Department of Biosciences
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

Computational genomics of lactobacilli

Matti Kankainen

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, for public examination in the lecture room 228 at Koetilantie 5, Viikki, on 24 April 2015, at 12 o'clock noon.

Finland 2015

Supervisor Professor Liisa Holm
Department of Biological and Environmental Sciences,
Faculty of Biosciences
University of Helsinki
Helsinki, Finland

Reviewers Professor Mauno Vihinen
Department of Experimental Medical Science
Lund University
Lund, Sweden

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

Opponent Adjunct Professor Laura Elo
Department of Mathematics and Statistics
University of Turku
Turku, Finland

Custos Professor Liisa Holm
Department of Biological and Environmental Sciences,
Faculty of Biosciences
University of Helsinki
Helsinki, Finland

Published in Dissertationes Scholae Doctoralis Ad Sanitatem Investigandam Universitatis Helsinkiensis

ISBN 978-951-51-0886-9 (pbk.)
ISBN 978-951-51-0887-6 (PDF)
ISSN 2342-3161 (print)
ISSN 2342-317X (online)

Hansaprint
Vantaa 2015

"*If you wish to make an apple pie from scratch,*
you must first invent the universe." - Carl Sagan

Contents

Abstract

Lactobacilli are gram-positive lactic acid bacteria (LAB) and have important implications for food production and preservation as well as human health and wellbeing. These bacteria occupy various niches in and on the human body, such as the gastrointestinal, respiratory, and urogenital tracts, and have been used for centuries in the fermentation of dairy products, the pickling of vegetables, baking, and curing fish, meats, and sausages. Recently, the use of lactobacilli as biotherapeutic agents has attracted interest. However, the molecular basis of host-microbe interactions, food production abilities and beneficial effects on health of lactobacilli are not well understood and deserve more research. In this thesis research, bioinformatics approaches were developed for genome-scale protein function classification, and the genetic composition of two also human-associated *Lactobacillus* species was determined by means of genome sequencing and computational genomics. Taken together, the results of these analyses illustrated that genome sequencing and computational genomics represent valuable approaches to the study of lactobacilli and to understanding their physiology. Furthermore, these methods provide effective means of identifying lactobacillar components that are involved in host-interactions.

Protein function prediction is one of the most crucial tasks of any genome sequencing project. In this thesis, two bioinformatics software tools were developed for the systematic analysis of protein function that are more advanced than current methods in several respects. The first automatic function prediction method, called LOCP, was developed to fulfil the need for rapid and accurate genome-wide identification of putative pilus operons in gram-positive bacteria. The computational resource was designed to support for both nucleotide sequence input or annotated bacterial protein sequence data and introduced a novel approach that combines similarity searches and statistical detection of sortase- and pilin-motif enriched regions for the prediction of putative pilus operons in gram-positive genomes. Markedly, the tool identified all genuine pilus gene clusters from the test genome sequences and made in the benchmarking test no false predictions. The second bioinformatics tool disclosed an improved homology-based function prediction solution and was created to offer an effective approach to the large-scale computational annotation of uncharacterised bacterial protein sequences. Compared to existing solutions for homology-based function prediction, BLANNOTATOR groups sequences that are found using sequence similarity searches into subsets according to their biological function and uses a set of matches with consistent functional information as the basis for annotation transfer to the query sequence instead of relying on a single match or all matches as many competing tools do. This procedure improved the functional classification substantially, producing consistent results and facilitating comparisons among various organisms. Overall, the two tools developed in this thesis are important additions to the current repertoire of function classification systems that are applicable to bacterial proteins and provided a novel means to classify bacterial proteins at the genome level. Most importantly, annotation accuracy was high, and both tools provided information that otherwise might have been ignored or considered too labour intensive to find.

Lactobacillus rhamnosus GG is a probiotic bacterium that has a long history of safe use in foods and that has a well-documented beneficial effect on human health. To gain

insight into its physiology and to elucidate the lactobacillar components that are involved in interactions with the host, the genomes of *L. rhamnosus* GG and its closely related dairy isolate *L. rhamnosus* LC705 were sequenced and analysed. Although the two genomes were shown to exhibit a high degree of synteny, altogether, nine regions of diversity punctuated the colinearity between the two genomes. The five GG-specific diversity regions included a number of genes encoding bacteriophage components and other genes that are implicated in sugar transport, metabolism, and exopolysaccharide (EPS) biosynthesis. In addition, genes for three pilus subunits and a pilin-dedicated sortase were identified in one of the diversity regions. Importantly, the presence of the pili on the cell surface of *L. rhamnosus* GG was confirmed and one of the GG-specific pilins was shown to be instrumental for the mucus interaction of *L. rhamnosus* GG. The diversity regions in LC705 strain encoded rhamnose, ribose, and maltose transporters and fermentative capacities that are missing from GG, thereby extending its metabolic versatility beyond that of GG and enabling LC705 to survive in a range of environments.

The genetic makeup of *Lactobacillus crispatus* was in turn explored by sequencing and analysing the genome of *L. crispatus* ST1 and performing a comparative genomic analysis of the chicken isolate ST1 and nine other *L. crispatus* genomes. The analyses revealed a rather compact genome and indicated that the genetic diversity present within *L. crispatus* has not yet been captured exhaustively. Specifically, the genomes of ST1 and nine vaginal strains were predicted to contain a pan-genome of 3,929 gene families, of which 1,224 families made up the *L. crispatus* conserved core. Mathematical extrapolations of these data to an infinite number of strains suggested that the core genome reaches a plateau of approximately 1,116 gene families and that the pan-genome doubles in size with the addition of the next 107 genomes, illustrating the value of sequencing many isolates. Interestingly, the comparison of protein-coding gene (CDS) contents revealed differences that are potentially relevant for the genetic adaptation of *L. crispatus* to different habitats, such as the existence of different EPS biosynthesis gene clusters in different strains and the Type II CRISPR (clustered regularly interspaced palindromic repeats)-Cas (CRISPR-associated) invader defence system that is specific to vaginal strains. In contrast, adhesin genes that are potentially involved in the exclusion and displacement of the bacterial vaginosis-associated species *Gardnerella vaginalis* were predicted to be present in the *L. crispatus* core genome, suggesting that all *L. crispatus* strains might have the potential to prevent bacterial vaginosis.

List of original publications

This thesis is based on the following publications:

I **Plyusnin I, Holm L, Kankainen M. (2009)** LOCP--locating pilus operons in gram-positive bacteria. *Bioinformatics* **25**:1187-1188.

II **Kankainen M, Ojala T, Holm L. (2012)** BLANNOTATOR: enhanced homology-based function prediction of bacterial proteins. *BMC bioinformatics* **13**:33

III **Kankainen M, Paulin L, Tynkkynen S, von Ossowski I, Reunanen J, Partanen P, Satokari R, Vesterlund S, Hendrickx APA, Lebeer S, De Keersmaecker SC, Vanderleyden J, Hämäläinen T, Laukkanen S, Salovuori N, Ritari J, Alatalo E, Korpela R, Mattila-Sandholm T, Lassig A, Hatakka K, Kinnunen KT., Karjalainen H, Saxelin M, Laakso K, Surakka A, Palva A, Salusjärvi T, Auvinen P, de Vos WM. (2009)** Comparative Genomic Analysis of *Lactobacillus rhamnosus* GG Reveals Pili Containing a Human Mucus-Binding Protein. *Proc Natl Acad Sci U S A* **106**:17193-17198.

IV **Ojala T, Kuparinen V, Koskinen JP, Alatalo E, Holm L, Auvinen P, Edelman S, Westerlund-Wikström B, Korhonen TK, Paulin L, Kankainen M. (2010)** Genome sequence of *Lactobacillus crispatus* ST1. *J Bacteriol* **192**:3547-3548.

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

The publications are referred to in the text by their roman numerals.

Abbreviations

Cas	CRISPR-associated protein
CDS	coding DNA sequence
COG	clusters of orthologous groups
Contig	contiguous sequenced region
CRISPR	clustered regularly interspaced palindromic repeat
DAG	directed acyclic graph
DE	description line
EC	enzyme commission
emPCR	emulsion polymerase chain reaction
EPS	exopolysaccharide
GIT	gastrointestinal tract
GO	gene ontology
GRAS	generally regarded as safe
HMM	hidden Markov model
HMP	human microbiome project
LAB	lactic acid bacteria
MGE	mobile genetic element
ncRNA	noncoding RNA
NGS	next generation sequencing
OLC	overlap-layout-consensus
ORF	open-reading frame
PCR	polymerase chain reaction
PTS	phosphotransferase system
RBH	reciprocal best hit
RBS	ribosomal binding site
RSD	reciprocal smallest distance
S-layer	surface layer
Sn	sensitivity
Sp	specificity
SVM	support vector machine
TC	transporter classification system
WGS	whole-genome shotgun

1 Review of the literature

This chapter introduces the objectives and scope of this dissertation and includes three sections: 1.1) a methodological review of DNA sequencing, sequence assembly and scaffolding, genomic feature calling, and functional classification techniques that are suitable for bacterial whole-genome sequencing; 1.2) an introduction to the general characteristics, phylogeny, taxonomy, and ecological distribution of lactobacilli; and 1.3) an overview of the existing literature on lactobacilli genomics.

1.1 Bacterial whole-genome sequencing

Whole-genome sequencing is the process of determining the order of nucleotides in the DNA molecules of an organism. Beginning with the first experimental determinations of DNA sequence using a location-specific primer extension strategy (Wu & Kaiser, 1968; Wu & Taylor, 1971) and following the introduction of the more practical 'plus and minus' (Sanger & Coulson, 1975), chemical degradation (Maxam & Gilbert, 1977), and chain-termination (Sanger *et al.*, 1977) sequencing strategies in the late 1970s, DNA sequencing has advanced to the point at which genomes can be sequenced rapidly and affordably (Mardis, 2008). This astounding growth in DNA sequencing capacity and speed has laid the foundation for determining the genomes of tens of thousands of bacterial (Figure 1) and thousands of other organisms in less than two decades and has permanently altered our understanding of microbial life.

Currently, the whole-genome shotgun (WGS) approach is the most widely used method for determining genome sequences (Anderson, 1981). When applied to microbial organisms (Figure 2), the initial standard approach is to grow the microbe from a single colony and then isolate a sufficient quantity of DNA for library construction. Depending on the protocol adopted, the amount used ranges from a few nanograms to tens of micrograms of DNA (Loman *et al*., 2012). In the second step, the DNA is fragmented into random overlapping DNA sequences that are used as templates for amplification and are sequenced from one (single-end reads) or both (paired reads) ends. For many years, the fragments were inserted into plasmids for cloning and then sequenced using the Sanger method. However, massively parallel sequencing methods are the current standard approach (Mardis, 2008). The third step of the process involves identifying overlaps between the reads and deriving contiguous consensus sequences from the reads, termed contigs. In most cases, however, full genome sequences cannot be built from a single shotgun read library. To improve the contiguity, paired reads from multiple libraries of different insert sizes are utilised or WGS reads are combined with mate-pair reads resulting from a process in which the ends of long (3-, 6-, or 8-kb) DNA fragments are brought together by circularisation and then sequenced across the ligation regions (Collins & Weissman, 1984). In addition, dedicated gap-closing techniques can be used to improve assembly contiguity (Figure 2). Previously, the closure of genomes was deemed

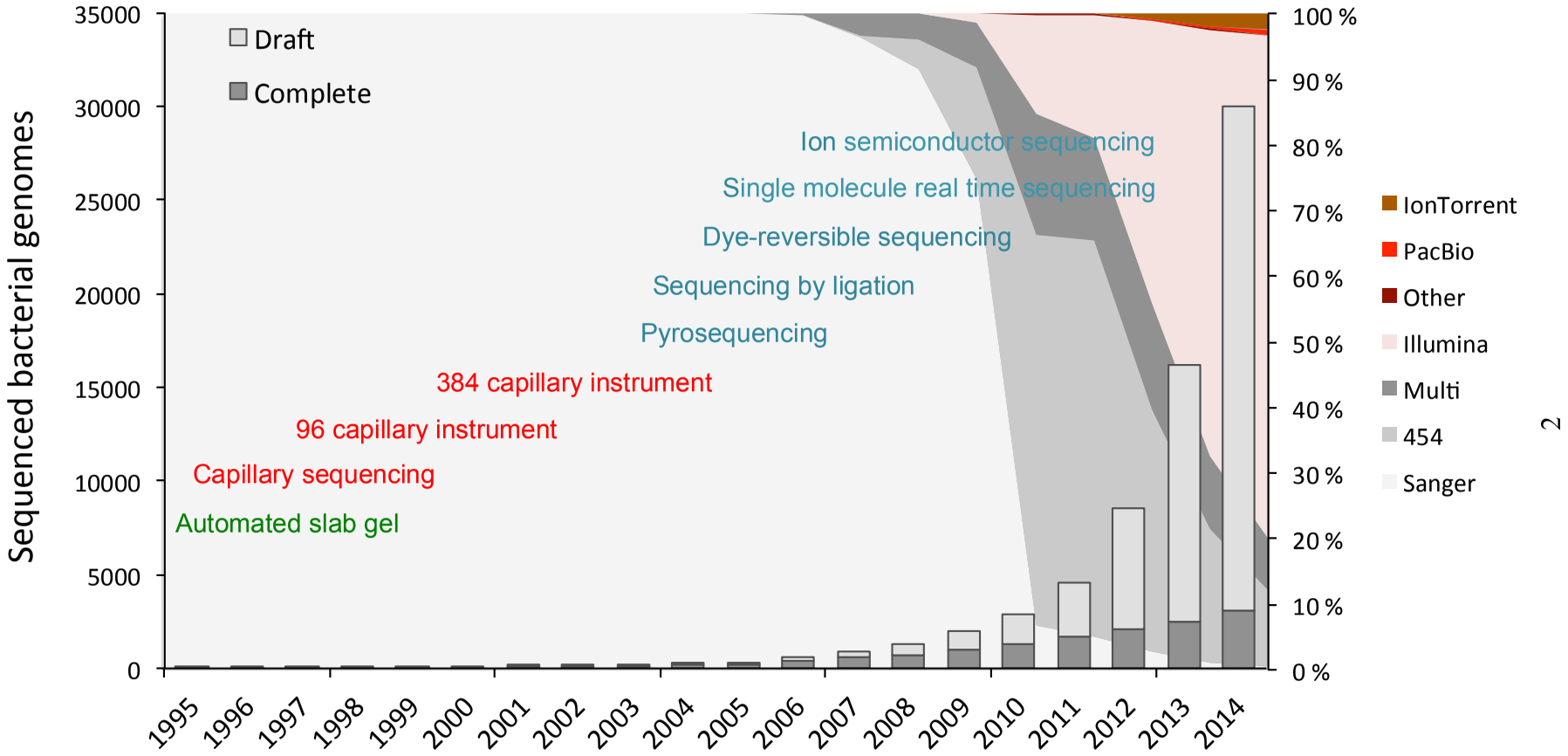


Figure 1. The cumulative number of bacterial genome entries in NCBI. Dark and light bars indicate the number of complete and draft genome assemblies, respectively. Background surfaces indicate the fraction of genomes that have been sequenced using a particular sequencing technique. Others consist of genomes sequenced using the SOLiD or Complete Genomics platform. Labels indicate advances involving gel-based systems (green), capillary sequencing (red), and massively parallel DNA sequencing (blue).

necessary by the scientific community; however, this finishing phase is no longer routine (Figure 1) because of the prohibitive cost and labour required by the gap-closing phase. In future, the emergence of sequencing techniques that produce reads over 10 kb in length might provide a novel means to close genomes. The final step of the standard protocol is genome annotation, followed by analysis of the resulting information.

