacking sequence for a classical N-terminal secretion signal, this was the obvious reason why this 379-residue protein had been absent in the SignalP output. Further predictions using the Secretome 2.0 and TatP 1.0 programs revealed it to be a probable non-classically secreted protein (data not shown). Although this collagen-binding protein is not encoded by the other 12 *L. rhamnosus* genomes, NCBI BlastP searches found that the most-related homologs are associated with the plasmid pLC705_00003 of the LC705 strain, several *L. casei* and *L. paracasei* strains, and the *Enterococcus faecium* gut pathobiont (data not shown). However, whereas the ability to bind collagen has yet to be determined for *L. rhamnosus* E800, one can anticipate that the E800_2250 protein is among the factors providing a competitive edge for this strain during the colonization of epithelium-rich host tissues.

For the remaining sections of study II, a pan-genomic assessment of potential variation in cell-surface protein phenotypes for *L. rhamnosus* will be focused on a number surface proteins that have already been characterized biochemically and functionally, which includes SpaCBA and SpaFED piliation, mucus-binding factor (MBF) protein, modulator of adhesion and biofilm (MabA) protein, major-secreted-proteins Msp1 and Msp2 (or p75 and p40, respectively), and fibronectin-binding protein (Fbp) (see **Table 2; Appendix II**). Among their posited or established roles, these different proteins are involved with cellular adhesion and/or microbe-host immune cell interactions. For more thorough background on each protein the reader should refer to information provided in Appendix II.

SpaCBA and SpaFED piliation. As already mentioned in the literature review section, uncovering the operonic genes for sortase-dependent SpaCBA and SpaFED piliation was a significant milestone in *L. rhamnosus* genomics analysis (Kankainen et al., 2009), since, in a wider sense, this discovery fundamentally changed the way these types of pili are viewed, i.e., no longer just as a factor of virulence, but now one of ecological niche adaptation. Found at the time, only the genes for the SpaCBA pilus in the *L. rhamnosus* GG strain were subject to native

(constitutive) expression (Kankainen et al., 2009), whereas those for the SpaFED pilus were otherwise silent, though later they were recombinantly expressible in another bacterial host (Rintahaka et al., 2014). However, as a further outcome, extensive functional and structural characterizations of these pilus structures were performed and reported by a number of research groups (for review, see von Ossowski, 2017).

Based on the pan-genome data for *L. rhamnosus* (**Table S1**), genes for the SpaCBA pilus were not a universal genomic feature of all 13 strains, which then put them among the accessory genes (see **Table 2; Appendix II**). All told, there were just four gut-derived strains (LMS2-1, E800, GG, and ATCC 53103) whose genomes encoded the genes for the tip SpaC (~90 kDa), basal SpaB (~20 kDa), and backbone SpaA (~30 kDa) pilins along with the SrtC1 (~40 kDa) sortase. In this context, the *spaCBA* pilus operon could be regarded a rare genetic attribute of the *L. rhamnosus* species. In any case, by having the genes for surface piliation, these four strains have evolved an "upper hand" for colonizing the host intestine and competing against the menagerie of bacteria therein. Among the *spaCBA* operon-containing strains, the E800 strain contained a constitutively active set of genes and was shown to assemble SpaCBA pili on its cell surface (**Figure S2**). Remarkably, despite the common occurrence of the *spaCBA* pilus operon in the four strains, as well as having similar source origins, other genomic differences had positioned these strains in unrelated lineages in the molecular phylogeny of *L. rhamnosus*, though expectedly with GG and ATCC 53103 paired together, insofar as these strains are the same (see **Figure 1; Appendix II**). However, it was noteworthy that for the other intestinal strains (ATCC 21052, PEL5, and PEL6) and the isolates from the human mouth (LRHMDP2 and LRHMDP3) and airways (ATCC 8530), all of which reside in a mucus-lined and epithelium-rich environment, their genomes lacked the *spaCBA* genes, even though the established mucus- and collagen-binding properties of the SpaCBA pilus would represent an adaptive advantage during host colonization. Clearly, for acquiring the *spaCBA* pilus operon, both mucus and collagen are not the prevailing drivers of evolution at the gene level. Likewise, this is presumably the case for the three milk-derived strains of *L. rhamnosus* (LC705, R0011, and HN001), as none of their genomes encode the *spaCBA* genes. While each of these strains appears to have adapted to a dairy environment ladened with the shed-off mucin (Newburg, 2013) and ECM proteins (Black et al., 1998) of the mammary epithelium, evolutionary pressure was insufficient to drive the selection of the SpaCBA piliation geno-phenotype.

However, the less-than-universal presence of the *spaCBA* pilus operon among the 13 strains offers the possibility that these genes were a more recent genomic acquisition. In support of this notion, the G+C content for the *spaC* (45%), *spaB* (45.3%), *spaA* (44.6%), and *srtC1* (43.3%) genes in the *L. rhamnosus*

GG strain appears lower than the rest of the genome (see **Table 1; Appendix II**), which over evolutionary time might reflect a shorter genetic presence. As a conceivable mechanism and source, the *spaCBA* genes might have been acquired by the *L. rhamnosus* strains through a LGT event involving the *spaCBA* operon-containing *L. casei* and *L. paracasei* species, or more distantly via the piliated *E. faecium* and *Enterococcus faecalis* species. Here, the transposon-like IS (insertion sequence) elements that flank the *spaCBA* pilus operon in *L. rhamnosus* GG (Douillard et al., 2013b) might have had a role in expediting lateral gene movement between bacteria.

On the other hand, pan-genome analysis of *L. rhamnosus* (**Table S1**) instead showed that the core genome contains the clustered genes (*spaFED* operon) for the SpaFED pilus, namely those encoding for the tip SpaF (~104 kDa), basal SpaE (~45 kDa), and backbone SpaD (~51 kDa) pilins and the SrtC2 (~30 kDa) sortase (see **Table 2; Appendix II**). Here, the *spaFED* pilus operon is present in all 13 *L. rhamnosus* strains, four out of which (GG, E800, LC705, and R0011) are confirmed to not constitutively produce the SpaFED pilus (data not shown). In fact, no study has thus far revealed the production of SpaFED pili in any *L. rhamnosus* strain or, for that matter, in the taxonomic close species of *L. casei* and *L. paracasei*. Thus, it would appear that a fully assembled SpaFED pilus in its native host is still under some conjecture. Still, although it seems that the unexpressed *spaFED* genes are not a genetic burden for the various strains, they also pose no added fitness benefit to the *L. rhamnosus* species. It is of speculative interest then to know why the *spaFED* pilus operon would be so widespread among *L. rhamnosus* genomes and not be subjected to the natural evolutionary processes of gene decay and loss. One possible explanation might lie with the G+C content of the genes for the SpaFED pilus. For instance, in the *L. rhamnosus* GG strain, the G+C content is noticeable higher for the *spaF* (49.3%), *spaE* (48.2%), *spaD* (48.9%), and *srtC2* (53.7%) genes than that of the entire genome (see **Table 1; Appendix II**) or the *spaCBA* genes (refer above for details). As inferred from this, the *spaFED* genes might then possess greater stability for better enduring certain DNA-level stresses, and therein is their sustained genomic presence in *L. rhamnosus* or related species.

Mucus-binding factor (MBF) protein. MBF (~38 kDa) is a characterized LPXTG-like protein in *L. rhamnosus* GG and, as its name suggests, it displays a binding affinity for mucus (von Ossowski et al., 2011). Analysis of the pan-genome (**Table S1**) revealed that the gene for the MBF protein is a commonality across the other *L. rhamnosus* strains and thus part of the core genome (see **Table 2; Appendix II**). Such a mucoadhesive geno-phenotype is in keeping with the source origins of strains from the mouth, gut, and airways (GG, ATCC 53103, PEL5, PEL6, LMS2-1, E800, ATCC 21052, and ATCC 8530), as this substrate-binding attribute would likely be expedient and required for host colonization of

the mucosal epithelium. Along the same lines, even those strains with dairy origins (R0011, HN001, and LC705) might also seem to benefit from the MBF, since it is reasonable that milk mucins (Newburg, 2013) would have delivered sufficient evolutionary or genetic pressure for the genomic presence of this mucus-specific protein. However, for those strains not showing the SpaCBA piliation geno-phenotype, it is likely that by relying on MBF adhesiveness they will be less effectual binders of mucus, thus suggesting the possibility of supporting a habitat colonization behavior that is more transient and less sustained.

Modulator of adhesion and biofilm (MabA) protein. MabA is one of the larger-sized LPXTG-like proteins (~250 kDa) that have been characterized in *L. rhamnosus* GG, showing both an adhesive capacity toward gut epithelial cells and an ability to promote biofilm growth (Vélez et al., 2010). According to the pan-genome gene content (**Table S1**), the gene for the MabA protein is present in all 13 *L. rhamnosus* strains and thereby included in the core genome (see **Table 2; Appendix II**). However, with only limited specificity about the MabA phenotype available, it becomes somewhat problematic to make any correlations with respect to the different strains and their source origins and habitat selection. Moreover, despite some sequence variability between the primary structures of MabA (**Figure S1F**), this does not translate to a recognizable association. Then again, if one speculates based on the thus far established properties of MabA (see above), this surface protein might have an adaptive role for those *L. rhamnosus* strains residing in an intestinal environment.

Major-secreted-proteins Msp1 and Msp2. Both Msp1 (p75) and Msp2 (p40) are characterized as cell wall-associated proteins (~47 and ~42 kDa, respectively) in *L. rhamnosus* GG, either being loosely bound to the outer cell-surface or else becoming detached and released from cells (Yan et al., 2007). Functionally, the Msp1 and Msp2 proteins display peptidoglycan hydrolase activities (Claes et al., 2012; Bäuerl et al., 2010), but as well they represent "moonlighting" proteins with additional immunogenic functions (Yan et al., 2007; Yan et al., 2011; Yan and Polk, 2012; Yan et al., 2013). Among the 13 genomes of the *L. rhamnosus* pan-genome (**Table S1**), all have present the genes for Msp1 and Msp2, which makes them part of the core gene pool (see **Table 2; Appendix II**). Interestingly, the presence of Msp1 and Msp2 in the core genome is consistent with the negligible amino acid variation among each of the two protein types (i.e., overall sequence identity of >99% and >99.5% for Msp1 and Msp2, respectively; see **Figures S1G and H**), thereby highlighting their functional and structural prominence as essential hydrolytic components of cell division and separation processes. In a broader context, although the Msp1 and Msp2 would then be key proteins for the different *L. rhamnosus* strains during host colonization and survival, their universal presence in this species would not

dictate any sort of adaptive preference toward a specific type of habitat or environment. However, toward a human or animal host colonized by *L. rhamnosus*, the immune functions of the Msp1 and Msp2 proteins might otherwise provide certain health-related benefits.

Fibronectin-binding protein (Fbp). Fibronectin is one of the ECM glycoproteins comprising the mucosal epithelium, often acting as a specific attachment site for various bacteria (Henderson et al., 2011). Though not yet characterized functionally in any *L. rhamnosus* strain, the Fbp (~64 kDa) in *L. casei* has been shown to be actively adhesive toward fibronectin (Muñoz-Provencio et al., 2010). Intriguingly, for the primary structure of Fbp in both *L. rhamnosus* and *L. casei*, an N-terminal secretion signal and C-terminal anchoring domain are not evident (Kankainen et al., 2009; Muñoz-Provencio et al., 2010), which causes this protein to have a weak cell wall association. Scrutiny of the pan-genome (**Table S1**) indicates that the gene for Fbp is found in each of the 13 *L. rhamnosus* genomes and thus within the core genome structure (see **Table 2; Appendix II**). Expectedly, due to the absence of a "classical" signal sequence at the N-terminus, Fbp went undetected by SignalP and was not among the predicted surface proteins in the pan-secretome of *L. rhamnosus*. However, also to be expected is the evolutionary selection for Fbp in the core genome, given that fibronectin is an endogenous component of epithelial cells, and in such an environment there would be a strong adaptive benefit to displaying a fibronectin-binding phenotype in *L. rhamnosus*.

4.3 Study III: Pan-genomics of *L. ruminis*

The research behind study III is based on the peer-reviewed article entitled "An *in silico* pan-genomic probe for the molecular traits behind *Lactobacillus ruminis* gut autochthony", which was published in PLoS ONE during 2017 (see **Appendix III**). As noted in the literature review section, gut-autochthonic *L. ruminis* has recently emerged as a species of some interest, both from the standpoint of its ecology and potential probiosis. While this has led to a number of published comparative genomics studies on *L. ruminis* (and by the same research group), none of them had tried to examine which of the genetic traits makes this species indigenously adapted (i.e., autochthonous) to the intestinal region of humans and animals. Moreover, among the comparative analyses performed with *L. ruminis* genome data, a pan-genomic comparison was yet to be done. This presented an interesting research opportunity and attainable goal, and thus study III became the merging of these two unfulfilled tasks.

At the time, the undertaking to build a pan-genome of *L. ruminis* was

somewhat held back by the few number of freely available genomes, i.e., only six. However, this was remedied through the in-house sequencing and annotation of three additional *L. ruminis* genomes, one from a porcine strain (GRL1172) and another two from bovine isolates (PEL65 and PEL66). Thus, a sufficient total of nine genomes were used for constructing the *L. ruminis* pan-genome. In addition to the three mentioned sources, the genomes also came from a single equine strain (DPC 6832) and three human strains (ATCC 25644, SPM0211, and S23), and an extra one each from other porcine and bovine strains (DPC 6830 and ATCC 27782, respectively). Moreover, with the inclusion of the bovine isolates of *L. ruminis*, this meant the pan-genome was representative of ruminant and non-ruminant gut ecologies. Among the general features of the nine *L. ruminis* genomes (see **Table 1; Appendix III**), these were consistent with those from other studies mentioned in the literature review (i.e., genome size, GC-content, and number of ORFs, and encoded proteins). Further, given that there was a close genomic likeness between the nine strains, this would suggest the *L. ruminis* species is genetically homogeneous.

As thus far it seems there is no natural habitat for the *L. ruminis* species other than intestinal surroundings, a reconstructed phylogeny was used to determine if the nine genomes show any source relatedness to a particular host gut. According to the phylogenomic tree of core-genome loci, the *L. ruminis* genomes were gathered into clades based on the host source of each strain (see **Figure 1; Appendix III**). Again, a noticeable peculiarity with the molecular phylogeny of *L. ruminis* was that the genomes of human and bovine strains showed a common lineage (as observed previously; O'Donnell et al., 2015), and thus at a genetic level there appears to be no evolutionary distinction in the lineage between the genomes of ruminant and non-ruminant sourced strains.

Once assembled, the estimated size of the *L. ruminis* pan-genome was calculated to be 4,301 protein-encoding genes, out of which the core and accessory genomes held 1,234 and 3,067 genes, respectively (see **Figure 2; Appendix III**, and **S1 Table**; available online as Supporting Information at https://doi.org/10.1371/journal.pone.0175541). Based on the Heaps' Law fitting of the pan-genome development plot (see **Figure 3; Appendix III**), the α-parameter was found to be less than one (i.e., 0.63). This suggested that the pan-genome of *L. ruminis* is open, which means the gene pool for this species has not yet been fully defined. Further still, the pan-genome curve has not plateaued, suggesting the gene content is theoretically unlimited. However, taken from the core genome development plot (see **Figure 3; Appendix III**), there appears to be a sharp leveling out of the core genome curve and the movement to a fixed number of core loci. Concerning the composition of the core genome, its gene pool was typical and meant for the general housekeeping functions necessary for cellular viability. Among these are included various different catabolic and

metabolic pathways and regulatory mechanisms, which as well were reflected by the COG functional classifications (see **Figure 4; Appendix III**). Based on the COG classification estimates, it was logically expected that two percent of the core genes were assigned to the cell motility category, as this would be consistent with the flagellate phenotype of *L. ruminis* cells. Interestingly, since the core genome accounted only for 28.7% of the pan-genome size, this limited number of core genes, and by default the larger accessory genome, would confirm a high genomic plasticity for the *L. ruminis* species. Here, the accessory genes comprise about 70% of the pan-genome, which thus strengthens the genetic diverseness of *L. ruminis* and would be of particular need for this species when adapting to the gut. Further, close to about 30% of the accessory genes were those that are unique or strain-specific (see **Figure 2; Appendix III**), and while most are annotated as hypothetical proteins, some have predicted roles in transport and metabolism, but as well in the lateral acquisition of genes.

By comparison, the overall focus of this *in silico* investigation had differed from that of study II (**Appendix II**), as the *L. ruminis* pan-genome data was used to mainly explore the genetics behind two kinds of molecular phenotypes that presumably help support an indigenous lifestyle in the digestive tract. One of these was the proteinaceous character of the outer surface of *L. ruminis* cells. Since such features would be exposed to the surrounding gut environment, they should have a key role in the way this species can interact with other intestinal bacteria and host cells, and in case of the latter particularly with those surfaces chosen for cellular attachment and colonization. The strategy here was to scrutinize the predicted genes for a broad category of surface-associated proteins (i.e., classical and non-classical secretory proteins, transmembrane proteins, lipoproteins, LPXTG-anchored surface proteins, and the flagellar and chemotaxis proteins) and pinpoint whether there are included any characteristic differences that correlate with strain-adapted ecological niche preference. For categorizing these surface proteins, most were done using specific *in silico* prediction tools (e.g., SignalP 4.1, TatP 1.0, SecretomeP 2.0, TMHMM 2.0, PRED-LIPO, and CW-PRED; see main text **Table 2** for reference citations), along with manual eyeballing of the data output. The tallied output from these various predictions is available online as Supporting Information (see https://doi.org/10.1371/journal.pone.0175541 for more details) and the specifics about their sorting into either core or accessory genomes can be found in Appendix III of this thesis. However, some notable findings and interpretations were obtained from each category of surface-associated protein, and these are worth mentioning (see subsections below).

Classical and non-classical secretory proteins. Among the best hits for predicted classical and non-classical secretory proteins (see **S2-S4 Tables**), many were related to an established outer surface function or activity. However,

one of the non-classically secreted proteins would seem to have a possible and relevant role in cellular adhesion, and thus might contribute to the autochthonic behavior of *L. ruminis*. Identified as a core gene (GRL1172_498, HMPREF0542_10570, LRC_RS05075, LRN_0851, PEL65_1842, PEL66_466, P869_04425, LRU_00261, and LRP_1613) and annotated as the fibronectin-binding protein (Fbp), this surface-associated protein would likely target the fibronectin component of the ECM of gut epithelial cells (Henderson et al., 2011), thereby providing a suitable attachment site for *L. ruminis* colonization within the digestive tract.

Transmembrane proteins. While many of the hits for transmembrane proteins corresponded to ORFs with predicted transporter functions in the metabolism of different macromolecules (see **S5 Table**), and as well various processes and activities involving cellular signaling and gene regulation, one accessory gene for a putative fucose permease (FucP) (Dang et al., 2010) was common to the human-derived ATCC 25644 and S23 strains (HMPREF0542_11925 and P869_03355, respectively). As a transporter protein involved with fucose uptake, this would give *L. ruminis* cells an alternative energy substrate in situations when nutrients are limiting (Hooper et al., 1999). Although the genomes of the ATCC 25644 and S23 strains do not encode the gene for an α-fucosidase enzyme, which catalyzes the removal of fucose from mucin glycans (Katayama et al., 2005), fucose might still be absorbed by *L. ruminis* as the hydrolyzed remnant of other gut bacteria. As this possibility can be interpreted as a habitat-specific fitness gain for the two human isolates of *L. ruminis*, it does remain puzzling why the fucose permease-encoding gene is not universally shared among the other animal strains when mucus is normally associated with the intestinal tract.