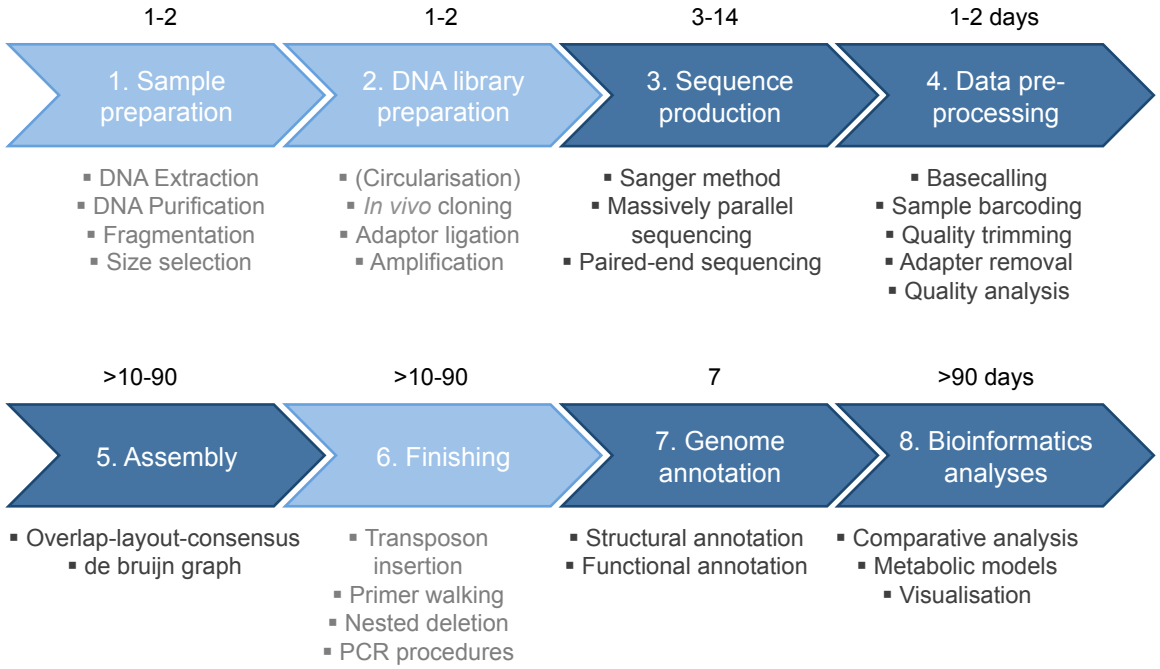


Figure 2. The principal steps involved in obtaining a bacterial genome sequence. In Sanger sequencing, genomic DNA is extracted from a single colony and then fragmented, ligated into a plasmid vector, and used to transform *Escherichia coli*. For each sequencing reaction, a single *E. coli* colony is selected, and the plasmid DNA is isolated. In massively parallel sequencing, common adaptors are ligated to fragmented genomic DNA, which is then subjected to PCR-based amplification and massively parallel sequencing. Following read preprocessing and assembly, unrevealed genome regions can be fixed using a variety of methods, wherein adjacent contigs are detected, and the gap between them is sequenced. Fixing gaps produces a finished genome. Genome annotation includes the identification of sequence features and the subsequent association of biological information with these features. These analyses are often followed by other computational analyses, such as whole-genome comparison, phylogenetic analysis, and metabolic network reconstruction.

1.1.1 DNA sequencing technologies

Six DNA sequencing technologies are currently used (Table 1). The Sanger method is the oldest of these and has been the workhorse of DNA sequencing for over 30 years (Sanger *et al.*, 1977; Mardis, 2008). In essence, the Sanger method uses mixtures of deoxy-

Table 1. Characteristics of widely-used DNA sequencing instruments (Liu *et al.*, 2012; Quail *et al.*, 2012). Genome price, genome coverage and number of genomes were computed assuming a bacterial genome of 3 Mb and a sequencing depth of ×40.

	Life technologies / Sanger 3730xl	Roche / 454's GS FLX+	Illumina / Hiseq2500	Life technologies / SOLiD 5500xl wildfire	Pacific Biosciences / PacBio RS II	Life technologies / IonTorrent PII
Released	1996	2005	2012	2006	2010	2014
Template preparation	PCR	Emulsion PCR on bead surface	BridgePCR on glass surface	Emulsion / Wildfire PCR	Single molecule	Emulsion PCR on bead surface
Sequencing chemistry	Dideoxy chain termination	Pyrosequencing	Reversible dye terminators	Sequencing by ligation and two-base coding	Sequencing by synthesis	Ion semiconductor sequencing
Machine cost ($)	95000	500000	740000	665000	700000	149000
Max read length (bases)	900	1000	125+125	50+50	11500	100
Fragments per run (M)	0.000096	1	4000	2400	0.04	320
Run time (h)	1	23	144	240	3	4
Sequence yield (Gb/run)	0.0000864	0.7	1000	240	0.375	32
Sequence yield (Mb/h)	0.09	30	6944	1000	125	8000
Accuracy (%)	99.999	99.997	98.000	98.000	87.140	98.290
Operating cost ($/Mb)	1111	8.86	0.03	0.02	1.07	0.01
Paired-end / Mate pairs / Multiplex	Yes / Yes / No	No / Yes / Yes	Yes / Yes / Yes	Yes / Yes / Yes	No / No / Yes	No / Yes / Yes
Pros	Read length, accuracy	Read length, rate of substitution errors	Throughput, operating costs	Accuracy, operating costs	Read length	Machine cost, no use of optics or fluorescence
Cons	Operating costs, throughput	Homopolymer errors, operating costs	Substitution errors	Color space data, read length	Operating costs, throughput, error rate	Homopolymer errors, throughput
Genome price ($)	133333.3	1062.9	4.6	2.5	208.7	47.4
Genome coverage (×) / Number of genomes	0 / 0	233 / 5	333333 / 8333	80000 / 2000	125 / 3	10667 / 266

nucleotides and chain-terminating dideoxy-nucleotides to generate copies from the template that differ in length from each other by one nucleotide. In the process, the DNA sample is divided into four DNA sequencing reactions, which contain all four normal nucleotides, DNA polymerase, and one of four chain-terminating dideoxy-nucleotides in low amounts. As DNA synthesis progresses, DNA polymerase adds nucleotides to the chain. However, the occasional incorporation of a chain-terminating nucleotide into the strand causes DNA polymerase to cease DNA extension, resulting in fragments of different lengths (Sanger *et al.*, 1977). These DNA fragments are then denatured and separated according to mass using gel- or capillary electrophoresis, and the species of terminal base present is identified by exciting the fluorophore attached to the primer (Smith *et al.*, 1986) or chain-terminating base (Prober *et al.*, 1987) using a laser. Notably, incremental improvements to the Sanger method have rendered the technique advantageous for a number of applications, including the sequencing of long polymerase chain reaction (PCR) products and closing genomes by PCR.

Recently, a family of novel DNA sequencing technologies has supplanted the Sanger method (Mardis, 2008). These next-generation sequencing (NGS) platforms can process millions of DNA molecules in parallel and enable inexpensive and rapid sequencing, although at the expense of lower read length and accuracy. During template preparation, millions of template clusters, each comprising a large number of copies of a given template DNA molecule, are created using emulsion PCR (emPCR, Dressman *et al.*, 2003) or bridge amplification (Adessi *et al.,* 2000). Template aggregations are then sequenced in parallel using pyrosequencing (Margulies *et al.*, 2005), sequencing by ligation (Shendure *et al.*, 2005), sequencing by synthesis (Bentley, 2006; Bentley *et al.*, 2008), or ion semiconductor sequencing (Rothberg *et al.*, 2011). A notable exception is the single-molecule real-time sequencing system, which was developed by Pacific Bioscience (Eid *et al.*, 2009) and involves sequencing single-template DNA molecules using DNA polymerases that are immobilised onto a zero-mode waveguide array. The key differences between the six sequencing technologies relate to the number and length of the reads produced (Table 1). In general, these two characteristics are negatively correlated, and technologies that provide long read lengths produce fewer data at higher cost than short-read sequencing instruments. However, the total sequence output from even the lowest capacity NGS instruments is far greater than the amount of sequence data needed to disclose a single bacterial genome; thus, these methods are an appealing choice for projects involving few isolates and benefiting from long contigs. In contrast, the Illumina and SOLiD platforms yield very large volumes of sequence data (Table 1) and are an affordable choice for large whole-genome projects that accommodate fragmented genome representations and for re-sequencing projects.

In terms of accuracy and error profiles, small but significant differences exist among the platforms (Mardis, 2008; Liu *et al.*, 2012; Quail *et al.*, 2012). The 454 and Ion Torrent platforms produce more nucleotide over- and under-calls than the other NGS platforms. These errors emerge from the methodology, which introduces all subsequent bases of one species at once, and from the difficulty in resolving the number of incorporated bases based on signal intensities. Especially troublesome are homopolymers of four or more bases (Voelkerding *et al.*, 2009). In contrast, the Illumina platform suffers from phasing,

fading, and crosstalk-generated noise (Erlich & Mitra, 2008). Phasing noise results from incorporation of none or more than one nucleotide during sequencing cycles and introduces lagging and leading nascent strands transmitting a mixture of signals; in contrast, fading noise arises from an exponential decay of the fluorescent signal intensity as a function of cycle number. The third noise factor arises from fluorophore crosstalk (Sheikh & Erlich, 2012). Regarding whole-genome sequencing, errors attributed to the Ion Torrent and 454 platforms are more fatal because the handling of insertion and deletion errors during genome assembly requires the use of computationally intensive gap-alignments. Furthermore, insertion and deletion errors in the final contigs hamper gene calling and can cause fragmentation of predicted open reading frames (ORFs). The SOLiD platform, however, is considered accurate (Li *et al.*, 2012). Indeed, the investigation of each dinucleotide by two ligation reactions and the requirement that adjacent colour-calls must agree guarantees high accuracy. Nonetheless, the use of a colour-space coding scheme complicates data analysis, *de novo* assembly, and integration of the SOLiD data into other genomic resources (Li *et al.*, 2012).

1.1.2 Preprocessing of sequencing data

Modern DNA sequencing instruments generate large amounts of data that have to be preprocessed in several steps to convert them to a usable form (Figure 2). The recovery of human-readable read sequences from sequencing instrument data is referred to as base calling and is the typical first step toward usable data. It includes the transformation of intensity signals to nucleotide calls and the assignment of quality scores (indicating the reliability of the call) to each base (Sheikh & Erlich, 2012). In most cases, quality scores are reported in terms of logarithmically linked error probabilities termed Phred scores (Ewing & Green, 1998). Exceptions are the 454 and Ion Torrent error probabilities, which represent the likelihood that a base is an overcall and divide single quality values between two or more bases.

The adjustment of signal data for platform-specific anomalies is another important task that is performed by base callers (Sheikh & Erlich, 2012). Typically, this step is addressed by the use of a list of signal processing techniques, each tackling a specific error and error source. For example, most Illumina base-calling algorithms correct for fluorescent decay, fluorophore crosstalk, and errors caused by the incomplete removal or incorporation of reversible terminators (Kao *et al.*, 2009; Erlich & Mitra, 2008; Kircher *et al.*, 2009), whereas Sanger data is corrected for shifts in peak locations, fluorophore crosstalk, and background noise (Ewing *et al.*, 1998). Typically, a vendor-supplied base-calling method is used, because the base-calling process can require substantial amounts of processing time. However, third party programs have been developed as an alternative to vendor software and have been shown to improve the accuracy of base calls. Some popular third-party base-calling approaches for the Illumina (Rougemont *et al.*, 2008; Erlich & Mitra, 2008; Kao *et al.*, 2009; Kircher *et al.*, 2009), SOLiD (Wu *et al.*, 2010) and Roche 454 sequencing platforms (Quinlan *et al.*, 2008; Beuf *et al.*, 2012) are listed in Table 2 and Appendix Table 1.

Table 2. Bioinformatic resources that are commonly used in the study of bacterial genomes. More information on the listed software and databases is available in Appendix Table 1.

Category	Software
Base callers (Sanger)	Phred (Ewing & Green, 1998), KB Basecaller
Base callers (NGS)	Rolexa (Rougemont *et al.,* 2008), Alta-Cyclic (Erlich & Mitra, 2008), BayesCall (Kao *et al.,* 2009), Ibis (Kircher *et al.,* 2009), Pyrobayes (Quinlan *et al.,* 2008), HPCall (Beuf *et al.,* 2012), Rsolid (Wu *et al.*, 2010), All Your Base (Massingham & Goldman, 2012), BM-BC (Ji *et al.*, 2012).
Quality analysis and read manipulation	FastQC, PrinSeq (Schmieder & Edwards, 2011), BIGpre (Zhang *et al.*, 2011), Cutadapt (Martin, 2011), fastx, Staden package (Staden *et al.,* 1999), NGS QC Toolkit (Patel & Jain, 2012), NARWHAL (Brouwer *et al.*, 2012)
Read correction	Reptile (Yang *et al.,* 2010), HiTEC (Ilie *et al.*, 2011), ECHO (Kao *et al.*, 2011), Hybrid-SHREC (Salmela, 2010)
Greedy assemblers	SSAKE (Warren *et al.,* 2007), VCAKE (Jeck *et al.,* 2007)
Overlap-based genome assembly	Newbler (Margulies *et al.*, 2005), EDENA (Hernandez *et al.,* 2008), SGA (Simpson & Durbin, 2012), MIRA (Chevreux *et al.*, 2004)
De Bruijn graph based genome assembers	SPAdes (Bankevich *et al.,* 2012), ALL-PATHS (Butler *et al.,* 2008), SOAPdenovo (Li *et al.,* 2010), Velvet (Zerbino & Birney, 2008), ABySS (Simpson *et al.,* 2009), MaSuRCA (Zimin *et al.*, 2013), RAY (Boisvert *et al.*, 2010)
Reference-based genome assemblers	VAAL (Nusbaum *et al.,* 2008), Amos-Cmp (Pop *et al.*, 2004b)
Scaffolders	Bambus (Pop *et al.*, 2004a), SSPACE (Boetzer *et al.*, 2011), SOMA (Nagarajan *et al.*, 2008), OSLay (Richter *et al.*, 2007), BACCardI (Bartels *et al.,* 2005), PAGIT (Swain *et al.*, 2012)
Assembly integrators	Minimus2 (Sommer *et al.,* 2007), MAIA (Nijkamp, *et al.*, 2010), GAA (Yao *et al.,* 2012)
Ab initio CDS predictors	Glimmer (Delcher *et al.,* 2007), GeneMark (Besemer *et al.,* 2001), EasyGene (Larsen & Krogh, 2003), Prodigal (Hyatt *et al.,* 2010), ZCURVE (Guo *et al.,* 2003)
Evidence-base CDS predictors	ORPHEUS (Frishman *et al.,* 1998), CRITICA (Badger & Olsen, 1999)
CDS model integrators	Reganor (McHardy *et al.,* 2004), YACOP (Tech & Merkl, 2003)
CDS model refinement tools	GenePRIMP (Pati *et al.*, 2010), Mugsy-Annotator (Angiuoli *et al.*, 2011), ORFCor (Klassen & Currie, 2013), MICheck (Cruveiller *et al.*, 2005)
ncRNA predictors	RNAmotif (Macke *et al.*, 2001), RNAmmer (Lagesen *et al.*, 2007), QRNA (Rivas & Eddy, 2001), RNAz (Washietl *et al.*, 2005), Aragorn (Laslett & Canback, 2004), tRNAscan-SE (Lowe & Eddy, 1997), Infernal (Nawrocki *et al.*, 2009), SRP-scan (Regalia *et al.*, 2002), Bcheck (Yusuf *et al.*, 2010), CMFinder (Yao *et al.*, 2006)
Intrinsic terminators	TransTermHP (Kingsford *et al.*, 2007), RNIE (Gardner *et al.*, 2011)

CRISPR arrays	CRT tool (Bland *et al.*, 2007), PILER-CR (Edgar, 2007), CRISPRFinder (Grissa *et al.*, 2007)
Repeats	REPuter (Kurtz & Schleiermacher, 1999), RepeatScout (Price *et al.*, 2005)
Insertion sequences	ISfinder database (Siguier *et al.*, 2006), IScan (Wagner *et al.*, 2007), ISsaga (Varani *et al.*, 2011)
Prophages	ACLAME (Leplae *et al.*, 2004), ProphageDB (Srividhya *et al.*, 2007), PHAST (Zhou *et al.*, 2011), Prophinder (Lima-Mendez *et al.*, 2008), PhiPsy (Akhter *et al.*, 2012), Phage_Finder (Fouts, 2006)
Genomic islands	SIGI-HMM (Waack *et al.*, 2006), IslandViewer (Langille & Brinkman, 2009), PAI-IDA (Tu & Ding, 2003), Alien_Hunter (Vernikos & Parkhill, 2006), IslandPick (Langille *et al.*, 2008)
Plasmids, integrative and conjugative elements, gene cassettes, integrons	ICEberg (Bi *et al.*, 2012), INTEGRALL (Moura *et al.*, 2009), ACID (Joss *et al.*, 2009), cBar (Zhou *et al.*, 2010)
Origin of replication	Ori-Finder (Gao & Zhang, 2008)
Sequence database search	BLAST and PSI-BLAST (Altschul *et al.*, 1997), FASTA (Pearson & Lipman, 1988), HMMER3 (Eddy, 2011)
Biological databases	GenBank (Benson *et al.*, 2013), UniProt (Bairoch *et al.*, 2008), PATRIC (Gillespie *et al.*, 2011), CharProtDB (Madupu *et al.*, 2012), Rfam (Griffiths-Jones *et al.*, 2003), COG (Tatusov *et al.*, 1997, Tatusov *et al.*, 2003), SEED (Overbeek *et al.*, 2005)
Protein signature databases	PFAM (Punta *et al.*, 2012), TigrFAM (Haft *et al.*, 2003), HAMAP (Lima *et al.*, 2009), Interpro (Hunter *et al.*, 2012), CDD (Marchler-Bauer *et al.*, 2011)
General function classification	FunCut (Abascal & Valencia, 2003), CLAN (Kunin & Ouzounis, 2005), Gotcha (Martin *et al.*, 2004), GOPET (Vinayagam *et al.*, 2006), PFP (Hawkins *et al.*, 2006), ConFunc (Wass & Sternberg, 2008), SIFTER (Engelhardt *et al.*, 2005), InterProScan (Zdobnov & Apweiler, 2001), Argot2 (Falda *et al.*, 2012), BLANNOTATOR (Study II), PANNZER (Koskinen *et al.*, 2015), Sma3s (Muñoz-Mérida *et al.*, 2014)
Advanced function classification	BAGEL3 (van Heel *et al.*, 2013), RASTA-Bacteria (Sevin & Barloy-Hubler, 2007), TADB (Shao *et al.*, 2011), ARGO (Scaria *et al.*, 2005), MvirDB (Zhou *et al.*, 2007), ARDB (Liu & Pop, 2009), DBD (Wilson *et al.*, 2008), antiSMASH (Medema *et al.*, 2011), Mist2 (Ulrich & Zhulin, 2010), LOCP (Study I)
Metabolism related genes	PRIAM (Claudel-Renard *et al.*, 2003), KAAS (Moriya *et al.*, 2007), TCDB (Saier *et al.*, 2006), CAZy (Cantarel *et al.*, 2009), MEROPS (Rawlings *et al.*, 2004), REBASE (Roberts *et al.*, 2010)
Context-based protein function prediction	ContextMirror (Juan *et al.*, 2008), String (von Mering *et al.*, 2005), Prolinks (Bowers *et al.*, 2004)
Ab initio protein function prediction	CSS-Palm (Ren *et al.*, 2008), DISIS (Ofran *et al.*, 2007), MetalDetector (Lippi *et al.*, 2008), VirulentPred (Garg & Gupta, 2008), SPAAN (Sachdeva *et al.*, 2005)

Subcellular location	PSORTb v3.0 (Yu *et al.*, 2010), tatP (Bendtsen *et al.*, 2005), LipoP (Rahman *et al.*, 2008), SignalP (Petersen *et al.*, 2011), TMHMM (Krogh *et al.*, 2001), LocateP (Zhou *et al.*, 2008), EffectiveT3 (Arnold *et al.*, 2009), CoBaltDB (Goudenège *et al.*, 2010)
Orthology prediction	RBH (Tatusov *et al.*, 1996), RSD (Wall *et al.*, 2003), RIO (Zmasek & Eddy, 2002), OrthoStrapper (Storm & Sonnhammer, 2002), InParanoid (Remm *et al.*, 2001), EggNOG (Jensen *et al.*, 2008), OrthoMCL (Li *et al.*, 2003), Protein cluster (Klimke *et al.*, 2009), OMA (Roth *et al.*, 2008), PoFF (Lechner *et al.*, 2014), morFeus (Wagner *et al.*, 2014)
Multiple sequence aligners	Muscle (Edgar, 2004), ProbCons (Do *et al.*, 2005), Clustal Omega (Sievers *et al.*, 2011)
Whole-genome aligners	MUMmer (Darling *et al.*, 2004; Darling *et al.*, 2010), TBA (Blanchette *et al.*, 2004), Pecan (Paten *et al.*, 2008), Mauve aligner (Rissman *et al.*, 2009), Gepard (Krumsiek *et al.*, 2007), BLASTatlas (Wassenaar *et al.*, 2010)
Prediction of constrained elements	GERP (Cooper *et al.*, 2005), SiPhy (Garber *et al.*, 2009)
Phylogenetic trees	PhyML (Guindon *et al.*, 2003), BioNJ (Gascuel, 1997), Phylip (Felsenstein, 1989)
Metabolic reconstruction	Pathway tools (Karp *et al.*, 2002), FMM (Chou *et al.*, 2009), MetaCyc (Krieger *et al.*, 2004), UniPathway (Morgat *et al.*, 2012), KEGG (Kanehisa *et al.*, 2004)
Annotation pipelines	IMG (Markowitz *et al.*, 2012), RAST (Aziz *et al.*, 2008), DOE-JGI MAP (Mavromatis *et al.*, 2009), CG pipeline (Kislyuk *et al.*, 2010), ERGO (Overbeek *et al.*, 2003), PGAAP, Broad Institute, BCM, JCVI (Tanenbaum *et al.*, 2010)
Visualization	Genome atlas (Wassenaar *et al.*, 2010), ACT (Carver *et al.*, 2005), Combo (Engels *et al.*, 2006), Circoletto (Darzentas, 2010)

Quality assessment is another important pre-processing step, which aims to pinpoint poor quality reads and typically includes the visualisation of base quality scores, sequence length distributions, and nucleotide distributions. In addition, sequence data can benefit from data cleaning and steps such as adapter and poor quality region trimming and the filtering of chimeric and short reads and other types of sequence artefacts. For example, the removal of bases with poor quality at the end of the reads has a positive impact on downstream analyses (Haridas *et al.*, 2011; Cox *et al.*, 2010). Adaptor cutting has also been shown to be advantageous and ensures that only the relevant part of the read is passed to and considered at the downstream analysis.

An alternative strategy for improving data quality is based on high base coverage and supposedly infrequent and random sequencing errors (Yang *et al.*, 2013). In this setting, sequencing errors are detected by aligning reads to a reference genome and by examining the alignment for uncommon base calls or, in its more generalised form, by decomposing reads into overlapping oligomers of the length *k* (*i.e., k*-mers) and identifying infrequent *k*-mers that resemble frequent *k*-mers (Pevzner *et al.*, 2001; Yang *et al.*, 2013). Typically, *k*-mers that occur only a few times in the data are considered in this spectral alignment approach to represent errors and are rectified by making a minimum number of nucleotide edits to the reads from which they emerged. At present, several spectral alignment methods have been developed that mainly differ with regard to choosing the coverage threshold for the identification of infrequent *k*-mers. However, this feature is important because overly low thresholds result in uncorrected errors, whereas overly high thresholds affect correct *k*-mers. Among various spectral alignment tools, Reptile (Yang *et al.*, 2010), HiTEC (Ilie *et al.*, 2011) and ECHO (Kao *et al.*, 2011) have been shown to reliably detect and correct erroneous reads (Yang *et al.*, 2013).

1.1.3 Genome assembly

Genome assembly is the process of assembling short reads into the largest possible continuous sequences, thus providing a representation of the expected genome (Pop, 2009; Flicek & Birney, 2009; Miller *et al.*, 2010). In essence, this process relies on the assumption that highly similar reads originate from the same position within a genome and involves joining individual overlapping reads into contigs (Pop, 2009; Flicek & Birney, 2009; Miller *et al.*, 2010). In addition, the process can include a separate scaffolding step to produce larger scaffold structures comprising an ordered set of contigs with intervening gaps representing DNA stretches that are not present in the reads. The sequential phases of current genome assembly methodology are illustrated in Figure 3.

The first phase of genome assembly is contig construction. Currently, this is achieved in most cases using algorithms that are based on either the overlap-layout-consensus (OLC) or the de Bruijn graph approaches (Pop, 2009; Flicek & Birney, 2009). Alternatively, genome sequences can be reconstructed using the greedy extension approach, in which the best-matching reads are iteratively joined together into contigs until no more reads or contigs can be joined (Miller *et al.*, 2010). Many early assemblers relied on the greedy approach, as do some modern ones (Warren *et al.*, 2007; Jeck *et al.*,

2007); however, this approach is no longer utilised due to the inherently local assembly process, which can become stuck at a local minimum if the sequence under assembly is extended with a read that would have been more beneficial to other joining operations (Pop, 2009; Miller *et al.*, 2010). Finally, contig building can rely on aligning reads to the reference sequence and on grouping reads by continuity (Pop *et al.*, 2004b; Nusbaum *et al.*, 2008). However, the comparative approach is valid only in the presence of a reference genome with substantial sequence similarity (≥90%) to the genome of interest (Pop *et al.*, 2004b). Table 2 and Appendix Table 1 provide partial lists of the tools that are currently available for genome assembly and scaffolding.

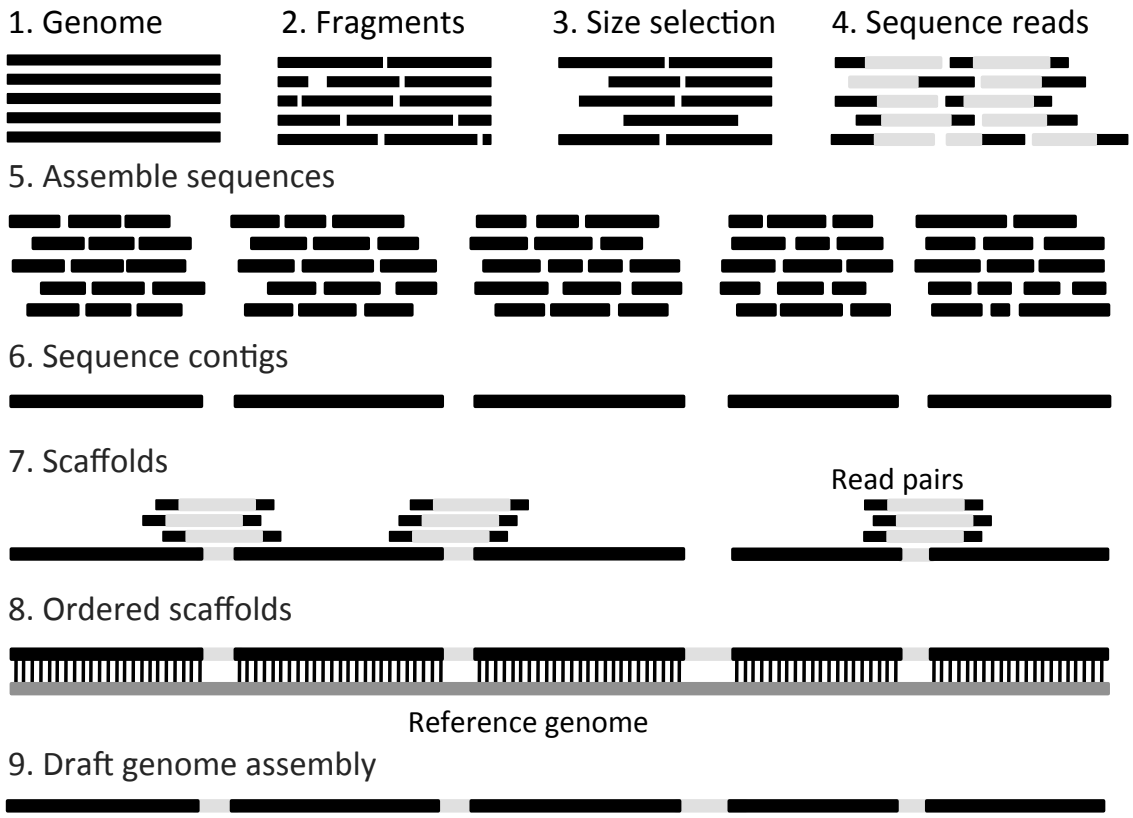


Figure 3. The sequential phases of current genome assembly methodology. Starting from a large amount of genomic DNA, the DNA is sheared (2) into random fragments, size selected (3), and amplified and sequenced from one or both ends (4). The sequencing reads are then assembled on the basis of sequence overlaps (5), thereby yielding sequence contigs (6). The contigs can then be oriented and ordered based on the read pairs that map to two contigs (7) or with the aid of reference genomes (8).