Lipoproteins. Somewhat expectedly, most of the hits for lipoproteins (Desvaux et al., 2006) were annotated as ABC transporters, potentially with various functions and different substrate-binding specificities (see **S6 Table**). Thus, it is by their very nature in the transport/exchange of essential metabolic products and nutrients that one might speculate these predicted lipoproteins will have had a key role in the adaptation of the *L. ruminis* species to the gut microcosm.

LPXTG-anchored surface proteins. Positive hits for genes encoding sortase-dependent proteins (SDPs) with an intact C-terminal LPXTG domain region (Desvaux et al., 2006) were nine in number, with three and six being allocated to the core and accessory genomes, respectively (see **S7 Table**). While the annotation for one of the core genes provided no more a functional description than as a cell wall-anchored protein (though an additional BlastP search suggested slight similarity to the mucin-binding protein pfam06458 domain), the two other core genes were functionally annotated as the tip LrpC and backbone LrpA pilin subunits of the LrpCBA pilus (Yu et al., 2015). However, it was not at all expected that the gene for the basal LrpB pilin would be missing from the core

genome, as the three LrpCBA pilus genes are part of an operon. This discrepancy turned out to be due to the genome of the bovine ATCC 27782 strain, where the ORF for the LrpB pilin was concealed as a pseudogene, but prior to that was shown to encode a truncated protein without the N-terminal 40 residues. For the latter, this was the result of a reading-frameshift change caused by two additional adenine bases in the *lrpB* coding sequence. Another discrepancy involved the genome of the equine DPC 6832 isolate, in which a cytosine was absent from a serine codon in the primary structure of the LrpB pilin (Yu et al., 2015). This gave the presence of two ORFs and each with an individual locus tag (i.e., LRN_0080 and LRN_0081). Since it is not certain whether these frameshift inconsistencies are due to an authentic indel mutation or a DNA sequencing mistake, the *lrpB* gene might in fact be part of the *L. ruminis* core genome. Among the six SDP genes in the accessory genome, three were found to be strain-specific (i.e., P869_01660 and P869_04430 from the human S23 strain and LRP_348 from the porcine DPC 6830 strain) and annotated as hypothetical protein. Another two accessory genes were each found to be shared by two different strains. These included LRP_1521 from DPC 6830 and PEL65_242 from PEL65 (both annotated with a starch-debranching pullulanase function) and LRU_00897 from SPM0211 and P869_09985 from S23 (both having hypothetical protein annotations). As the remaining accessory genes (i.e., P869_01660 and P869_04430 from S23) are also assigned as a hypothetical protein, it is largely unknown what functional role is played by most of the accessory SDP genes, particularly in the niche specialization of gut indigenous *L. ruminis*. Nonetheless, based on these pan-genomic findings, it can be concluded that there is a genetic scarcity in the variety of SDPs available to *L. ruminis* cells, and though many of the related genes have uninformative annotations, those encoding for the LrpCBA pilus would seem to play a vital part in bringing about the autochthonic behavior of this intestinal species.

Flagellar and chemotaxis proteins. Motility and chemotaxis functions are both complex and involve a diverse variety of different proteins (Baker et al., 2006; Chaban et al., 2015). Based on the *L. ruminis* pan-genome, a total of 39 core genes are annotated with a flagellar motility function (see **S9 Table**). However, as most of the genomes (GRL1172, ATCC 25644, ATCC 27782, DPC 6832, PEL65, and S23) encode for just one accessory gene, and where the genome of the DPC 6830 strain encodes none, it would seem that *L. ruminis* motility shows little genetic variation from strain-to-strain, and thus relies on a common core of loci for encoding the components and assembly of a flagellum structure. For those ORFs whose annotation indicates a chemotaxis function, there are 16 core genes, but as well, three to six accessory genes (see **S10 Table**). Key core-gene annotations included those for the Che proteins (i.e., CheA, CheB, CheC, CheD, CheR, CheW, and CheY) that regulate the flagellar tumbling frequency (Baker et al., 2006; Chaban et al., 2015), and as well the transmembrane methyl-accepting

chemotaxis proteins (MCPs) (Baker et al., 2006; Chaban et al., 2015), which represent outer surface sensory receptors that detect and bind different types of external chemoattractants or chemorepellents. As the various Che proteins and MCPs interact with each other, both sets of proteins would be needed to yield functional flagella in *L. ruminis* cells. Moreover, since most of the accessory genes encoding chemotaxis proteins have MCP annotations, this suggests the flagella of the *L. ruminis* strains are potentially responsive to stimuli only available from a given host source. Speculatively, the possibility for such phenotype evolvability might help shape *L. ruminis* adaptation to a prevailing gut environment.

As for the second kind of molecular phenotype that was investigated using the pan-genome data, this dealt with the genetic adaptation of the *L. ruminis* species to an oxygen-free environment. What was explored here was how the "strictly anaerobic" *L. ruminis* bacterium is able to cope metabolically with the demands associated with an autochthonous lifestyle in the gut. As the pan-genome data had revealed none of the genomes from the various strains encode for antioxidant enzymes like superoxide dismutase and catalase (i.e., which degrade the superoxide radical and hydrogen peroxide, respectively) (Espey, 2013), *L. ruminis* must have evolved an energy yielding metabolism that allows it to survive in an anoxic microenvironment. Thus, for this part of the study the genetic fitness for utilizing energy substrates via fermentative and anaerobic respiratory processes was examined (see subsections below).

Fermentation. As indicated in the literature review section, *L. ruminis* is an obligate homolactic bacterium and will produce its energy needs through lactic acid fermentation, a process involving the metabolism of hexose sugars (glucose) via the EMP pathway (Kandler, 1983). Based on the pan-genome analysis of the nine *L. ruminis* strains, all of them contain the annotated genes for the enzymatic steps of the EMP pathway, most notably for the indispensable fructose-1,6-bisphosphate (FBP) aldolase enzyme (see **Figure 5A; Appendix III**). Since this fermentative process appears to be a core genome feature, such a finding not only substantiates the genetic basis for the obligately homofermentative metabolism seen with *L. ruminis*, but it implies a fundamental role in the metabolic physiology of this species, thus reflecting the particular demands of the deoxygenated intestinal milieu. Alternatively, various other lactobacilli (i.e., facultative or obligate heterofermenters) can produce energy through the PK pathway (Kandler, 1983; Pessione et al., 2010), which allows for the utilization of both the hexose and pentose carbohydrates (see Appendix III for further details). The ability of *L. ruminis* to ferment the pentose sugar ribose has been thus far inconclusive, as there are two conflicting studies on this point (Tanasupawat et al., 2000; O'Donnell et al., 2011), this concerning the bovine strain NRIC 1689 (i.e., ATCC 27780 or PEL65). With this discrepancy in mind, an examination of

the pan-genome data revealed that the *L. ruminis* core genome holds all of the genes for the PK pathway enzymes (see **Figure 5B; Appendix III**). At first, the gene for the pentose-cleaving phosphoketolase enzyme seemed to be missing from the equine DPC 6832 strain, but this was later cleared up after the locus tags for its genome were updated in the NCBI database and an ORF subsequently became annotated with a phosphoketolase function. However, while the enzymatic components comprising the PK pathway are present in the various *L. ruminis* strains, their functional expression would not yet be certain. Moreover, of the four proteins typically needed during the active cellular uptake of ribose (McLeod et al., 2011), annotated ORFs for only ribokinase (GRL1172_644, HMPREF0542_10159, LRC_RS07720, LRN_1311, PEL65_950, PEL66_2068, P869_06985, LRU_01270, and LRP_932) and RbsR repressor (GRL1172_1281, HMPREF0542_12070, LRC_RS08440, LRN_1438, PEL65_1929, PEL66_741, P869_08795, LRU_01112, and LRP_465) were among the core genes, whereas those for ribose transporter and D-ribose pyranase were absent from all nine genomes. As the genetics for an intact ribose transport system is lacking in *L. ruminis*, it is unlikely that this species can support growth on ribose, which is in keeping with previous findings of others (O'Donnell et al., 2011). Nonetheless, with the presence of the genes for a complete heterofermentative PK pathway in *L. ruminis*, there remains the possibility that some strains might be able to use this alternative way to break down certain pentose and hexose sugars, and ultimately this can represent an adaptive metabolic advantage during autochthonous growth and colonization in the host gut environment.

Anaerobic respiration. For some LAB, another route for procuring energy is through a respiratory metabolism, which would yield no lactate as an end product, but normally requires heme and menaquinone be made accessible to cells (Pedersen et al., 2012). Under anoxic conditions, other compounds as an alternative to oxygen must be used as the final electron acceptor during the respiration process (e.g., nitrate or fumarate), and then complemented by the presence of a corresponding oxidoreductase enzyme (González et al., 2006; Arkhipova and Akimenko, 2005). At a genetic level, only a few LAB encode the needed genes for carrying out anaerobic respiration with nitrate (e.g., *L. plantarum*, *L. fermentum*, and *L. reuteri*) (Brooijmans et al., 2009) or fumarate (e.g., *E. faecalis*) (Huycke et al., 2001) as the terminal oxidant. As for those bacteria capable of anaerobic nitrate respiratory growth, a nitrate reductase can be identified in the genome either as a single gene (for a monomeric intracellular form) or as an operonic cluster of genes (for a membrane-bound complex of three different subunits) (González et al., 2006). With respect to *L. ruminis*, the predicted gene pool of the pan-genome offered no evidence of a nitrate reductase among the nine strains. An ORF with a nitrate reductase annotation was observed in the bovine ATCC 27782 strain (i.e., LRC_RS08915), but it appears to be mis-annotated, as the Pfam domain assignment of its predicted protein

product revealed it is a member of the amidinotransferase family. Clearly, based on these results, *L. ruminis* cells are unlikely to respire using nitrate as their anaerobic electron acceptor. On the other hand, analysis of the pan-genome data disclosed the possibility of anaerobic fumarate respiratory growth by the *L. ruminis* species, as one gene in the core genome appeared to be annotated as a fumarate reductase (i.e., GRL1172_1261, HMPREF0542_10777, LRC_RS01705, LRN_0304, PEL65_229, PEL66_1400, P869_08075, LRU_02203, and LRP_1424) (see **Figure 6; Appendix III**), and whose identity was further supported by an extra BlastP search. While some bacterial fumarate reductases are encoded in the genome as an operon (e.g., for a membrane-located form of four subunits), others are found as a single gene for monomeric protein (Arkhipova and Akimenko, 2005) and this would seem the case for the putative fumarate reductase in *L. ruminis*. Presumably as a metabolic benefit to *L. ruminis* cells, anaerobic fumarate respiration would yield more energy than via lactic acid fermentation (Payne, 2001), with enhanced cellular biomass and long-term survival as a possible outcome. Further, this might allow *L. ruminis* to outcompete other non-respiring gut microbes and, as a possible fitness advantage, thereby help support a gut-indigenous existence.