The OLC approach is one of the two main approaches used and is useful for the assembly of long reads (Pop, 2009; Miller *et al.*, 2010). The OLC process begins with an all-against-all read comparison. The reads and relationships between reads are then structured into a graph with a node for every read and an edge between any pair of reads

that overlap sufficiently well (Pop, 2009; Miller *et al.*, 2010). Using this structure, finding contigs becomes equivalent to finding a path through a graph that visits each node exactly once, termed a Hamiltonian circuit. The OLC approach is highly suitable for reads of varying length because it can employ all of the information obtained from long and short reads (Miller *et al.*, 2010). This approach captures repetitive regions in nodes with multiple connections, thus providing a means by which to exclude these regions and effectively handle sequencing errors. Regardless, identifying the Hamiltonian circuit is an intensive task, and storing each read in a separate node requires memory, thus making the OLC approach somewhat impractical for modern sequencing projects (Pop, 2009; Miller *et al.*, 2010; Flicek & Birney, 2009). An exception is the SGA assembly package, which exhibits significantly reduced memory requirements and computing time (Simpson & Durbin, 2012).

The second widely used assembly approach is known as the De Bruijn graph approach and is considered ideal for short read-length, high-coverage data (Pevzner *et al.*, 2001). This strategy relies on deconstructing reads into short *k*-mer fragments before assembling them into contigs with the help of De Bruijn graphs, in which nodes represents unique *k*-mers and an edge connects all node pairs that overlap by *k*-1 bases (Pop, 2009; Miller *et al.*, 2010; Flicek & Birney, 2009). In theory, the number of unique *k*-mers in the genome determines the number of nodes, thereby allowing the use of a small memory footprint for genomes of limited size and in the presence of repetitive elements. Moreover, De Bruijn-based assemblies are fast to compute because overlaps between the reads are implicitly captured by the graph rather than computed individually; furthermore, a linear-time algorithm exists for finding the Eulerian paths that visit each edge once (Fleischner, 1990). This approach is also considered useful for short-read data. Theoretically, 16-mers should yield reasonable assemblies, and larger *k*-mers should entail longer contigs; however, parameter optimisation tests have shown that the optimal performance is obtained using moderate (approximately 50) to large (approximately 100) *k* values (Peng *et al.*, 2012; Jünemann *et al.*, 2014) or a combination of multiple *k*-mers (Peng *et al.*, 2012; Bankevich *et al.*, 2012). The greatest disadvantage of the De Bruijn graph paradigm is its insensitivity to repetitive regions and read errors. The use of even large *k*-mers fails to resolve repeat regions, if repeats are longer than the given *k*-mer value. In the presence of such repeats, the graph will have branching vertices and contain multiple Eulerian paths. In these cases, continuity can however be improved by adding constrains to the graph and finding Eulerian superpaths that traverse the graph via predetermined sub-paths (Pevzner *et al.*, 2001; Pop, 2009). Erroneously called bases near read ends on the other hand result in dead-end "tips", whereas errors in the middle lead to branching nodes and alternate paths termed "bubbles" that start and terminate in the same nodes, each providing an equally good solution. To compensate for these problems, many assemblers employ the spectral alignment approach to correct reads before beginning the assembly (Pevzner *et al.*, 2001; Li *et al.*, 2010; Butler *et al.*, 2008). This approach has been shown to remove up to 66% of the *k*-mers (Li *et al.*, 2010). The concept of De Bruijn graphs is implemented in several assemblers; of these, ALL-PATHS (Butler *et al.*, 2008) and SPAdes (Bankevich *et al.*, 2012) have performed well in recent genome assembler surveys (Earl *et al.*, 2011; Magoc *et al.*, 2013).

Repetitive regions that are longer than the read length are problematic for genome assembly and cannot be resolved by simply investigating read overlaps regardless of the coverage depth. The position of these repeats and other unrevealed genome regions in the genome can, however, be approximated by supplementing the assembly process with a scaffolding phase in which individual contigs are joined into larger sequence structures that comprise contigs and the intervening gaps (Pop, 2009; Miller *et al.*, 2010). Chiefly, scaffolding exploits information that is derived from paired reads (libraries with insert sizes of approximately 300 to 1,000 bp) or mate-pair reads (libraries with insert sizes of 3-, 6-, or 8-kb; Figure 3). Pairs of contigs that are likely to reside side-by-side in the genome are detected by aligning reads to contigs and by identifying read pairs that match different contigs. The optimal order and orientation of contigs are then solved using heuristic or graph-based approaches that rely on majority voting from a large number of read pairs (Pop, 2009; Miller *et al.*, 2010). Importantly, the length of the gaps can be estimated based on the positions of the paired reads in contigs and expected insert sizes (Pop, 2009). The major disadvantage of this approach is related to the quality of the data and to the difficulty in minimising inconsistency between the assembled contigs and read pair constraints; for this reason, the joining of contigs requires multiple coherently aligned read pairs (Li *et al.*, 2010; Simpson *et al.*, 2009).

Alternative approaches for organising contigs involve the use of PCR-assisted or other gap-closing techniques (Figure 2), optical maps, reference genomes, and multiple genome assemblies. Optical mapping is a technique for generating whole-genome ordered restriction endonuclease maps (Samad *et al.*, 1995). These optical maps not only provide information about restriction fragment sizes, but also provide information regarding the order in which these fragments occur in the DNA. Regarding scaffolding, matches between *in silico* restriction-digested sequences and fragments in the optical map provide means for stitching contigs together (Nagarajan *et al.*, 2008). The use of reference genome alignments is another method for organising contigs into larger units (Richter *et al.*, 2007; Rissman *et al.*, 2009; Bartels *et al.*, 2005). This approach is increasingly common as the number of sequenced genomes in databases increases; however, the accuracy of the scaffolds depends on the alignment quality and the level of sequence similarity between the contigs and reference genomes. Indeed, even closely related reference genomes are useless in the presence of horizontal transfer and genomic rearrangements. Finally, scaffolds can be constructed by integrating assemblies. For example, methods such as Minimus (Sommer *et al.*, 2007), MAIA (Nijkamp *et al*., 2010) and GAA (Yao *et al*., 2012) make use of multiple assemblies to produce meta-assemblies with increased contiguity and accuracy.

1.1.4 Structural annotation

Structural annotation is the aspect of genome annotation that consists of the identification of genomic features (Koonin & Galperin, 2003; Angelova *et al.*, 2010). In theory, this can be achieved using experimental techniques but is typically attained using computational approaches followed by manual curation due to time and cost. This chapter introduces

bioinformatic strategies for calling CDSs, ncRNA genes, and MGEs in bacterial genomes. Table 2 and Appendix Table 1 list some popular computational tools for structural annotation.

CDS calling is one of the most important steps of structural annotation (Angelova *et al.*, 2010). In its simplest form, CDS calling could consist of scanning genomes for sufficiently long (≥ 90 bases) uninterrupted stretches of DNA between a start codon and a stop codon. However, the screening of such ORFs can yield a certain number of incorrect gene predictions (Koonin & Galperin, 2003) because true CDSs often echo long ORFs in neighbouring reading frames. An alternative strategy is to take advantage of the statistical properties of the coding sequences and estimate the coding potentials of ORFs or to infer genes based on the similarity of the encoded protein sequences to those of other proteins in public database (Koonin & Galperin, 2003). *Ab initio* programmes make predictions using a probabilistic model that distinguishes CDSs from noncoding sequences based on sequence composition and estimates the coding potentials of ORFs. The model employs, for example, the infrequencies of Gs and As at the first codon positions, the infrequency of Gs at the second codon position, and/or amino acid composition differences between coding and noncoding genome regions. Typically, the parameters used in probabilistic models are trained from separately prepared training datasets, which comprise sufficiently long and non-overlapping ORFs from the sequence in question (Delcher *et al.*, 2007), ORFs from the sequence in question showing homology to known proteins in public databases (Larsen & Krog, 2003), and sequences of a known type. An alternative to the *ab initio* prediction of genes is to search the target genome for CDSs that are similar to extrinsic evidence (Koonin & Galperin, 2003). Extrinsic methods include the BLAST-type mapping of ORFs against known gene products and gene callers such as ORPHEUS (Frishman *et al.*, 1998) and CRITICA (Badger & Olsen, 1999), which infer CDSs based on coding potential and sequence similarity. Finally, software packages exist for refining gene call anomalies (Pati *et al.*, 2010; Cruveiller *et al.*, 2005) and combining evidence from individual gene-finding systems into consensus CDS models (Tech & Merkl, 2003; McHardy *et al.*, 2004).

To date, no systematic analysis of the gene calling accuracies of different gene finders is available. However, comparisons of methods given in the original papers show that bacterial gene finder algorithms boast an average accuracy of 90% or better (Delcher *et al.*, 2007; Hyatt *et al.*, 2010; Besemer *et al.*, 2001). Moreover, function assignments can be attached to most gene products, indicating that bacterial gene callers are likely accurate. The major hurdles in the use of modern *ab initio* gene callers involve the identification of short genes (≤150 bp), over-annotation, the false prediction of pseudo-genes, and the presence of longer than anticipated overlaps between CDS calls (Besemer *et al.*, 2001; Delcher *et al.*, 2007; Hyatt *et al.*, 2010). The scarcity of stop codons in GC-rich genomes can also impair the accuracy of gene callers that give value for gene length in estimating coding potentials (Hyatt *et al.*, 2010). It is also possible that the probabilistic model may fail in genomic islands with atypical base compositions. In contrast, evidence-based methods can ignore novel CDSs and have longer runtimes than *ab initio* methods, which scan millions of bases in minutes. Evidence-based gene callers are nonetheless useful in calling CDSs that are ignored by *ab initio* gene finders and that exhibit homology to

known proteins. For example, current genome annotations appear to lack, on average, 30 *bona fide* genes that can be identified using a BLAST procedure (Warren *et al.*, 2010).

Another standard analysis in the process of structural genome annotation is the identification of tRNA, rRNA, and other types of ncRNA genes that function directly as RNA rather than being translated to proteins. Traditionally, ncRNA genes are called using comparative genomics methods (Eddy, 2002; Pichon & Felden *et al.*, 2008). In the case of rRNAs, genes can be called using primary sequence similarity with known rRNA genes (Lagesen *et al.*, 2007). However, other types of ncRNAs often lack common statistical signals in their primary sequences that could be exploited for detection. Instead, their calling often requires methods that use sequence and structural conservation, such as the tRNA prediction system tRNAscan-SE or the general ncRNA search suite Infernal. Both of these algorithms use covariance models to capture the primary consensus and secondary structure information of an RNA family and are very accurate (Lowe & Eddy, 1997; Nawrocki *et al.*, 2009). However, this performance comes at the expense of runtime, and the analysis requires a template ncRNA structure. The more general approaches for identifying ncRNA genes rely on genomic variations in sequence composition statistics, predict transcripts without long ORFs and with initiation and termination sites, search for sequences that have the ability to adopt given secondary structure patterns (Macke *et al.*, 2001), or use a combination of RNA structure prediction and comparative sequence analyses to test for a characteristic signal (Rivas & Eddy, 2001; Washietl *et al.*, 2005). These programmes are considered useful for defining the structure of a sequence that is already known to be an RNA gene but are largely immature for use in genome-wide scans (Eddy, 2002; Pichon & Felden, 2008).

Approximately one tenth of a bacterial replicon is intergenic. Although this portion of the genome is commonly referred to as noncoding, it contains a variety of important sequence features (Madigan *et al.*, 2010). Intergenic regions contain, for example, transcriptional regulatory elements and basal promoter elements that are key players in gene regulation. Such regions are also rich in repetitive elements and contain motifs contributing to the coordination of replication, cell division, DNA segregation, and DNA repair (Touzain *et al.*, 2010). Also visible in the genome are mRNA stem-and-loop structures that control of gene expression. Depending on the type of stem-and-loop structure, these motifs can be called with the help of general ncRNA annotation algorithms and by inferring genomes for rho-independent terminators using software tools such as RNIE (Gardner *et al.*, 2011) and TransTermHP (Kingsford *et al.*, 2007). Intriguingly, stem-and-loop structure annotations provide valuable clues about genome structure and enable the marking of operon endpoints and start sites, because rho-independent terminators are mainly located at transcription termini and transcriptional attenuators are located between the basal promoter elements and the start codons of the 5'-most genes of operons. Finally, intergenic regions can be annotated for CRISPR arrays. Because CRISPR arrays consist of short (approximately 20-50 bp) direct repeats that are interspaced by variable sequences called spacers, they are routinely inferred using repeat finding algorithms or their modified versions (Bland *et al.*, 2007; Edgar, 2007; Grissa *et al.*, 2007).

MGEs are DNA segments that encode proteins that mediate the movement of DNA within genomes or between bacterial cells. These sequence features can have a tremendous impact on the transfer, recombination, and deletion of host genes, and traces of MGE activity are present in nearly all bacterial genomes (Frost *et al.*, 2005). MGEs are typically annotated based on their similarity to the known MGE members. This paradigm is implemented in the IScan (Wagner *et al.*, 2007) and ISsaga (Varani *et al.*, 2011) methods, enabling the identification of insertion elements using curated references from the ISfinder database (Siguier *et al.*, 2006). This approach can also be used to identify clusters of genes that exhibit similarity to known phage genes (Lima-Mendez *et al.*, 2008; Zhou *et al.*, 2011), integrons and gene cassettes (Moura *et al.*, 2009; Joss *et al.*, 2009), and integrative and conjugative elements (Bi *et al.*, 2012). Furthermore, some computational resources for genomic island annotation build on sequence similarity searches (Langille *et al.*, 2008). Alternatively, the automated detection of MGEs can rely on sequence composition characteristics. The methods included in this category are based on the notion that MGEs are often acquired horizontally and that MGEs can be identified by searching for local variations in sequence composition, such as variations in the G+C ratio and dinucleotide bias. Several programmes for the automated detection of MGEs adopt this approach, including those designated for the annotation of plasmids (Zhou & Xu, 2010), phage-like regions (Srividhya *et al.*, 2007), and genomic islands (Tu & Ding, 2003; Waack *et al.*, 2006; Vernikos & Parkhill, 2006). Although these bioinformatic tools are also able to call novel MGEs, sequence composition skew appears to be a less reliable predictor than sequence similarity. For example, genomic island detectors that rely on base composition statistics exhibit less agreement with a dataset of known genomic islands than their homology-based counterparts (Langille *et al.*, 2010).