5. Conclusions

The main crux of this thesis was to adopt a pan-genome approach for exploring the genetic potential of bacteria in the Gram-positive genus *Lactobacillus*. This was undertaken as three separate investigations that focused on the comparative genomics of the *Lactobacillus* species as a whole (study I) and of two individual species, habitat-promiscuous *L. rhamnosus* (study II) and gut-autochthonic *L. ruminis* (study III). From the outset, it should be understood and emphasized that the conclusions drawn from a pan-genomic appraisal are based on *in silico* predictions and theoretical assumptions, and thus it is only after further experimentation and testing that definitive proof will be forthcoming. Nevertheless, the analysis and interpretation of the genomic data extracted from a bacterial pan-genome not only establishes pertinent genetic correlations involving certain phenotypic traits and characteristics, but it also helps to provide the initial groundwork for new hypotheses and concepts, which will ultimately pave the way for empirical study and new research. General conclusions from the three studies of this thesis are already well laid out in their corresponding articles, and for this the reader is referred to appendices I, II, and III. Thus, instead, a brief summary highlighting the key points of each study is outlined below.

The pan-genome investigation of study I was done at the genus level, involving 14 different *Lactobacillus* species and 20 fully sequenced genomes, and its results offered an additional glimpse into the inherent geno-phenotypic prowess of the lactobacilli, and which can now serve as a genetic foothold for further studies with other *Lactobacillus* species. As derived from the pan-genome of 14,000 genes, a small 383-gene core genome was defined for the 14 *Lactobacillus* species and this allowed for a molecular phylogenetic reconstruction and taxonomic grouping of the 20 genomes and their genetic information.

The *L. rhamnosus* pan-genome of study II was comprised of 13 genomes from strains of varied origins, and its genetic content was analyzed for geno-phenotypic variations related to the outer cell surface and how this might impact on the preference for colonizing a particular host or environmental habitat. Here, the prime focus of the pan-genome was on seven functionally characterized surface proteins (i.e., SpaCBA and SpaFED pili, MBF, MabA protein, Msp1 and Msp2 proteins, and Fbp), and out of these only the genes for SpaCBA piliation were part of the accessory genome, which makes the corresponding *spaCBA* pilus operon a genetic rarity in the *L. rhamnosus* species. However, for the SpaCBA-piliated strains of *L. rhamnosus*, such bacteria have evolved a unique mucoadhesive phenotype and one that will protract transient (allochthonous) colonization of the mucosa epithelium, either in the gut or elsewhere. Of the remaining six surface proteins, their genes were found in the core genome, and consequently can be regarded as common or housekeeping in nature. However, because these genes are a universal attribute of the 13 genomes, their protein products would be essential for the physical contact and interactions that the various *L. rhamnosus* strains have with the host and environment.

In study III, the pan-genome of the *L. ruminis* species was built from the genomes of nine different strains in an effort to establish the molecular basis for the autochthonic (indigenous) lifestyle of this gut bacterium. While the genetics behind intestinal indigenousness is no doubt a complex and multifactorial trait, the *L. ruminis* pan-genomic data was analyzed for the geno-phenotypes involved with cellular surface morphology and anaerobic fermentative-respiratory metabolism, both of which are perceived as plausible host colonization determinants. Inferences made from the pan-genome suggest the possibility that a number of surface proteins and an alternative energy-yielding pathway for anaerobic growth combine to give the *L. ruminis* species the necessary fitness advantage for colonizing the oxygen-depleted crevices of the intestinal epithelium. In terms of genomics, these sorts of adaptive phenotypes in *L. ruminis* help make this species an ecological specialist custom-tailored for the gut environment.

6. Acknowledgements

This thesis is in partial fulfillment of the requirements for my degree of Doctor of Philosophy and was done under the auspices of the Faculty of Veterinary Medicine at the University of Helsinki. For this, I extend my sincere appreciation to **Professors Antti Sukura** (Dean of the Faculty of Veterinary Medicine), **Tomi Taira** (Head of the Department of Veterinary Biosciences), and **Olli Vapalahti** (Head of the Division of Microbiology and Epidemiology) for providing the scientific infrastructure and facilities that allowed me to complete my doctoral studies.

I also express my utmost gratitude to my two doctoral supervisors, **Professor Emeritus Airi Palva** and **Docent Ingemar von Ossowski**, though particularly to Airi for her financial support and discussions and Ingemar for his patience and willingness to guide and help me through the thesis-writing process.

Additionally, I offer my genuine thanks to **Docents Jaana Mättö and David Fewer**, **Professor Olli Vapalahti**, and **Professor Per Saris** for their roles and duties as pre-examiners, custos, and opponent, respectively.

Likewise, my appreciative thanks are given to the various co-authors of my three thesis publications, but especially to **Professor Willem de Vos** (study I) and **Docent Ingemar von Ossowski** (study II and III) for their positive driving efforts in seeing that these articles were published.

My added appreciation is extended to my past and present colleagues within the department and beyond, as you have all made my doctoral experience a "memorable" one.

As well, I gratefully acknowledge the University of Helsinki Applied Bioscience (ABS) Graduate School (now known as the Doctoral Program in Food Chain and Health) for one year of funding.

Finally, to my family, I thank my loving wife **Evi** for her unconditional support during these many years of study, and as well my sweet daughters **Liisa** and **Saara** for bringing so much joy into my life.

Röykkä, May 2018

7. References

Adams, J. (2008) Complex genomes: shotgun sequencing. Nat Educ 1(1): 186.

Ahrne, S., E. Lonnermark, A. E. Wold, N. Aberg, B. Hesselmar, R. Saalman, I. L. Strannegard, G. Molin and I. Adlerberth (2005) Lactobacilli in the intestinal microbiota of Swedish infants. Microbes Infect 7(11-12): 1256-1262.

Altermann, E., W. M. Russell, M. A. Azcarate-Peril, R. Barrangou, B. L. Buck, O. McAuliffe, N. Souther, A. Dobson, T. Duong, M. Callanan, S. Lick, A. Hamrick, R. Cano and T. R. Klaenhammer (2005) Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. Proc Natl Acad Sci USA 102(11): 3906-3912.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17): 3389-3402.

Arkhipova, O. V. and V. K. Akumenko (2005) Unsaturated organic acids as terminal electron acceptors for reductase chains of anaerobic bacteria. Microbiology (Translated) 74(6): 629-639.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9: 75.

Bagos, P. G., K. D. Tsirigos, T. D. Liakopoulos and, S. J. Hamodrakas (2008) Prediction of lipoprotein signal peptides in Gram-positive bacteria with a hidden Markov model. J Proteome Res 7: 5082-5093.

Baker, M. D., P. M. Wolanin and J. B. Stock (2006) Signal transduction in bacterial chemotaxis. Bioessays 28(1): 9-22.

Bakke, P., N. Carney, W. Deloache, M. Gearing, K. Ingvorsen, M. Lotz, J. McNair, P. Penumetcha, S. Simpson, L. Voss, M. Win, L. J. Heyer and A. M. Campbell (2009) Evaluation of three automated genome annotations for *Halorhabdus utahensis*. PLoS ONE 4(7): e6291.

Barrangou, R. and P. Horvath (2012) CRISPR: new horizons in phage resistance and strain identification. Annu Rev Food Sci Technol 3: 143-162.

Bauerl, C., G. Perez-Martinez, F. Yan, D. B. Polk and V. Monedero (2010) Functional analysis of the p40 and p75 proteins from *Lactobacillus casei* BL23. J Mol Microbiol Biotechnol 19(4): 231-241.

Bashir, A., A. A. Klammer, W. P. Robins, C. S. Chin, D. Webster, E. Paxinos, D. Hsu, M. Ashby, S. Wang, P. Peluso, R. Sebra, J. Sorenson, J. Bullard, J. Yen, M. Valdovino, E. Mollova, K. Luong, S. Lin, B. LaMay, A. Joshi, L. Rowe, M. Frace, C. L. Tarr, M. Turnsek, B. M. Davis, A. Kasarskis, J. J. Mekalanos, M. K. Waldor and E. E. Schadt (2012) A hybrid approach for the automated finishing of bacterial genomes. Nat Biotechnol 30(7): 701-707.

Batzoglou, S. (2005) The many faces of sequence alignment. Brief Bioinform 6(1): 6-22.

Becerra, J. E., M. J. Yebra and V. Monedero (2015) An L-fucose operon in the probiotic *Lactobacillus rhamnosus* GG is involved in adaptation to gastrointestinal conditions. Appl Environ Microbiol 81(11): 3880-3888.

Beckloff, N., S. Starkenburg, T. Freitas and P. Chain (2012) Bacterial Genome Annotation. In: Microbial Systems Biology. Methods in Molecular Biology (Vol. 881). Humana Press, Totowa, NJ.

Bendtsen, J. D., L. Kiemer, A. Fausbøll and, S. Brunak (2005a) Non-classical protein secretion in bacteria. BMC Microbiol 5: 58.

Bendtsen, J. D., H. Nielsen, D. Widdick, T. Palmer and S. Brunak (2005b) Prediction of twin-arginine signal peptides. BMC Bioinformatics 6: 167.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler, (2008) GenBank Nucleic Acids Resm36 (Database issue): D25-D30.

Berger, B., R. D. Pridmore, C. Barretto, F. Delmas-Julien, K. Schreiber, F. Arigoni and H. Brussow (2007) Similarity and differences in the *Lactobacillus acidophilus* group identified by polyphasic analysis and comparative genomics. J Bacteriol 189(4): 1311-1321.

Bernardeau, M., J. P. Vernoux, S. Henri-Dubernet and M. Gueguen (2008) Safety assessment of dairy microorganisms: the *Lactobacillus* genus. Int J Food Microbiol 126(3): 278-285.

Black, B. F., L. Jarman, J. Simpson (1998) The Science of Breastfeeding (Vol. 3). Jones and Bartlett Publishers, Sudbury, MA, USA.

Blom, J., S. P. Albaum, D. Doppmeier, A. Puhler, F. J. Vorholter, M. Zakrzewski and A. Goesmann (2009) EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics 10: 154.

Boekhorst, J., R. J. Siezen, M. C. Zwahlen, D. Vilanova, R. D. Pridmore, A. Mercenier, M. Kleerebezem, W. M. de Vos, H. Brussow and F. Desiere (2004) The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive

differences in chromosome organization and gene content. Microbiology 150(Pt 11): 3601-3611.

Broadbent, J. R., E. C. Neeno-Eckwall, B. Stahl, K. Tandee, H. Cai, W. Morovic, P. Horvath, J. Heidenreich, N. T. Perna, R. Barrangou and J. L. Steele (2012) Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. BMC Genomics 13: 533.

Brooijmans, R., W. M. de Vos and J. Hugenholtz (2009) Electron transport chains of lactic acid bacteria: walking on crutches is part of their lifestyle. F1000 Biol Rep 1: 34.

Burland, V., G. Plunkett, 3rd, D. L. Daniels and F. R. Blattner (1993) DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. Genomics 16(3): 551-561.

Canchaya, C., M. J. Claesson, G. F. Fitzgerald, D. van Sinderen and P. W. O'Toole (2006) Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. Microbiology 152(Pt 11): 3185-3196.

Cavanagh, D., G. F. Fitzgerald and O. McAuliffe (2015) From field to fermentation: the origins of *Lactococcus lactis* and its domestication to the dairy environment. Food Microbiol 47: 45-61.

Ceapa, C., M. Davids, J. Ritari, J. Lambert, M. Wels, F. P. Douillard, T. Smokvina, W. M. de Vos, J. Knol and M. Kleerebezem (2016) The variable regions of *Lactobacillus rhamnosus* genomes reveal the dynamic evolution of metabolic and host-adaptation repertoires. Genome Biol Evol 8(6): 1889-1905.

Chaban, B., H. V. Hughes and M. Beeby (2015) The flagellum in bacterial pathogens: for motility and a whole lot more. Semin Cell Dev Biol 46: 91-103.

Chen, C., H. Huang and C. H. Wu (2017) Protein Bioinformatics Databases and Resources. Methods Mol Biol 1558: 3-39.

Chin, C. S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner and J. Korlach (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10(6): 563-569.

Chun, B. H., K. H. Kim, H. H. Jeon, S. H. Lee and C. O. Jeon (2017) Pan-genomic and transcriptomic analyses of *Leuconostoc mesenteroides* provide insights into its genomic and metabolic features and roles in kimchi fermentation. Sci Rep 7(1): 11504.

Claes, I. J. J., M. E. Segers, T. L. A. Verhoeven, M. Dusselier, B. F. Sels, S. C. J. De Keersmaecker, J. Vanderleyden and S. Lebeer (2012) Lipoteichoic acid is an

important microbe-associated molecular pattern of *Lactobacillus rhamnosus* GG. Microb Cell Fact 11: 161.

Claesson, M. J., D. van Sinderen and P. W. O'Toole (2008) *Lactobacillus* phylogenomics: towards a reclassification of the genus. Int J Syst Evol Microbiol 58(Pt 12): 2945-2954.

Collins, M. D., S. Wallbanks, D. J. Lane, J. Shah, R. Nietupski, J. Smida, M. Dorsch and E. Stackebrandt (1991) Phylogenetic analysis of the genus *Listeria* based on reverse transcriptase sequencing of 16S rRNA. Int J Syst Bacteriol 41(2): 240-246.

Computational Pan-Genomics Consortium (2018) Computational pan-genomics: status, promises and challenges. Brief Bioinform 19(1): 118-135.

Dang, S., L. Sun, Y. Huang, F. Lu, Y. Liu, H. Gong, J. Wang and N. Yan (2010) Structure of a fucose transporter in an outward-open conformation. Nature 467: 734-738.

Danne, C. and S. Dramsi (2012) Pili of gram-positive bacteria: roles in host colonization. Res Microbiol 163(9-10): 645-658.

Davidsen, T., E. Beck, A. Ganapathy, R. Montgomery, N. Zafar, Q. Yang, R. Madupu, P. Goetz, K. Galinsky, O. White and G. Sutton (2010) The comprehensive microbial resource. Nucleic Acids Res 38(Database issue): D340-D345.

Delcher, A. L., K. A. Bratke, E. C. Powers and S. L. Salzberg (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23(6): 673-679.

Desvaux, M., E. Dumas, I. Chafsey and M. Hébraud (2006) Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. FEMS Microbiol Lett 256: 1-15.

Dettman, J. R., N. Rodrigue, S. D. Aaron and R. Kassen (2013) Evolutionary genomics of epidemic and nonepidemic strains of *Pseudomonas aeruginosa*. Proc Natl Acad Sci USA 110(52): 21065-21070.

Di Cerbo, A., B. Palmieri, M. Aponte, J. C. Morales-Medina and T. Iannitti (2016) Mechanisms and therapeutic effectiveness of lactobacilli. J Clin Pathol 69(3): 187-203.

Douillard, F. P., A. Ribbera, H. M. Jarvinen, R. Kant, T. E. Pietila, C. Randazzo, L. Paulin, P. K. Laine, C. Caggia, I. von Ossowski, J. Reunanen, R. Satokari, S. Salminen, A. Palva and W. M. de Vos (2013b) Comparative genomic and functional analysis of *Lactobacillus casei* and *Lactobacillus rhamnosus* strains marketed as probiotics.

Appl Environ Microbiol 79(6): 1923-1933.

Douillard, F. P., A. Ribbera, R. Kant, T. E. Pietila, H. M. Jarvinen, M. Messing, C. L. Randazzo, L. Paulin, P. Laine, J. Ritari, C. Caggia, T. Lahteinen, S. J. Brouns, R. Satokari, I. von Ossowski, J. Reunanen, A. Palva and W. M. de Vos (2013a) Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. PLoS Genet 9(8): e1003683.

Dressman, D., H. Yan, G. Traverso, K. W. Kinzler and B. Vogelstein (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA 100(15): 8817-8822.

Duar, R. M., X. B. Lin, J. Zheng, M. E. Martino, T. Grenier, M. E. Perez-Munoz, F. Leulier, M. Ganzle and J. Walter (2017) Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. FEMS Microbiol Rev 41(Supp_1): S27-S48.

Eddy, S. R. (1998) Profile Hidden Markov Models. Bioinformatics 14: 755-763.

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

Edwards, D. J. and K. E. Holt (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. Microb Inform Exp. 3(1): 2.

Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach and S. Turner (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910): 133-138.

El Kafsi, H., J. Binesse, V. Loux, J. Buratti, S. Boudebbouze, R. Dervyn, S. Kennedy, N. Galleron, B. Quinquis, J. M. Batto, B. Moumen, E. Maguin and M. van de Guchte (2014) *Lactobacillus delbrueckii* ssp. *lactis* and ssp. *bulgaricus*: a chronicle of evolution in action. BMC Genomics 15: 407.

Emanuelsson, O., S. Brunak, G. von Heijne and H. Nielsen (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4): 953-971.

Espey, M. G. (2013) Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota. Free Radic Biol Med 55: 130-140.

Espino, E., K. Koskenniemi, L. Mato-Rodriguez, T. A. Nyman, J. Reunanen, J. Koponen, T. Ohman, P. Siljamaki, T. Alatossava, P. Varmanen and K. Savijoki (2015) Uncovering surface-exposed antigens of *Lactobacillus rhamnosus* by cell shaving proteomics and two-dimensional immunoblotting. J Proteome Res 14(2): 1010-1024.

Felis, G. E. and F. Dellaglio (2007) Taxonomy of lactobacilli and bifidobacteria. Curr Issues Intest Microbiol 8(2): 44-61.

Felsenstein, J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6.

Fimereli, D. K., K. D. Tsirigos, Z. I. Litou, T. D. Liakopoulos, P. G. Bagos and S. J. Hamodrakas (2012) CW-PRED: a HMM-based method for the classification of cell wall-anchored proteins of Gram-positive bacteria. In: Artificial Intelligence: Theories and Applications (Vol. 7297). Springer, Berlin-Heidelberg, Germany.

Finn, R.D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young and A. L. Mitchell (2017) InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45(D1): D190-D199.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate and A. Bateman (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1): D279-D285.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223): 496-512.

Flicek, P. and E. Birney (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6(11 Suppl): S6-S12.

Forde, B. M. and P. W. O'Toole (2013) Next-generation sequencing technologies and their impact on microbial genomics. Brief Funct Genomics 12(5): 440-453.

Forde, B. M., B. A. Neville, M. M. O. Donnell, E. Riboulet-Bisson, M. J. Claesson, A. Coghlan, R. P. Ross and P. W. O. Toole (2011) Genome sequences and comparative genomics of two *Lactobacillus ruminis* strains from the bovine and human intestinal tracts. Microb Cell Fact 10(Suppl 1): S13.