1.1.5 Protein function prediction

Protein function prediction can be defined as the inference and assignment of specific biological and biochemical roles to proteins (Koonin & Galperin, 2003). This stage of genome annotation attempts to compile a definitive catalogue of the protein functions of the organism and provides researchers with specific testable hypotheses about the roles of proteins in the cell. This knowledge is critical for understanding life at the molecular level (Koonin & Galperin, 2003; Rost *et al.*, 2003; Friedberg, 2006; Valencia, 2005). However, genome-scale protein function prediction is a challenging process (Rost *et al.*, 2003; Valencia, 2005; Friedberg, 2006; Schnoes *et al.*, 2009) and involves the use of many functional classification schemes and computational procedures (Table 2 and Appendix Table 1), as discussed in detail below. As with structural genome annotation, the use of automated prediction methods is preferably performed in conjunction with manual annotation.

Functional classification schemes are used to capture biological knowledge in a form that is suitable for computational processing. Regarding bacteria, the functions of proteins are most often available as description lines (DEs, Bairoch *et al.*, 2008). These natural language function labels are the traditional way of describing functional information and

can be very informative. However, DEs, like the natural language itself, are rife with synonyms and ambiguity, making comparison of DEs notoriously difficult (Friedberg, 2006). Another challenge with regard to DEs such as '*DNA gyrase subunit B*' is that protein function is a subjective concept, and different researchers may denote the functions of proteins differently (Friedberg, 2006). In addition to DEs, standardised functional labelling schemes have been developed to unify information about protein function. These schemes employ only certain words and punctuation and describe functional information in a controlled and computationally amenable fashion. Such resources include TIGRFAM (Haft *et al.*, 2003), SEED (Overbeek *et al.*, 2005), the keyword catalogues of UniProt (Bairoch *et al.*, 2008), and orthologous groups of proteins (COG) (Tatusov *et al.*, 1997; Tatusov *et al.*, 2003) that cover general functional aspects and providing a means by which to overview and compare the functional contents of organisms. Additional classification schemes have also been developed for the classification of enzymes (Webb, 1992), enzyme-related functions (Rawlings *et al.*, 2004; Roberts *et al.*, 2010; Cantarel *et al.*, 2009), and transporters (Saier *et al.*, 2006). Noteworthy examples include the Enzyme Classification (EC, Webb, 1992) and Transporter Classification (TC, Saier *et al.*, 2006) systems, which follow a hierarchical structure that allows the measurement of the functional similarity of genes. Another intensively used functional classification scheme is Gene Ontology (GO, Ashburner *et al.*, 2000). Machine-readable GO encompasses three ontologies that describe three aspects of function: molecular function, biological process, and cellular location. These ontologies are non-redundant and are implemented as a directed acyclic graph (DAG) that represents general terms as nodes near the root of the ontology and specific terms as nodes near the leaves of the ontology. Differing from the hierarchical structure, a node can have multiple parents. By definition, if a gene is associated with a term, it is associated with all of its broader level GO terms. Furthermore, each GO term assignment has an evidence code attributed to it, thereby providing a means of separating computationally derived information from manually curated information (Ashburner *et al.*, 2000). Although GO is the most widespread functional classification scheme in use today, its use in bacterial genome annotation is limited by the low number of specific GO descriptions that have been associated with bacterial proteins.

The conventional approach for protein functional classification involves the transfer of biological information on the basis of sequence similarity. The rationale for this approach is that sequences with a high degree of similarity are likely to have evolved from a common ancestor and will also share functional roles (Friedberg, 2006). This type of homology-based functional classification has been shown to be rather accurate and to be able to assign a function to approximately 73% of CDSs in the average genome (Raes *et al.*, 2007); however, this approach is not without problems. Principally, sequence similarity does not imply similarity of function (Eisen, 1998), and transferring annotations based on sequence similarity can propagate existing annotation errors (Schnoes *et al.*, 2009). This analysis method also fails with regard to orphan genes and often results in a set of differently characterised sequences. It also appears that the quality of the prediction depends on the sequence database used (Schnoes *et al.*, 2009), the sequence similarity search method used, and the type of function information to be transferred (Clark &

Radivojac, 2011). Another shortcoming is the need to separate useful and spurious sequences. Some studies have proposed that as low as 30% amino acid sequence identity is required for assigning complete enzyme function (Valencia, 2005; Devos & Valencia, 2001) and that inference by similarity performs approximately equally well for all matches above this sequence identity threshold (Clark & Radivojac, 2011; Altenhoff *et al.*, 2012). In contrast, other researchers argue that below a 70% sequence identity threshold, EC numbers start to diverge rapidly, such that at 30% sequence identity, only a tenth of pairs of proteins share all EC numbers (Rost, 2002; Rost *et al.*, 2003). It appears that the latter conclusions are more accurate regarding bacterial enzymes (Figure 4). Among the manually annotated proteins entries available in the UniProt database, the proportion of pairs in which both proteins are described using the same enzyme code became high above the 50% sequence identity level, whereas at the 30% identity level, only every fifth protein pair was observed to share the same enzyme class.

The functional classification of proteins on a genome-wide scale typically involves a direct sequence-sequence comparison of query proteins against large sequence databases using BLAST (Altschul *et al.*, 1997), FASTA (Pearson & Lipman, 1988), or PSI-BLAST (Altschul *et al.*, 1997). The function of the prototype sequence is then transferred to the protein under consideration. In its simplest form, predictions are based on information that is associated with the top hit (*i.e.,* the best-BLAST approach) or the most informative description. Based on Figure 4, this approach is accurate provided that it is performed for highly similar enzymes. More elaborate methods have also been proposed that make use of more than one sequence (Abascal & Valencia, 2003; Kunin & Ouzounis, 2005, Martin *et al.*, 2004; Vinayagam *et al.*, 2006; Hawkins *et al.*, 2006; Wass & Sternberg, 2008) or that transfer functions within an evolutionary context (Engelhardt *et al.*, 2005; Zmasek & Eddy, 2002; Storm & Sonnhammer, 2002). According to the method comparisons provided in the original papers, the pooling of information over multiple sequences and the additional use of weakly similar sequences as the source of function information appear to be beneficial and increase prediction accuracy. An alternative but fundamentally similar approach is to search for sequence-based signatures. Examples of tools that are designed for finding signatures and resources that archive signature profiles are listed in Table 2 and Appendix Table 1. Overall, the use of protein signatures appears to be more powerful and sensitive for detecting remote homologues than the use of pairwise sequence similarity search tools. Nevertheless, the combination of several methods for functional annotation is recommended because the most suitable method for a particular sequence set cannot be known *a priori*.

Homology transfer is also the principal method for identifying proteins with specific and predetermined functional roles, such as signal transduction system proteins, virulence factors, and adhesins. For example, function-specialised databases have implemented the possibility of conducting a BLAST-like search against their high-quality sequence sets. Some representative examples in the field of microbiology include databases and services that were developed for mining bacteriocins (van Heel *et al.*, 2013), carbohydrate-active enzymes (Cantarel *et al.*, 2009), *cas* genes (Haft *et al.*, 2005), signal transduction system proteins (Ulrich & Zhulin, 2010), antibiotic resistance factors (Scaria *et al.*, 2005; Zhou *et al.*, 2007; Liu & Pop, 2009), type II Toxin-antitoxin systems (Sevin & Barloy-Hubler,

2007; Shao *et al.*, 2011), pilus components (developed in Study I), restriction modification system components (Roberts *et al.*, 2010), DNA-binding transcription factors (Wilson *et al.*, 2008), and secondary metabolite biosynthetic loci (Medema *et al.*, 2011).

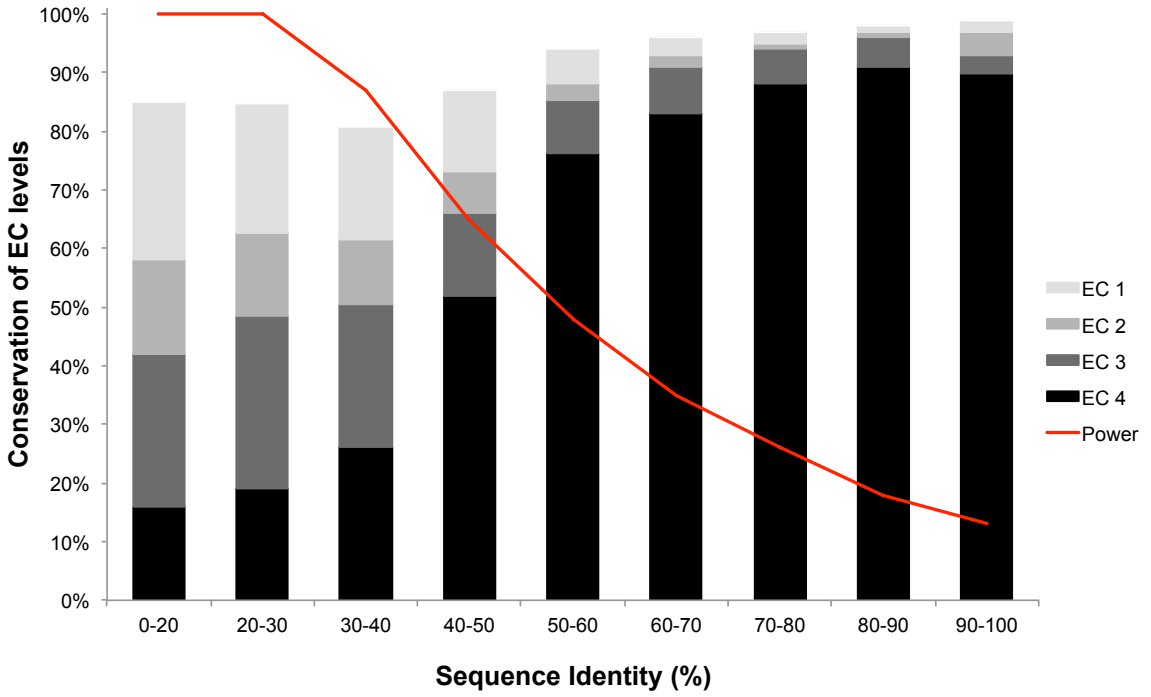


Figure 4. Conservation of EC number and the power of homology transfer for bacterial enzymes. Protein sequence data was extracted from the UniProt database, and only entries that had been verified either at the RNA level or at the protein level were accepted. Sequence-level identity was quantified using BLAST (default settings), and the fraction of pairwise sequence matches sharing a certain number of EC number digits was tabulated at different levels of pairwise sequence identity. The red line indicates the fraction of sequences that match other sequences at the given or a higher identity range.

Apart from evolutionary counterparts, functional classification can rely on the genomic context of genes. This approach utilises the co-localisation and/or co-evolution of genes to predict functional linkages between encoded proteins and is also practical for genes that match only other uncharacterised genes because the transfer of functional information between organisms is not necessary. Instead, the use of context information provides an opportunity for also transferring information between genes within a single organism. The four major methods for context-based prediction are the phylogenetic profile method (which employs the fact that functionally associated genes appear to be preserved or eliminated in concert during evolution; Pellegrini *et al.*, 1999), the gene fusion method (which relies on gene pairs that occur in parallel as a larger composite gene in other genomes; Marcotte *et al.*, 1999), the gene neighbour method (which searches for genes

with preserved physical genomic proximity; Dandekar *et al.*, 1998), and the co-evolution method (which uses the correlation between phylogenetic trees; Juan *et al.*, 2008). Among these methods, the phylogenetic profile method is considered the most precise and provides comparable performance to homology-based functional classification (von Mering *et al.*, 2005).

Ab initio protein function prediction methods predict protein function based on sequence information alone. Unlike other functional prediction methods, these methods are resistant to existing error-prone information in databases and have the advantage of being also suitable for orphan proteins as the analysis does not involve transfer of function from unreliable sources (Ofran *et al.*, 2005; Punta & Ofran, 2008). However, *ab initio* methods currently provide only rough approximations of various aspects of function. Nevertheless, promising methods exist for the prediction of subcellular localisation and transmembrane topologies of proteins (Table 2 and Appendix Table 1); these methods providing valuable clues and rather accurate predictions about bacterial secretomes that comprise secreted or surface-associated proteins. With regard to bacterial genome annotation, other aspects of protein function that can be predicted using *ab initio* methods include amino acids that are involved in DNA binding (Ofran *et al.*, 2007) and binding to various metals (Lippi *et al.*, 2008), virulence-associated factors (Garg & Gupta, 2008), the probability of a protein being an adhesin (Sachdeva *et al.*, 2005), and palmitoylation sites in protein sequences (Ren *et al.*, 2008). In addition, specialised bioinformatic resources are available that locate proteins that are secreted by type III (Arnold *et al.*, 2009) or IV (Burstein *et al.*, 2009) secretion systems.

1.1.6 Summarisation of genome annotation results

The summarisation and visualisation of genome data are key parts of any genome project and one of the first tasks to be performed following annotation. Typically, genome data are visualised to obtain an overview and to compare genome annotations. This step is also an important quality control step and can reveal assembly errors and genome regions that may need attention. Picture-based structural DNA analysis (*e.g.,* the visualisation of GC-skew, AT-content, GC-content, and gene-strand bias) can also be used to identify genomic features such as MGEs and origins of replication. Tools for data visualisation are listed in Table 2 and include standalone genome browsers, such as ACT (Carver *et al.*, 2005), and web services, such as the BLASTatlas (Wassenaar *et al.*, 2010). Additionally, genome annotation results can be recapitulated into genomic feature tables that, typically, provide information about the genome size, GC-content sequence elements, and coding density of the organism in question and of other, closely related organisms. These tables can also highlight the biological aspects of the organism. Genome size is, for example, an indicator of adaptive potential; large rRNA and tRNA gene numbers suggest a short doubling time, and codon usage can be used to shed light on the organism's likely environmental niche (Wassenaar *et al.*, 2010).