Friedberg, I. (2006) Automated protein function prediction: the genomic challenge. Brief Bioinform 7(3): 225-242.

Fujisawa, T., Y. Benno, T. Yaeshima and T. Mitsuoka (1992) Taxonomic study of the *Lactobacillus acidophilus* group, with recognition of *Lactobacillus gallinarum* sp. nov. and *Lactobacillus johnsonii* sp. nov. and synonymy of *Lactobacillus acidophilus* group A3 (Johnson Et-Al 1980) with the type strain of *Lactobacillus amylovorus* (Nakamura 1981). Int J Syst Bacteriol 42(3): 487-491.

Garrido-Cardenas, J. A., F. Garcia-Maroto, J. A. Alvarez-Bermejo and F. Manzano-Agugliaro (2017) DNA sequencing sensors: an overview. Sensors 17: 588.

Giraffa, G., N. Chanishvili and Y. Widyastuti (2010) Importance of lactobacilli in food and feed biotechnology. Res Microbiol 161(6): 480-487.

Glaser, P., F. Kunst, M. Arnaud, M. P. Coudart, W. Gonzales, M. F. Hullo, M. Ionescu, B. Lubochinsky, L. Marcelino, I. Moszer, E. Presecan, M. Santana, E. Schneider, J. Schweizer, A. Vertes, G. Rapoport and A. Danchin (1993) *Bacillus subtilis* genome project: cloning and sequencing of the 97kb region from 325-degrees to 333 degrees. Mol Microbiol 10(2): 371-384.

Gonzalez, P. J., C. Correia, I. Moura, C. D. Brondino and J. J. Moura (2006) Bacterial nitrate reductases: molecular and biological aspects of nitrate reduction. J Inorg Biochem 100(5-6): 1015-1023.

Guimaraes, L. C., L. B. de Jesus, M. V. C. Viana, A. Silva, R. T. J. Ramos, S. de Castro Soares and V. Azevedo (2015) Inside the pan-genome–methods and software overview. Curr Genom 16: 245-52.

Guo, J., N. Xu, Z. Li, S. Zhang, J. Wu, D. H. Kim, M. Sano Marma, Q. Meng, H. Cao, X. Li, S. Shi, L. Yu, S. Kalachikov, J. J. Russo, N. J. Turro and J. Ju (2008) Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci USA 105(27): 9145-9150.

Gupta, N., S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith and P. A. Pevzner (2007) Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. Genome Res 17: 1362-1377.

Haakensen, M., A. Schubert and B. Ziola (2009) Broth and agar hop-gradient plates used to evaluate the beer-spoilage potential of *Lactobacillus* and *Pediococcus* isolates. Int J Food Microbiol 130(1): 56-60.

Haft, D.H., J. D. Selengut and O. White (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31(1): 371-373.

Hawkins, T., S. Luban and D. Kihara (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15(6): 1550-1556.

Henderson, B., S. Nair, J. Pallas and M. A. Williams (2011) Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins. FEMS Microbiol Rev 35(1): 147-200.

Hevia, A., S. Delgado, B. Sanchez and A. Margolles (2015) Molecular players involved in the interaction between beneficial bacteria and the immune system. Front Microbiol 6: 1285.

Hooper, L. V., J. Xu, P. G. Falk, T. Midtvedt and J. L. Gordon (1999) A molecular sensor that allows a gut commensal to control its nutrient foundation in a competitive ecosystem. Proc Natl Acad Sci USA 96: 9833-9838.

Huang, Y. F., S. C. Chen, Y. S. Chiang, T. H. Chen and K. P. Chiu (2012) Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst Biol 6(Suppl 2):S10.

Huycke, M. M., D. Moore, W. Joyce, P. Wise, L. Shepard, Y. Kotake and M. S. Gilmore (2001) Extracellular superoxide production by *Enterococcus faecalis* requires demethylmenaquinone and is attenuated by functional terminal quinol oxidases. Mol Microbiol 42(3): 729-740.

Hyatt, D., G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

Jorth, P., U. Trivedi, K. Rumbaugh and M. Whiteley (2013) Probing bacterial metabolism during infection using high-resolution transcriptomics. J Bacteriol 195(22): 4991-4998.

Kandler, O. (1983). Carbohydrate metabolism in lactic acid bacteria. Antonie Van Leeuwenhoek 49(3): 209-224.

Kankainen, M., L. Paulin, S. Tynkkynen, I. von Ossowski, J. Reunanen, P. Partanen, R. Satokari, S. Vesterlund, A. P. Hendrickx, S. Lebeer, S. C. De Keersmaecker, J. Vanderleyden, T. Hamalainen, S. Laukkanen, N. Salovuori, J. Ritari, E. Alatalo, R. Korpela, T. Mattila-Sandholm, A. Lassig, K. Hatakka, K. T. Kinnunen, H. Karjalainen, M. Saxelin, K. Laakso, A. Surakka, A. Palva, T. Salusjarvi, P. Auvinen and W. M. de Vos (2009) Comparative genomic analysis of *Lactobacillus rhamnosus* GG reveals pili containing a human-mucus binding protein. Proc Natl Acad Sci USA 106(40): 17193-17198.

Katayama, T., K. Fujita and K. Yamamoto. Novel bifidobacterial glycosidases acting on sugar chains of mucin glycoproteins. J Biosci Bioeng 99: 457-465.

Kleerebezem, M., J. Boekhorst, R. van Kranenburg, D. Molenaar, O. P. Kuipers, R. Leer, R. Tarchini, S. A. Peters, H. M. Sandbrink, M. W. Fiers, W. Stiekema, R. M. Lankhorst, P. A. Bron, S. M. Hoffer, M. N. Groot, R. Kerkhoven, M. de Vries, B. Ursing, W. M. de Vos and R. J. Siezen (2003) Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc Natl Acad Sci USA 100(4): 1990-1995.

Klimke, W., C. O'Donovan, O. White, J. R. Brister, K. Clark, B. Fedorov, I. Mizrachi, K. D. Pruitt and T. Tatusova (2011) Solving the problem: genome annotation standards before the data deluge. Stand Genomic Sci 5(1): 168-193.

Koonin, E. V. and M. Y. Galperin (2003) Sequence - Evolution - Function: computational approaches in comparative genomics. Kluwer Academic, Boston, MA, USA.

Korithoski, B., C. M. Levesque and D. G. Cvitkovitch (2007) Involvement of the detoxifying enzyme lactoylglutathione lyase in *Streptococcus mutans* aciduricity. J Bacteriol 189(21): 7586-7592.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.

Krumsiek, J., R. Arnold and T. Rattei (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23(8): 1026-1028.

Lagesen, K., P. Hallin, E. A. Rodland, H. H. Staerfeldt, T. Rognes and D. W. Ussery (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35(9): 3100-3108.

Langille, M. G., W. W. Hsiao and F. S. Brinkman (2010) Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol 8(5): 373-382.

Laslett, D. and B. Canback (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32(1): 11-16.

Lawley, B., I. M. Sims and G. W. Tannock (2013) Whole-transcriptome shotgun sequencing (RNA-seq) screen reveals upregulation of cellobiose and motility operons of *Lactobacillus ruminis* L5 during growth on tetrasaccharides derived from barley beta-glucan. Appl Environ Microbiol 79(18): 5661-5669.

Lebeer, S., T. L. Verhoeven, G. Francius, G. Schoofs, I. Lambrichts, Y. Dufrene, J. Vanderleyden and S. C. De Keersmaecker (2009) Identification of a gene cluster for the biosynthesis of a long, galactose-rich exopolysaccharide in *Lactobacillus*

rhamnosus GG and functional analysis of the priming glycosyltransferase. Appl Environ Microbiol 75(11): 3554-3563.

Ledergerber, C. and C. Dessimoz (2011) Base-calling for next-generation sequencing platforms. Brief Bioinform 12(5): 489-497.

Lee, B., T. Kim, S. K. Kim, K. H. Lee and D. Lee (2007) Patome: a database server for biological sequence annotation and analysis in issued patents and published patent applications. Nucleic Acids Res 35(Database issue): D47-50.

Lerat, E., V. Daubin and N. A. Moran (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLoS Biol 1: 101–109.

Lerche, M. and G. Reuter (1961) Isolierung und Differenzierung anaerober Lactobacillaceae aus dem Darm erwachsener Menschen (Beitrag zum *Lactobacillus bifidus*-Problem). Zentralbl Bakteriol 180: 324–356.

Leroy, F. and L. De Vuyst (2004) Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends Food Sci. Technol 15(2): 67-78.

Lima-Mendez, G., J. Van Helden, A. Toussaint and R. Leplae (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24(6): 863-865.

Liu, B. and M. Pop (2009) ARDB: antibiotic resistance genes database. Nucleic Acids Res 37(Database issue): D443-447.

Liu, L., Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu and M. Law (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.

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

Lombard, V., H. Golaconda Ramulu, E. Drula, P. M. Coutinho and B. Henrissat (2014) The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 42(Database issue): D490-495.

Lowe, T. M. and S. R. Eddy (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5): 955-964.

Lukjancenko, O., D. W. Ussery and T. M. Wassenaar (2012) Comparative

genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. Microb Ecol 63(3): 651-673.

MacLean, D., J. D. Jones and D. J. Studholme (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol 7(4): 287-296.

Makarova, K., A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D. M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J. H. Lee, I. Diaz-Muniz, B. Dosti, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson, S. Kozyavkin, B. Weimer and D. Mills (2006) Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci USA 103(42): 15611-15616.

Marchler-Bauer, A., M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng and S. H. Bryant (2015) CDD: NCBI's conserved domain database. Nucleic Acids Res 43(Database issue): D222-226.

Marchler-Bauer, A., A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer, and S. H. Bryant (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30(1): 281-283.

Mardis, E. R. (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376-380.

Markowitz, V. M., I. M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova and N. C. Kyrpides (2012) IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40(Database issue): D115-122.