1.1.7 Comparative genomics

Whole-genome comparisons are a powerful approach for understanding genomic diversity and the relatedness of organisms (Ali *et al.*, 2013). This approach can reveal fascinating differences and similarities between genomes, is a prerequisite for profiling rapidly evolving sequences (Cooper *et al.*, 2005; Garber *et al.*, 2009), and provides a means to classify species and to describe horizontal gene transfer events (Darling *et al.*, 2004). The two underlying paradigms that are used to identify regions of similarity are local and global sequence alignment (Frazer *et al.*, 2003). The first of these reports all similar subregions of the sequence and also identifies homology in the presence of rearrangements. However, local alignment cannot suggest how the subregions have evolved from their ancestors (Frazer *et al.*, 2003). In contrast to local alignments, global alignments describe an end-to-end alignment of sequences. Aligned regions need to be conserved in both order and orientation (Frazer *et al.*, 2003); thus, this approach is optimal for the comparison of genomes with a high degree of synteny, but it is less good for outlining genomic relatedness of organisms in the presence of horizontal transfer and genomic rearrangements. Nevertheless, both alignment strategies can be useful, especially when explored graphically using comparative alignment viewers such as Combo (Engels *et al.*, 2006) and ACT (Carver *et al.*, 2005). Intuitive visualisation of the whole-genome homology can also be achieved by using dot-plot visualisation software (Krumsiek *et al.*, 2007) and by mapping and visualising genome homology of genes and proteins within a reference strain in comparison to other prokaryotes (Wassenaar *et al.*, 2010). Additionally, whole-genome comparisons can base on advanced genome alignment methods (Darling *et al.*, 2010; Blanchette *et al.*, 2004; Paten *et al.*, 2008; Dubchak *et al.*, 2009, Angiuoli *et al.*, 2011; Rissman *et al.,* 2009) that often mix global and local alignment procedures and can align long sequences while detecting the presence of inversions, translocations, duplications, and gains and losses. However, the use of even the most sophisticated software entails the selection of many mundane parameters (Frith *et al.*, 2010), and these methods perform best with sequences that exhibit significant nucleotide-level similarity and colinearity.

The identification of orthologue groups is a central part of functional classification (Li *et al.*, 2003; Koonin & Galperin, 2003; Sonnhammer & Koonin, 2002) and underpins the delineation of phenotype-genotype relationships among bacteria by providing a means of listing genes that are present only in the genomes of those isolates that express the phenotype of interest (Korbel *et al.*, 2005). In general, methods for constructing orthologue groups are classified into distance and tree-based methods. Distance-based methods include the reciprocal best hit (RBH, Tatusov *et al.*, 1996) and reciprocal smallest distance (RSD, Wall *et al.*, 2003) approaches. These methods build on BLAST scores or maximum likelihood estimations of evolutionary distances and they resolve orthologous (*i.e.,* homologues that have evolved by speciation from a single ancestral gene) between two organisms by finding two-way best genome-wide similarities. The software InParanoid extends this concept further (Remm *et al.*, 2001) and exploits the RBH strategy to identify orthologues between two species while applying additional rules to accommodate paralogues that arise from duplication after speciation (*i.e.*, inparalogues).

The rationale behind this approach is that orthologues and inparalogues are more likely to perform the same function than outparalogues that result from duplication preceding the speciation event (Sonnhammer & Koonin, 2002). Distance-based methods have also been proposed for finding orthologue groups across multiple genomes. These approaches typically use pairwise sequence similarity search methods and cluster RBHs that span multiple genomes using triangular linkage clustering (Tatusov *et al.*, 1997; Jensen *et al.*, 2008), the Markov clustering procedure (Li *et al.*, 2003), maximum weight cliques (Roth *et al.*, 2008), or fully connected sub-graphs (Klimke *et al.*, 2009). Alternatively, orthologue detection can be based on tree reconstruction and on gene and species tree reconciliation (Zmasek & Eddy, 2002; Storm & Sonnhammer, 2002). Regardless, assessments of orthologue prediction tools have resulted in contradictory results. One study concluded that EggNOG is the best and OrthoMCL the worst prediction method (Trachana *et al.*, 2011); however, another study endorsed OrthoMCL and InParanoid over other methods (Chen *et al.*, 2007). Confusingly, a recent study proposed that the RBH method might be more accurate and specific than any complex method (Salichos & Rokas, 2011). Overall, phylogenetic approaches are powerful for capturing evolutionary relationships; however, the computational costs of multiple sequence alignment, the lack of an accurate species trees for a given collection of bacteria, and the complexity of tree reconciliation preclude the use of this approach for the genome-wide identification of orthologues (Rentzsch & Orengo, 2009; Trachana *et al.*, 2011).

Molecular phylogenetic analyses are frequently used to depict the evolutionary history of a given set of organisms (Williams & Sarah, 2014). In addition, they can be applied to a single gene family, with each copy of the gene in each organism being included in the analysis (Yang & Rannala, 2012). The reconciliation of the gene tree with a species tree can then be used to time gene gains, duplications, and losses (Eisen, 1998). Some methods have even been proposed for the genome-scale inference of gene gains and deletions and for the inference of ancestral gene inventories (Mirkin *et al.*, 2003). Relatedness among organisms has typically been estimated by comparing molecular sequences, mostly small-subunit ribosomal RNAs (Woese, 1987) or ubiquitous housekeeping genes (Konstantinidis & Tiedje, 2005). The process begins with the extraction of sequences from all species under examination. After obtaining a multiple sequence alignment for the sequences, several phylogenetic methods can be used to infer the phylogeny. These methods can be broadly classified into maximum parsimony, maximum likelihood, Bayesian inference, and distance methods (Yang & Rannala, 2012), which differ largely in the way in which they choose the best among all possible trees. Maximum parsimony selects the tree that requires the minimum number of character changes from the common ancestral sequences. In maximum likelihood methods, the probability that a certain tree with a set of parameters produces a given set of data is computed, and the tree that makes the given data most probable is chosen. In Bayesian analysis, inferences of phylogeny are based upon the posterior probabilities of phylogenetic trees, whereas distance-based methods calculate the evolutionary distance among sequences of interest and construct a distance matrix that is used to cluster sequences hierarchically (Yang & Rannala, 2012). Additionally, organism-level evolutionary histories can be estimated from genome-scale datasets. These methods are believed to generate a more accurate picture of evolution,

especially for bacteria that have a high incidence of gene movement from one lineage to another (Brown *et al.*, 2001) and include the derivation of phylogenies from gene content (Snel *et al.*, 1999), gene order (Korbel *et al.*, 2002), and compositional signatures (Fox *et al.*, 1980). Further, phylogenies can be determined on the basis of multiple source-trees (Sanderson *et al.*, 1998) or concatenated single-copy orthologue sequences (Brown *et al.*, 2001; Rokas *et al.*, 2003).

1.1.8 Metabolic and regulatory reconstructions

Metabolic reconstructions provide a starting point for the study of the metabolism of an organism and attempt to include all of the relevant metabolic information of an organism (Oberhardt *et al.*, 2009; Durot *et al.*, 2009). In general, the metabolic information of an organism is represented in terms of metabolic pathways, each of which describes a set of chemical interactions and transformations for the conversion of compounds (Durot *et al.*, 2009). Metabolism-related data can also be conceptualised, and pathways can be integrated into genome-wide networks that usually envision metabolites as nodes and reactions as edges of the network. Although metabolic networks are not generated as often as metabolic pathways in bacterial genome studies, this approach offers a realistic view of metabolism, whereby, in theory, any reaction can have implications for other reactions (Oberhardt *et al.*, 2009; Durot *et al.*, 2009). Typically, metabolic pathway and network reconstructions stem from genome annotations (Oberhardt *et al.*, 2009). Initially, a preliminary list of metabolic and transport reactions relevant for the given organism are harvested; gene name, TC number and EC number assignments and transmembrane protein predictions being the main information sources. Metabolic pathways and networks are then formed by linking individual reactions into increasingly complex structures. Reaction sets can be assembled *ab initio* by tracking the movement of atoms through the network (Arita, 2004; Heath *et al.*, 2010) or by projecting the metabolic data onto reference metabolic pathways (Moriya *et al.* 2007; Karp *et al.*, 2002). Finally, the initial metabolic networks are refined based on literature information, manual inspection, experimental data, and bioinformatic tools that annotate inconsistencies between metabolic models and substrate utilisation predictions and that can resolve metabolites or reactions that are disconnected from the rest of the metabolism. For example, Pathway Tools can include missing reactions in pathways if a significant fraction of the remaining reactions of the pathways are supported by genome annotations (Karp *et al.*, 2002). In addition to reconstructing metabolic networks, genome annotation allows the reconstruction of transcriptional regulatory networks (Barabási & Oltvai, 2004). These networks provide a global picture of the transcriptional machinery of the cell and are constructed by integrating existing knowledge of regulons, operons, and transcriptional regulation interactions with the results of operon and transcription factor binding site screens (Baumbach *et al.*, 2009; Ravcheev *et al.*, 2013). Bioinformatics resources that are relevant to the study of metabolism and transcriptional regulation are listed in Table 2 and Appendix Table 1.

1.1.9 Genome annotation pipelines

Several genome annotation systems that are intended for the automated, in-depth annotation of prokaryotic genomes have been designed and presented in recent years, including the IMG (Markowitz *et al.*, 2012), RAST (Aziz *et al.*, 2008), DOE-JGI MAP (Mavromatis *et al.*, 2009), CG pipeline (Kislyuk *et al.*, 2010), ERGO (Overbeek *et al.*, 2003), and JCVI (Tanenbaum *et al.*, 2010) genome annotation pipelines. Some of these systems are completely automatic online services that have the advantage of simplicity, whereas others are standalone tools that require maintenance but provide an extra level of confidentiality. Some popular bacterial genome annotation systems are listed in Table 2 and Appendix Table 1. Typically, genome annotation pipelines involve a wide array of methods, through which they identify genomic features and assign functional information that describe the biological role of these features. For example, the annotation service of the Broad Institute calls ncRNA genes and CDSs using three automated methods for each. Gene models are clustered, and representative models are selected using heuristics, such as the relative overlap with BLAST hits. Finally, gene models with problems are filtered out (as described in the human microbiome web page; Nelson *et al.*, 2010). The JCVI genome annotation system (Tanenbaum *et al.*, 2010) is another popular annotation system and shares similarities with that of the Broad Institute. CDS prediction employs one *ab initio* and two evidence-based prediction methods, and the pipeline runs three RNA gene callers over the genome: tRNAScan-SE, ARAGORN, and BLAST searches against Rfam. Other modern annotation pipelines call genes in a highly similar fashion. After gene calling, most bacterial genome annotation systems search the set of predictions against one or more protein databases using BLAST (Markowitz *et al.*, 2012; Aziz *et al.*, 2008; Mavromatis *et al.*, 2009; Kislyuk *et al.*, 2010; Overbeek *et al.*, 2003; Tanenbaum *et al.*, 2010). Further, the gene products are usually also searched via InterProScan against a set of sequence profile databases. In addition to these basic analyses, some annotation pipelines include additional analysis modules and perform protein subcellular localisation prediction (Kislyuk *et al.*, 2010; Tanenbaum *et al.*, 2010; Markowitz *et al.*, 2012) and reconstruct metabolic pathways (Aziz *et al.*, 2008). Some annotation pipelines also have viewers that permit users to rectify old calls and introduce new gene and function calls. Surprisingly, modern genome assemblers and assembly pipelines are scarce in bacterial genome annotation systems; the CG pipeline is the sole exception (Kislyuk *et al.*, 2010). Although the various annotation servers are largely based on the same bioinformatic tools, pipelines appear to produce rather different annotation results. Importantly, the one study that has systematically compared the results of annotation services has documented distinct differences in annotation outputs (Bakke *et al.*, 2009).

1.2 Lactobacilli

The genus *Lactobacillus* comprises a large, heterogeneous group of gram-positive, non-sporulating, rod-shaped bacteria that have complex nutritional requirements and a low GC-content genome (less than 50 mol%). These bacteria are acid-tolerant, aero-tolerant or

anaerobic, and aciduric or acidophilic and employ a strictly fermentative metabolism (Hammes & Vogel, 1995; Felis & Dellaglio, 2007; Salvetti *et al.*, 2012). They are part of the lactic acid bacteria (LAB) group, which is characterised by the production of lactic acid as the main by-product of carbohydrate fermentation (Kandler & Weiss, 1986). In general, lactobacilli are encountered in an array of plant-, food-, and animal-related habitats that are rich in carbohydrates (Pot *et al.*, 1994; Hammes & Vogel, 1995; Felis & Dellaglio, 2007; Salvetti *et al.*, 2012). Some species, such as *Lactobacillus iners*, are restricted to specific niches, whereas others demonstrate a notable ability to adapt to a diverse set of environments. Lactobacilli also have a beneficial effect on our daily life and are of great economic importance. Several species are encountered on and in the human body, including the oral cavity, gastrointestinal tract (GIT), and vagina (Hammes & Vogel, 1995; Walter, 2008; Salvetti *et al.*, 2012); other species are essential in the fermentation of food, beverage, and feed products (Leroy & Vuyst, 2004; Bernardeau *et al.*, 2006; Giraffa *et al.,* 2010). Recently, members of *Lactobacillus* have been added to dietary and dairy products for probiotic purposes and to offer benefits for health and wellbeing (Saxelin *et al*., 2005; Giraffa *et al.* 2010).

1.2.1 Cellular characteristics of lactobacilli

The cell structure of lactobacilli is typical for that of a gram-positive bacterium and is often organised into three basic architectural regions: the cytoplasm, cell envelope, and surface appendages (Figure 5). The innermost region is the cytoplasm, which is largely the site of metabolism and replication. The cytoplasm contains the nucleoid, which is an irregularly shaped and non-membrane bound region that contains the chromosomal DNA, ribosomes, which are essential for protein synthesis, and an array of non-coding RNA molecules (ncRNA), which are not translated into proteins but function at the RNA level (Madigan *et al.*, 2010). Also included in the cytoplasm are various proteins that are responsible for important functions, such as forming structural components (structural proteins), the catalysis of biochemical reactions (enzymes), and the transmission of molecular signals (transcription factors and signal transducers). The innermost compartment can also carry one or more independently replicating extra-chromosomal DNA molecules (plasmids) that are not essential for survival but can comprise a notable fraction of the total genome (Madigan *et al.*, 2010). Of note, a sizable part of the genome is present as fragments of DNA that are capable of moving around within the genome or between genomes. A wide variety of such mobile genetic elements (MGEs) have been characterised in a number of different lactobacillus strains and include plasmids (Claesson *et al.*, 2006), genomic islands (genome regions that exhibit evidence of horizontal origins; Kleerebezem *et al.*, 2003), transposons (Callanan *et al.*, 2008), and prophages (lysogenic phages that can switch under some conditions to a lytic lifestyle and then infect other bacteria; Ventura *et al.*, 2006).