Martin, D. M., M. Berriman and G. J. Barton (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5: 178.

Martin, R., G. H. Heilig, E. G. Zoetendal, H. Smidt and J. M. Rodriguez (2007) Diversity of the *Lactobacillus* group in breast milk and vagina of healthy women and potential role in the colonization of the infant gut. J Appl Microbiol 103(6): 2638-2644.

Maze, A., G. Boel, M. Zuniga, A. Bourand, V. Loux, M. J. Yebra, V. Monedero, K. Correia, N. Jacques, S. Beaufils, S. Poncet, P. Joyet, E. Milohanic, S. Casaregola, Y. Auffray, G. Perez-Martinez, J. F. Gibrat, M. Zagorec, C. Francke, A. Hartke and J. Deutscher (2010) Complete genome sequence of the probiotic *Lactobacillus casei* strain BL23. J Bacteriol 192(10): 2647-2648.

McLeod, A., L. Snipen, K. Naterstad and L. Axelsson (2011) Global transcriptome response in *Lactobacillus sakei* during growth on ribose. BMC Microbiol 11: 145.

Metzker, M. L. (2005) Emerging technologies in DNA sequencing. Genome Res 15(12): 1767-1776.

Miller, J. R., S. Koren and G. Sutton (2010) Assembly algorithms for next-generation sequencing data. Genomics 95(6): 315-327.

Munoz-Provencio, D., G. Perez-Martinez and V. Monedero (2010) Characterization of a fibronectin-binding protein from *Lactobacillus casei* BL23. J Appl Microbiol 108(3): 1050-1059.

Nadkarni, M. A., Z. L. Chen, M. R. Wilkins and N. Hunter (2014) Comparative genome analysis of *Lactobacillus rhamnosus* clinical isolates from initial stages of dental pulp infection: identification of a new exopolysaccharide cluster. PloS ONE 9(3): e90643.

Nelson, K. E., G. M. Weinstock, S. K. Highlander, K. C. Worley, H. H. Creasy, J. R. Wortman, D. B. Rusch, M. Mitreva, E. Sodergren, A. T. Chinwalla, M. Feldgarden, D. Gevers, B. J. Haas, R. Madupu, D. V. Ward, B. Birren, R. A. Gibbs, B. Methe, J. F. Petrosino, R. L. Strausberg, G. G. Sutton, O. R. White, R. K. Wilson, S. Durkin, S. Gujja, C. Howarth, C. D. Kodira, N. Kyrpides, R. Madupu, T. Mehta, M. Mitreva, D. M. Muzny, M. Pearson, K. Pepin, A. Pati, X. Qin, C. Yandava, Q. D. Zeng, L. Zhang, A. M. Berlin, L. Chen, T. A. Hepburn, J. Johnson, J. McCorrison, J. Miller, P. Minx, C. Nusbaum, C. Russ, G. G. Sutton, S. M. Sykes, C. M. Tomlinson, S. Young, W. C. Warren, J. Badger, J. Crabtree, R. Madupu, V. M. Markowitz, J. Orvis, D. B. Rusch, G. G. Sutton, A. Cree, S. Ferriera, M. Gillis, L. D. Hemphill, V. Joshi, C. Kovar, K. A. Wetterstrand, A. Abouellleil, A. M. Wollam, C. J. Buhay, Y. Ding, S. Dugan, L. L. Fulton, R. S. Fulton, M. Holder, J. Hostetler, G. G. Sutton, E. Allen-Vercoe, J. Badger, S. W. Clifton, A. M. Earl, C. N. Farmer, M. G. Giglio, K. Liolios, M. G. Surette, G. G.

Sutton, M. Torralba, Q. Xu, C. Pohl, S. Durkin, G. G. Sutton, K. Wilczek-Boney, D. H. Zhu and H. M. Jumpstart (2010) A catalog of reference genomes from the human microbiome. Science 328(5981): 994-999.

Newburg, D. S. (2013) Glycobiology of human milk. Biochemistry (Mosc) 78(7): 771-785.

O'Callaghan, J. and P. W. O'Toole (2013) *Lactobacillus*: host-microbe relationships. Curr Top Microbiol Immunol 358: 119-154.

O'Donnell, M. M., B. M. Forde, B. Neville, P. R. Ross and P. W. O'Toole (2011) Carbohydrate catabolic flexibility in the mammalian intestinal commensal *Lactobacillus ruminis* revealed by fermentation studies aligned to genome annotations. Microb Cell Fact 10 Suppl 1: S12.

O'Donnell, M. M., H. M. B. Harris, D. B. Lynch, R. P. Ross and P. W. O'Toole (2015) *Lactobacillus ruminis* strains cluster according to their mammalian gut source. BMC Microbiology 15: 80.

Oh, J., A. L. Byrd, C. Deming, S. Conlan, N. C. S. Program, H. H. Kong and J. A. Segre (2014) Biogeography and individuality shape function in the human skin metagenome. Nature 514(7520): 59-64.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Rückert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko and V. Vonstein (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33(17): 5691-5702.

Overbeek, R., N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov Jr, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, A. Bhattacharyya, H. Burd, W. Gardner, P. Hanke, V. Kapatral, N. Mikhailova, O. Vasieva, A. Osterman, V. Vonstein, M. Fonstein, N. Ivanova and N. Kyrpides (2003) The ERGO genome analysis and discovery system. Nucleic Acids Res 31(1): 164-171

Pace, F., M. Pace and G. Quartarone (2015) Probiotics in digestive diseases: focus on *Lactobacillus* GG. Minerva Gastroenterol Dietol 61(4): 273-292.

Payne, W. J. (2001) Anaerobic respiration. eLS.

Pedersen, M. B., P. Gaudu, D. Lechardeur, M. A. Petit and A. Gruss (2012) Aerobic respiration metabolism in lactic acid bacteria and uses in biotechnology. Annu Rev Food Sci Technol 3: 37-58.

Pessione, A., C. Lamberti and E. Pessione (2010) Proteomics as a tool for studying energy metabolism in lactic acid bacteria. Mol Biosyst 6: 1419-1430.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10): 785-786.

Placzek, S., I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack and D. Schomburg (2017) BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res 45: D380-388.

Pop, M. (2009) Genome assembly reborn: recent computational challenges. Brief Bioinform 10(4): 354-366.

Pot, B., G. E. Felis, K. De Bruyne, E. Tsakalidou, K. Papadimitriou, J. Leisner and P. Vandamme (2014) The genus *Lactobacillus*. In: Lactic Acid Bacteria: Biodiversity and Taxonomy. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Pridmore, R. D., B. Berger, F. Desiere, D. Vilanova, C. Barretto, A. C. Pittet, M. C. Zwahlen, M. Rouvet, E. Altermann, R. Barrangou, B. Mollet, A. Mercenier, T. Klaenhammer, F. Arigoni and M. A. Schell (2004) The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. Proc Natl Acad Sci USA 101(8): 2512-2517.

Proft, T. and E. N. Baker (2009) Pili in Gram-negative and Gram-positive bacteria: structure, assembly and their role in disease. Cell Mol Life Sci 66(4): 613-635.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33 (Database issue): D501-D504.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, H. I. T. C. Meta, P. Bork, S. D. Ehrlich and J. Wang (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285): 59-65.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow and Y. Gu (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Rawlings, N. D., M. Waller, A. J. Barrett and A. Bateman (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res 42(Database issue): D503-509.

Reuter, G. (2001) The *Lactobacillus* and *Bifidobacterium* microflora of the human intestine: composition and succession. Curr Issues Intest Microbiol 2(2): 43-53.

Rintahaka, J., X. Yu, R. Kant, A. Palva and I. von Ossowski (2014) Phenotypical analysis of the *Lactobacillus rhamnosus* GG fimbrial *spaFED* operon: surface expression and functional characterization of recombinant SpaFED pili in *Lactococcus lactis*. PLoS ONE 9(11): e113922.

Rokas, A., B. L. Williams, N. King and S. B. Carroll (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960): 798-804.

Rost, B., J. Liu, R. Nair, K. O. Wrzeszczynski and Y. Ofran (2003) Automatic prediction of protein function. Cell Mol Life Sci 60(12): 2637-2650.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth and J. Bustillo (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356): 348-352.

Saier, M.H., V. S. Reddy, B. V. Tsu, M. S. Ahmed, C. Li and G. Moreno-Hagelsieb (2016) The transporter classification database (TCDB): recent advances. Nucleic Acids Res 44: D372-379.

Salvetti, E. and P. W. O'Toole (2017) The genomic basis of lactobacilli as health-promoting organisms. Microbiol Spectr 5(3).

Salvetti, E., S. Torriani and G. E. Felis (2012) The genus *Lactobacillus*: a taxonomic update. Probiotics Antimicrob Proteins 4(4): 217-226.

Sanger, F., S. Nicklen and A. R. Coulson (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12): 5463-5467.

Schnoes, A. M., S. D. Brown, I. Dodevski and P. C. Babbitt (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5(12): e1000605.

Schuster, S. C. (2008) Next-generation sequencing transforms today's biology. Nat Methods 5(1): 16-18.

Sheikh, M. A. and Y. Erlich (2012) Base-calling for bioinformaticians. In: Bioinformatics for High Throughput Sequencing. Springer, New York, NY, USA.

Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra and G. M. Church (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741): 1728-1732.

Siezen, R. J., V. A. Tzeneva, A. Castioni, M. Wels, H. T. Phan, J. L. Rademaker, M. J. Starrenburg, M. Kleerebezem, D. Molenaar and J. E. van Hylckama Vlieg (2010) Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches. Environ Microbiol 12(3): 758-773.

Smokvina, T., M. Wels, J. Polka, C. Chervaux, S. Brisse, J. Boekhorst, J. E. van Hylckama Vlieg and R. J. Siezen (2013) *Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity. PLoS ONE 8(7): e68731.