The cell envelope is a structural compartment that protects the cytoplasm and mediates interactions with the host and environment. In gram-positive lactobacilli, the cell envelope contains a cytoplasmic cell membrane and a thick peptidoglycan layer that is decorated

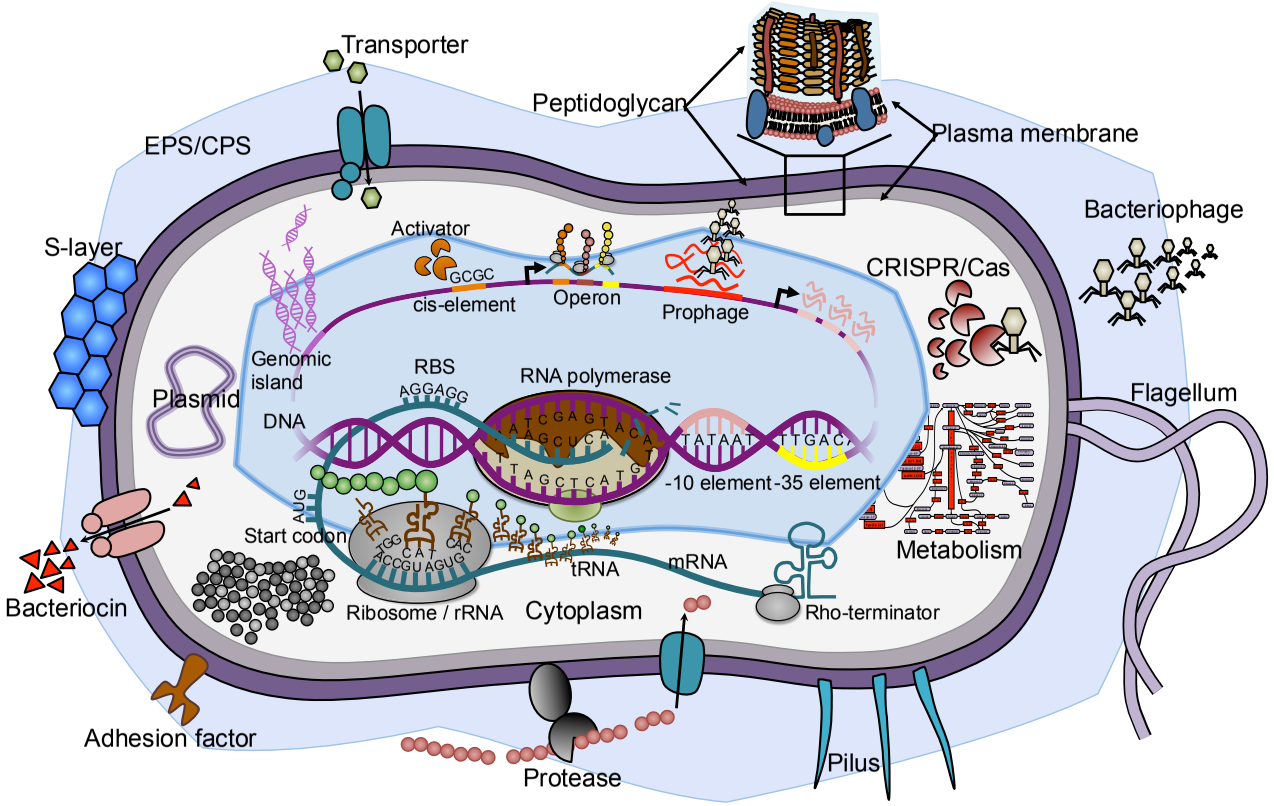


Figure 5. Schematic representation of a *Lactobacillus* cell. Included are the CRISPR (clustered regularly interspaced palindromic repeats)-Cas (CRISPR-associated) adaptive immunity system (which predominantly targets foreign genetic material), bacteriocins (proteinaceous antibacterial compounds that inhibit the growth of mainly closely related bacteria), pilus fibres (by which bacterial cells are attached to surfaces), and flagella (which are involved in cell movement). Genomic islands, clusters of co-transcribed genes (operons), prophages, *cis*-elements (DNA-binding sites for transcription factor proteins), -10 and -35 elements (attachment sites for σ-factor of RNA polymerase), ribosomal binding sites (RBSs), and rho-terminator sites (which control transcription termination through RNA hairpin formation) are also indicated in the figure. The cells are between 1 and 10 microns in length and are approximately 1 micron in width.

with proteins, teichoic acids, and polysaccharides (Lebeer *et al.*, 2008; Madigan *et al.*, 2010). In addition, some lactobacilli express an outermost coat, the Surface-layer (S-layer), which is composed of a single protein that completely encases the cell and appears to regulate bacterial contacts with the human dendritic cell (Konstantinov *et al.*, 2008). Some lactobacilli also surround themselves with an exopolysaccharide (EPS), which is either tightly associated with the cell wall or secreted into the surroundings (Lebeer *et al.*, 2008). Interestingly, these ubiquitous components of the cell envelope of lactobacilli have been shown to exert some immune responses and are therefore attractive candidates as probiotic effector molecules (Lebeer *et al.*, 2008). Finally, molecular and genomic studies of lactobacilli have characterised the presence of surface appendages in lactobacilli as being involved in movement (flagella; Forde *et al.*, 2011) or adhesion to surfaces (sortase-dependent pili and fimbriae; von Ossowski *et al.*, 2011; Lebeer *et al.*, 2012; Pridmore *et al.*, 2004). The cell envelope also contains membrane-associated proteins that are responsible for such processes as nutrient acquisition, adhesion, cell communication, microbe-host interactions, and stress sensing (Lebeer *et al.*, 2008).

1.2.2 Sugar fermentation

Lactobacilli exhibit a strong ability to degrade various carbohydrates and derive energy mainly from the conversion of sugars into lactic acid. In general, these bacteria degrade hexoses through homo- or heterofermentative carbohydrate fermentation pathways (Hammes & Vogel, 1995; Pot *et al.*, 1994). First, homofermentative lactobacilli use glycolysis (Embden-Meyerhof-Parnas pathway) to ferment hexoses primarily into lactic acid; most *Lactobacillus* species fall into this group (Salvetti *et al.*, 2012). In comparison, obligately heterofermentative lactobacilli ferment hexoses and pentoses via the pentose phosphate pathway. During this process, half of the substrate is converted into lactic acid and the rest is metabolised in equimolar amounts to ethanol (or to acetic or formic acid) and carbon dioxide. Finally, facultatively heterofermentative species utilise both pathways and are almost as common as homofermentative lactobacilli species (Salvetti *et al.*, 2012). Genetically, the three modes of sugar fermentation are explained by the absence or presence of genes encoding aldolase and phosphoketolase. Aldolase is present in homofermentative and facultatively heterofermentative species; phosphoketolase is present in obligately and facultatively heterofermentative lactobacilli (Salvetti *et al.*, 2012).

1.2.3 Taxonomy

The genus *Lactobacillus,* as currently circumscribed, contains over 152 species that exhibit wide phenotypic and genotypic variation. The genus is polyphyletic with the genus *Pediococcus* and is the largest genus within the family *Lactobacillaceae*, which in turn belongs to the order Lactobacillales, class Bacilli, and phylum Firmicutes. According to the most recent systematic study (Salvetti *et al.*, 2012), *Lactobacillus* species can be

subdivided into 29 distinct phylogenetic groups (Figure 6). These groups represent a distinct cluster in the 16S rRNA gene phylogeny and were named in the study according to the first recognised species of the given group. Some economically and biomedically important phylogenetic *Lactobacillus* groups are introduced below.

The *Lactobacillus delbrueckii* group is the largest of the phylogenetic *Lactobacillus* groups (Salvetti *et al.*, 2012) and contains many species that are essential in food production. For example, *L. delbrueckii* is widely used as a starter culture in yoghurt manufacturing, whereas *Lactobacillus helveticus* is important in the manufacture of a range of Swiss- and Italian-type cheeses (Leroy & De Vuyst, 2004; Giraffa *et al.,* 2010). Included in the group are also GIT-associated species, such as *Lactobacillus acidophilus* and *Lactobacillus johnsonii* (Altermann *et al.*, 2005; Pridmore *et al.*, 2004), and species such as *Lactobacillus crispatus*, *Lactobacillus jensenii*, and *L. iners,* which are major constituents of the healthy adult female urogenital tract and important agents of urogenital health (Ma *et al.*, 2012; Martin, 2012). Notably, the *L. delbrueckii* group includes several commercially distributed probiotic strains that appear to benefit health (Saxelin *et al*., 2005; Giraffa *et al.* 2010).

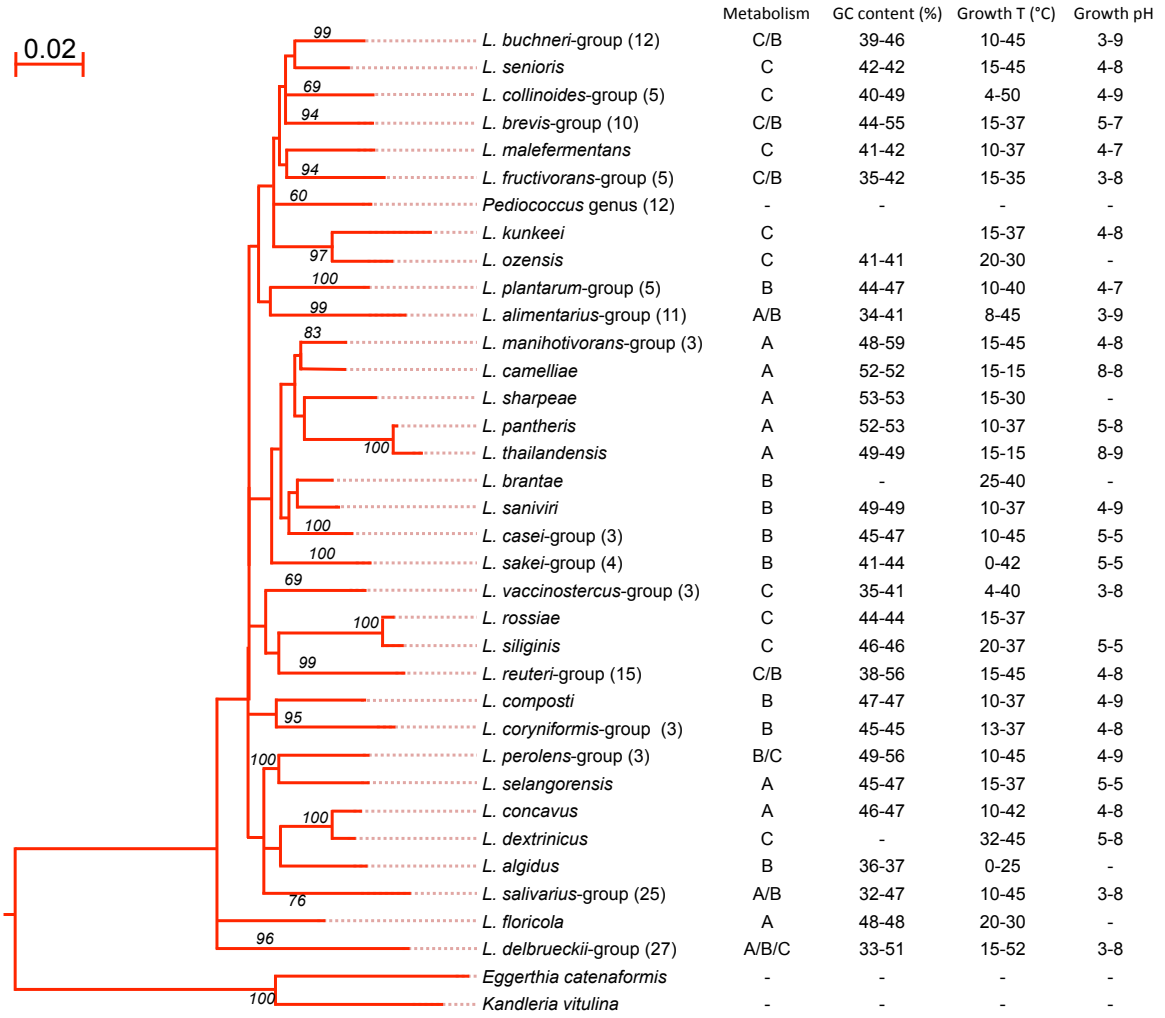
The *Lactobacillus salivarius* group is a heterogeneous group of 16 homofermentative and 9 facultatively heterofermentative species. The GC content within this group varies widely from 32 to 47 mol% (Figure 6), reflecting the fact that the members of this group occupy a wide variety of habitats, including human saliva, vertebrate intestine, soil, water, plants, and food (Forde *et al.*, 2011). Some species of the group appear to be motile and contain genes that encode the flagellar apparatus (Forde *et al.*, 2011).

The *Lactobacillus casei* group comprises three species; namely, *L. casei*, *Lactobacillus paracasei*, and *Lactobacillus rhamnosus*. These species have a broad ecological distribution and are frequently found in plant material as well as in the oral cavity and GIT of humans and animals (Kandler & Weiss, 1986). Industrially, *L. paracasei*, *L. rhamnosus*, and *L. casei* have applications as acid-producing starter cultures for milk fermentation and as starter adjunct cultures for the intensification and acceleration of flavour development in bacterial-ripened cheeses (Broadbent *et al.*, 2012; Mäyrä-Mäkinen & Bigret, 1998). Selected strains, such as *L. rhamnosus* GG, *L. casei* Shirota, and *L. casei* DN114-001*,* are commercially important probiotic that are added to various products for their potential to enhance the health of humans (Saxelin *et al*., 2005; Giraffa *et al.,* 2010; Siezen & Wilson 2010).

The *Lactobacillus reuteri* group contains 15 species that cover a broad host range. These species have been isolated from foods such as rye-bran fermentations and sourdough, and some are frequent in the GIT of birds, pigs, mice, and rats (Forde *et al.*, 2011; Frese *et al.*, 2011). *L. reuteri* is also considered indigenous to humans according to some investigations (Walter, 2008).

The *Lactobacillus sakei* group contains four facultatively heterofermentative species. The best known of the four species is *L. sakei*, a psychrotrophic bacterium that is found naturally on fresh meat and fish and that is used widely in their fermentation (Chaillou *et al.*, 2005). The *Lactobacillus plantarum* is another well-studied phylogenetic lactobacilli group. Out of its five facultatively heterofermentative species, the most noteworthy is the genetically heterogeneous *L. plantarum*, which exhibits remarkable ecological adaptability

and can be recovered from a variety of habitats including fermented foods, vegetables, and the human GIT (Siezen *et al.*, 2010; Siezen & van Hylckama Vlieg, 2011). Of note, some strains of *L. plantarum* have probiotic applications (Kleerebezem *et al.*, 2003).