Snart, J., R. Bibiloni, T. Grayson, C. Lay, H. Zhang, G. E. Allison, J. K. Laverdiere, F. Temelli, T. Vasanthan, R. Bell and G. W. Tannock (2006) Supplementation of the diet with high-viscosity beta-glucan results in enrichment for lactobacilli in the rat cecum. Appl Environ Microbiol 72(3): 1925-1931.

Sonnhammer, E. L. and R. Durbin (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167(1-2): GC1-10.

Spinler, J. K., A. Sontakke, E. B. Hollister, S. F. Venable, P. L. Oh, M. A. Balderas, D. M. Saulnier, T. A. Mistretta, S. Devaraj, J. Walter, J. Versalovic and S. K. Highlander (2014) From prediction to function using evolutionary genomics: human-specific ecotypes of *Lactobacillus reuteri* have diverse probiotic functions. Genome Biol Evol 6(7): 1772-1789.

Sun, Z. H., H. M. B. Harris, A. McCann, C. Y. Guo, S. Argimon, W. Y. Zhang, X. W. Yang, I. B. Jeffery, J. C. Cooney, T. F. Kagawa, W. J. Liu, Y. Q. Song, E. Salvetti, A. Wrobel, P. Rasinkangas, J. Parkhill, M. C. Rea, O. O'Sullivan, J. Ritari, F. P. Douillard, R. P. Ross, R. F. Yang, A. E. Briner, G. E. Felis, W. M. de Vos, R. Barrangou, T. R. Klaenhammer, P. W. Caufield, Y. J. Cui, H. P. Zhang and P. W. O'Toole (2015) Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun 6: 8322.

Talavera, G. and J. Castresana (2007) Improvement of phylogenies after

removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56: 564-577.

Tanasupawat, S., O. Shida, S. Okada and K. Komagata (2000) *Lactobacillus acidipiscis* sp. nov. and *Weissella thailandensis* sp. nov., isolated from fermented fish in Thailand. Int J Syst Evol Microbiol 50 Pt 4: 1479-1485.

Tatusov, R. L., A. R. Mushegian, P. Bork, N. P. Brown, W. S. Hayes, M. Borodovsky, K. E. Rudd and E. V. Koonin (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr Biol 6(3): 279-291.

Tatusov, R.L., E. V. Koonin and D. J. Lipman (1997) A genomic perspective on protein families. Science 278: 631-637.

Tatusova, T., M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky and J. Ostell (2016) NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44(14): 6614-6624.

Taweechotipatr, M., C. Iyer, J. K. Spinler, J. Versalovic and S. Tumwasorn (2009) *Lactobacillus saerimneri* and *Lactobacillus ruminis*: novel human-derived probiotic strains with immunomodulatory activities. FEMS Microbiol Lett 293(1): 65-72.

Tettelin, H., D. Riley, C. Cattuto and D. Medini (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11(5): 472-477.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli and C. M. Fraser (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102(39): 13950-13955.

Toh, H., K. Oshima, A. Nakano, M. Takahata, M. Murakami, T. Takaki, H. Nishiyama, S. Igimi, M. Hattori and H. Morita (2013) Genomic adaptation of the *Lactobacillus casei* group. PLoS ONE 8(10): e75073.

Ussery, D. W., T. M. Wassenaar and S. Borini (2009) Microbial communities: core and pan-genomics. In: Computing for Comparative Microbial Genomics: Computational Biology (Vol. 8). Springer, London, UK.

van Dijk, E. L., H. Auger, Y. Jaszczyszyn and C. Thermes (2014) Ten years of next-generation sequencing technology. Trends Genet 30(9): 418-426.

van Heel, A. J., A. de Jong, M. Montalban-Lopez, J. Kok and O. P. Kuipers (2013) BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res 41(Web Server issue): W448-453.

Vancanneyt, M., G. Huys, K. Lefebvre, V. Vankerckhoven, H. Goossens and J. Swings (2006) Intraspecific genotypic characterization of *Lactobacillus rhamnosus* strains intended for probiotic use and isolates of human origin. Appl Environ Microbiol 72(8): 5376-5383.

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

Velez, M. P., M. I. Petrova, S. Lebeer, T. L. Verhoeven, I. Claes, I. Lambrichts, S. Tynkkynen, J. Vanderleyden and S. C. De Keersmaecker (2010) Characterization of MabA, a modulator of *Lactobacillus rhamnosus* GG adhesion and biofilm formation. FEMS Immunol Med Microbiol 59(3): 386-398.

Ventura, M., S. O'Flaherty, M. J. Claesson, F. Turroni, T. R. Klaenhammer, D. van Sinderen and P. W. O'Toole (2009) Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat Rev Microbiol 7(1): 61-71.

Vernikos, G., D. Medini, D. R. Riley and H. Tettelin (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23: 148-154.

Voelkerding, K. V., S. A. Dames and J. D. Durtschi (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4): 641-658.

von Ossowski, I. (2017) Novel molecular insights about lactobacillar sortase-dependent piliation. Int J Mol Sci 18(7): 1551.

von Ossowski, I., R. Satokari, J. Reunanen, S. Lebeer, S. C. De Keersmaecker, J. Vanderleyden, W. M. de Vos and A. Palva (2011) Functional characterization of a mucus-specific LPXTG surface adhesin from probiotic *Lactobacillus rhamnosus* GG. Appl Environ Microbiol 77(13): 4465-4472.

Wagner, A., C. Lewis and M. Bichsel (2007) A survey of bacterial insertion sequences using IScan Nucleic Acids Res 35(16): 5284-5293.

Wass, M. N. and M. J. Sternberg (2008) ConFunc: functional annotation in the twilight zone. Bioinformatics 24(6): 798-806.

Wilson, D., V. Charoensawan, S. K. Kummerfeld and S. A. Teichmann (2008) DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 36(Database issue): D88-92.

Wuyts, S., S. Wittouck, I. De Boeck, C. N. Allonsius, E. Pasolli, N. Segata and S. Lebeer (2017) Large-scale phylogenomics of the *Lactobacillus casei* group highlights taxonomic inconsistencies and reveals novel clade-associated features. mSystems 2(4): e00061-17.

Xiao, J. F., Z. W. Zhang, J. Y. Wu and J. Yu (2015) A brief review of software tools for pangenomics. Genomics Proteomics & Bioinformatics 13(1): 73-76.

Yadav, A. K., A. Tyagi, A. Kumar, S. Panwar, S. Grover, A. C. Saklani, R. Hemalatha and V. K. Batish (2017) Adhesion of lactobacilli and their anti-infectivity potential. Crit Rev Food Sci Nutr 57(10): 2042-2056.

Yan, F., H. Cao, T. L. Cover, R. Whitehead, M. K. Washington and D. B. Polk (2007) Soluble proteins produced by probiotic bacteria regulate intestinal epithelial cell survival and growth. Gastroenterology 132(2): 562-575.

Yan, F., H. W. Cao, T. L. Cover, M. K. Washington, Y. Shi, L. S. Liu, R. Chaturvedi, R. M. Peek, K. T. Wilson and D. B. Polk (2011) Colon-specific delivery of a probiotic-derived soluble protein ameliorates intestinal inflammation in mice through an EGFR-dependent mechanism. J Clin Invest 121(6): 2242-2253.

Yan, F., L. Liu, P. J. Dempsey, Y. H. Tsai, E. W. Raines, C. L. Wilson, H. Cao, Z. Cao, L. Liu and D. B. Polk (2013) A *Lactobacillus rhamnosus* GG-derived soluble protein, p40, stimulates ligand release from intestinal epithelial cells to transactivate epidermal growth factor receptor. J Biol Chem 288(42): 30742-30751.

Yan, F. and D. B. Polk (2012) Characterization of a probiotic-derived soluble protein which reveals a mechanism of preventive and treatment effects of probiotics on intestinal inflammatory diseases. Gut Microbes 3(1): 25-28.

Yu, X., A. Jaatinen, J. Rintahaka, U. Hynönen, O. Lyytinen, R. Kant, S. Åvall-Jääskeläinen, I. von Ossowski and A. Palva (2015) Human gut-commensalic *Lactobacillus ruminis* ATCC 25644 displays sortase-assembled surface piliation: phenotypic characterization of its fimbrial operon through *in silico* predictive analysis and recombinant expression in *Lactococcus lactis*. PLoS ONE 10(12): e0145718.

Yu, X., S. Åvall-Jääskeläinen, J. Koort, A. Lindholm, J. Rintahaka, I. von Ossowski, A. Palva and U. Hynönen (2017) A comparative characterization of different host-sourced *Lactobacillus ruminis* strains and their adhesive, inhibitory, and immunomodulating functions. Front Microbiol 8: 657.

Yun, J. H., D. S. Yim, J. Y. Kang, B. Y. Kang, E. A. Shin, M. J. Chung, S. D. Kim, D. H. Baek, K. Kim and N. J. Ha (2005) Identification of *Lactobacillus ruminus* SPM0211

isolated from healthy Koreans and its antimicrobial activity against some pathogens. Arch Pharm Res 28(6): 660-666.

Zheng, J., L. Ruan, M. Sun and M. Ganzle (2015) A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. Appl Environ Microbiol 81(20): 7233-7243.

Zhou, C. E., J. Smith, M. Lam, A. Zemla, M. D. Dyer and T. Slezak (2007) MvirDB: a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res 35(Database issue): D391-394.

Zhou, F. and Y. Xu (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 26(16): 2051-2052.

Zhou, M., J. Boekhorst, C. Francke and R. J. Siezen (2008) LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics 9: 173.

Zhou, Y., Y. Liang, K. H. Lynch, J. J. Dennis and D. S. Wishart (2011) PHAST: a fast phage search tool. Nucleic Acids Res 39(Web Server issue): W347-352.

Zotta, T., A. Ricciardi, E. Parente, A. Reale, R. G. Ianniello and D. Bassi (2016) Draft genome sequence of the respiration-competent strain *Lactobacillus casei* N87. Genome Announc 4(3): e00348.