	Metabolism	GC content (%)	Growth T (°C)	Growth pH
L. buchneri-group (12)	C/B	39-46	10-45	3-9
L. senioris	C	42-42	15-45	4-8
L. collinoides-group (5)	C	40-49	4-50	4-9
L. brevis-group (10)	C/B	44-55	15-37	5-7
L. malefermentans	C	41-42	10-37	4-7
L. fructivorans-group (5)	C/B	35-42	15-35	3-8
Pediococcus genus (12)	-	-	-	-
L. kunkeei	C		15-37	4-8
L. ozensis	C	41-41	20-30	-
L. plantarum-group (5)	B	44-47	10-40	4-7
L. alimentarius-group (11)	A/B	34-41	8-45	3-9
L. manihotivorans-group (3)	A	48-59	15-45	4-8
L. camelliae	A	52-52	15-15	8-8
L. sharpeae	A	53-53	15-30	-
L. pantheris	A	52-53	10-37	5-8
L. thailandensis	A	49-49	15-15	8-9
L. brantae	B	-	25-40	-
L. saniviri	B	49-49	10-37	4-9
L. casei-group (3)	B	45-47	10-45	5-5
L. sakei-group (4)	B	41-44	0-42	5-5
L. vaccinostercus-group (3)	C	35-41	4-40	3-8
L. rossiae	C	44-44	15-37	
L. siliginis	C	46-46	20-37	5-5
L. reuteri-group (15)	C/B	38-56	15-45	4-8
L. composti	B	47-47	10-37	4-9
L. coryniformis-group (3)	B	45-45	13-37	4-8
L. perolens-group (3)	B/C	49-56	10-45	4-9
L. selangorensis	A	45-47	15-37	5-5
L. concavus	A	46-47	10-42	4-8
L. dextrinicus	C	-	32-45	5-8
L. algidus	B	36-37	0-25	-
L. salivarius-group (25)	A/B	32-47	10-45	3-8
L. floricola	A	48-48	20-30	-
L. delbrueckii-group (27)	A/B/C	33-51	15-52	3-8
Eggerthia catenaformis	-	-	-	-
Kandleria vitulina	-	-	-	-

Figure 6. A phylogenetic tree illustrating the evolutionary relationship between *Lactobacillus* and *Pediococcus* species based on 16S rRNA gene sequence similarity. The tree was calculated using Tamura Three Parameters as the distance matrix formula and minimum evolution as the tree reconstruction method. The scale bar represents the number of substitutions per site. Bootstrap values are reported in percentages at nodes if ≥60 %. Clusters containing more than three species were condensed and given the name of the first species described. The number of species in each group is indicated in parentheses, followed by the fermentation mode (A, homofermentative; B, facultatively heterofermentative; and C, obligately heterofermentative), GC content, temperature growth range, and pH growth range. The tree and phenotype data were adapted from Salvetti *et al.*, 2012.

1.2.4 Industrial applications

The large-scale production of dairy products such as cheeses, yogurts, and fermented milks is the best-known industrial application of lactobacilli (Leroy & De Vuyst, 2004; Bernardeau *et al.*, 2006; Giraffa *et al.,* 2010). However, lactobacilli also play an important role in other types of food fermentation (du Toit *et al.*, 2010, Leroy & Vuyst, 2004; Bernardeau *et al.*, 2006) and are used for the production and preservation of foods of plant (*e.g.*, pickles, olives, sauerkraut, sourdough bread, and Korean kimchi) and animal (*e.g.,* fermented and dry sausages, salami, and fermented fish) origin, even though lactobacilli can in some cases cause spoilage of meat and seafood products (Varnam, 2002). These processes typically employ lactobacilli as starter cultures that are added to a raw material to accelerate and drive fermentation. Alternatively, preservation can be based on microbes that are naturally present in the raw material (Pfeiler & Klaenhammer, 2007; Leroy & De Vuyst, 2004). However, the use of spontaneous fermentation results in less control over the fermentation process and varying product quality,.

Lactobacilli are mainly used in food production for the fermentative conversion of sugars into organic acids, mostly lactic acid. This reduces the sugar content of the product and acidifies the raw material below the pH ranges within which most food-spoilage microbes can grow (Leroy & De Vuyst, 2004). Acid production can also influence the organoleptic properties of the final product by contributing to the coagulation of milk proteins (Heller, 2001; Giraffa *et al.,* 2010). In addition to acids, lactobacilli are known to produce a variety of other compounds of relevance to food industry. Among these are aroma compounds, lipases, and proteases that play a role in flavour development in cheese (Steele *et al.,* 2012), EPS molecules that can enhance the mouthfeel of yogurts (Vuyst & Degees, 1999), and proteinaceous antibacterial peptides (*i.e.* bacteriocins) that can be used as natural food preservatives (Leroy & De Vuyst, 2004). Some lactobacilli are also able to produce health-enhancing ingredients (*e.g.*, vitamins, bioactive peptides, and antioxidants) that have potential applications as biotherapeutic agents (Saxelin *et al*., 2005; Giraffa *et al.,* 2010).

Chiefly, lactobacilli are generally regarded as safe (GRAS) organisms and have a long and safe history of application and consumption in the production of fermented foods. Under rare and unusual circumstances, the consumption of *Lactobacillus* products has been associated with infections in humans (Bernardeau *et al.*, 2008). However, the risk of *Lactobacillus* infection was estimated in that study to be unequivocally negligible, with approximately only one case per 10 million people over more than a century. In addition to food-related applications, lactobacilli are used to ferment animal feeds, to produce chemicals (Saxena *et al.*, 2009), to produce antibiotics, and as live vaccine carriers (Giraffa *et al.* 2010).

1.2.5 Lactobacilli in and on animals and humans

Lactobacilli are closely associated with humans and animals. In humans, they are notably abundant in breast milk (Collado *et al.,* 2009), the oral cavity (Walter, 2008), and the

mucosal surfaces of the vagina (Ravel *et al.*, 2011). Importantly, *Lactobacillus* species that are prevalent in breast milk, such as *L. gasseri* and *Lactobacillus fermentum* (Martín *et al.*, 2005), or in the vagina, such as *L. crispatus*, *Lactobacillus gasseri, L. iners*, and *L. jensenii* (Ravel *et al.*, 2011), are often regarded as offering a number of benefits for health and wellbeing (Martín *et al.*, 2005; Martin, 2012).

The human GIT is another site that hosts lactobacilli (Molin *et al.*, 1993; Ahrne *et al.*, 1998; Reuter, 2001; Vaughan *et al.*, 2005; Dal Bello & Hertel, 2006; Walter, 2008; Ryan *et al.*, 2008; Matsuda *et al.*, 2009) and at least 12 *Lactobacillus* species have been associated with the human GIT (Table 3). However, the *Lactobacillus* species composition varies among subjects, and approximately 25% of human faecal samples lack lactobacilli entirely (Walter, 2008). Furthermore, in faecal samples in which lactobacilli are detected, these organisms are found at low levels and account for a minor portion (approximately 0.01% to 0.6%) of the microbiota present (Lebeer *et al.*, 2008). It is therefore believed that only a small number of GIT-associated species are genuine residents of the GIT, and most are considered simply as allochthonous members that are passing through the GIT after originating from fermented food, the oral cavity, or more proximal parts of the GIT (Walter, 2008). In contrast, the presence of lactobacilli in animals and their GITs is more pronounced than in humans. For example, lactobacilli form stable populations in the GITs of pigs, mice, rats, and chickens at sites that are lined with stratified squamous epithelium (Walters, 2008). Although this epithelium type is absent from the human GIT, it is present in the buccal and vaginal cavities, sites at which lactobacilli are abundant (Walter, 2008).

Table 3. *Lactobacillus* species that are commonly detected in food, human faeces, and in various parts of the human GIT. The data were compiled from Molin *et al.*, 1993; Ahrne *et al.*, 1998; Reuter, 2001; Vaughan *et al.*, 2005; Dal Bello & Hertel, 2006; Walter, 2008; Ryan *et al.*, 2008; and Matsuda *et al.*, 2009. Brackets indicate the overall size of the resident lactobacillar populations according to Walter, 2005.

	Food	Oral cavity ($<10^3$-10^6)	Stomach ($<10^3$)	Small intestine ($<10^3$-10^8)	Colon ($<10^3$-10^9)	Faeces ($<10^3$-10^9)
L. acidophilus	+	+				+
L. crispatus		+				+
L. casei	+	+				+
L. fermentum	+	+				+
L. gasseri		+	+	+		+
L. paracasei	+	+			+	+
L. plantarum	+	+			+	+
L. reuteri	+		+	+		+
L. rhamnosus	+	+		+	+	+
L. ruminis			+			+
L. sakei	+					+
L. salivarius		+				+

1.2.6 Probiotic lactobacilli

By definition, probiotics are live microorganisms that have a beneficial effect on the health of the host when administered in adequate amounts (Lee & Salminen, 1995; Saxelin *et al.*, 2005; Giraffa *et al.* 2010). Although in theory any microorganism could be identified as probiotics, most probiotics in use today are lactobacilli and bifidobacteria (Siezen & Wilson, 2010). However, not all members of these genera are equally useful as probiotic additives. Preferably, the strains used should be of human origin and should have been proven safe for consumption (Lee & Salminen, 1995). In addition, probiotic additives should tolerate bile and acid to survive passage through the upper GIT and adhere to human tissues (Giraffa *et al.* 2010). Commercially distributed strains also should not adversely affect taste and should exhibit good growth characteristics and survive the production and storage processes used (Lee & Salminen, 1995). The last characteristic is particularly important because a bacterial concentration of $\geq 10^8$ colony-forming units per gram appears to be the efficacious dosage (Aureli *et al.*, 2011). Although the list of functional requirements is lengthy, at least some *Lactobacillus* strains appear to fulfil these criteria; see Table 4 for selected examples. These strains are typically delivered to customers as fermented dairy product or dietary supplements (Saxelin *et al.*, 2005; Giraffa *et al.* 2010); vaginal suppositories, cereals, and skin lotions representing other examples of products to which probiotic lactobacilli have been added (Krutmann, 2009; Rivera-Espinoza & Gallardo-Navarro, 2010; Salvatore *et al.*, 2011).

Many types of health benefit have been associated with the consumption of probiotic lactobacilli. These include the treatment of gastrointestinal infection and inflammatory bowel disease, the prevention of respiratory tract infections, the treatment of atopic diseases and allergy, the suppression of *Helicobacter pylori* infection, the prevention of urinary tract infections, anti-diarrheal properties and the treatment of bacterial vaginosis (Saxelin *et al.*, 2005; Uehara *et al.*, 2006; Anukam *et al.*, 2006; Siezen & Wilson, 2010; Aureli *et al.*, 2011). The potential mechanisms of action involved include a strengthening of the cell barrier function, vitamin supply, the enhancement of healthy microbiota, and antagonism against pathogens via the production of antimicrobials and competitive exclusion (Lebeer *et al.*, 2008; Ventura *et al.*, 2008; Oelschlaeger, 2010). Many probiotics are also believed to secrete immunomodulatory molecules that interact directly with the host (Yan *et al.*, 2007; Lebeer *et al.*, 2008). The physical interaction of cell-surface components with host tissues can also reinforce immune stimulation and provide positive health benefits. However, despite extensive research in the area of probiotics and the fact that probiotic products have been on the market since the creation of Yakult in 1935, none of the numerous commercial probiotic *Lactobacillus* products have been officially approved for their health-promoting claims in Europe. To date, the European Food Safety Authority has deemed only one general health claim valid; namely, “live cultures in yogurt or fermented milk improve lactose digestion”.

Table 4. Representative *Lactobacillus* strains associated with health-enhancing products. The data were compiled from Siezen & Wilson 2010 and Saxelin *et al.*, 2005.

Strain	Brand name	Claimed effect
L. casei Shirota	Yakult®	Alleviation of acute diarrhoea
L. rhamnosus GG	Gefilus®	Immune stimulation, alleviation of atopic eczema, prevention of diarrhoea, aleviation of symptoms associated with irritable bowel syndrome
L. acidophilus NCFM	Howaru®	Improvement of intestinal health, alleviation of symptoms associated with irritable bowel syndrome, gastrointestinal ecology
L. casei DN114-001	Actimel®	Diarrhoea treatment, gut infections, strengthening of the body's natural defences
L. reuteri 55730	Boost®	Alleviation of colic, pathogen inhibition

1.3 *Lactobacillus* genomes

The first 135 *Lactobacillus* genomes to be published included 38 finished and 97 draft genomes (Table 5 and Appendix Table 2). These genomes have enabled researchers to gain insight into the ecology, evolution, and biological role of lactobacilli and represent in total 46 different species that vary widely in GC content (32-53%), tRNA gene number (25 to 98), coding efficiency (49-91%), and genome size (1.2-3.8 Mb). Below, some of the most valuable discoveries from the first 135 *Lactobacillus* genome projects are outlined in detail.

L. delbrueckii is the largest of the 29 phylogenetic *Lactobacillus* groups (Salvetti *et al.*, 2012). The first genomes of this phylogenetic group to be resolved were the approximately 2.0-Mb genomes of *L. johnsonii* NCC 533 (Pridmore *et al.*, 2004) and *L. acidophilus* NCFM (Altermann *et al.*, 2005). Notably, these two GIT-associated isolates possess a bile salt hydrolase and various types of adhesins that support their persistence in the GIT. Their ability to synthesise cofactors and vitamins is in contrast limited and both isolates can synthesise only some amino acids *de novo*. These deficiencies are, however, alleviated by their broad repertoires of peptidases, proteases, and transporters, which allow efficient amino acid acquisition from the surrounding medium. As for these GIT-associated strains, genomes are available for several dairy-related isolates of the *L. delbrueckii* group. These include *L. delbrueckii* ATCC11842, which is known for its worldwide application in yogurt production (van de Guchte *et al.*, 2006), and *L. helveticus* DPC 4571, a Swiss cheese isolate that is recognised for its ability to reduce bitterness and increase flavour development in cheese (Callanan *et al.*, 2008). Intriguingly, *L. helveticus* DPC 4571 has few cell-surface-protein-encoding genes and does not encode bile salt hydrolase, despite its striking genome conservation with *L. acidophilus* NCFM (Callanan *et al.*, 2008). Analysis of the *L. delbrueckii* ATCC11842 genome has revealed an exceptionally high number of rRNAs, tRNAs, and partial carbohydrate-utilisation pathways, suggesting that this genome has undergone a recent phase of size reduction (van

Table 5. Characteristics of sequenced *Lactobacillus* groups and their genomes. See Appendix Tables 2 and 3 for a full description of each genome.

Clade name as in Salvetti *et al.*, 2012	Sequenced strains	Species with sequenced strains	Source	Length (Mb)	CDSs	Strain specific orthologue groups	RNA genes	16S
L. alimentarius	2	2/11	Sausage	2.5-2.5	2354-2440	290-349	55-57	1-2
L. brevis	2	1/10	Wine, silage	2.3-3.1	2218-3041	235-254	61-81	1-5
L. buchneri	5	4/12	Corn steep liquor, ethanol production plant, human, wine, oral cavity	2.6-3.0	2392-3325	134-574	44-78	0-5
L. casei	17	3/3	Dairy, corn steep liquor, human, koumiss, laboratory strain, starter culture	2.5-3.1	2719-3255	15-942	0-76	0-5
L. delbrueckii	61	11/27	Beer, dairy, fermented yak milk, human, kefir grain, porcine, poultry, starter culture	1.2-2.4	1144-2330	3-421	0-126	0-9
L. fructivorans	2	2/5	Sake, sourdough	1.4-1.4	1284-1358	127-163	39-82	1-7
L. malefermentans	1	1/1	Beer	2.0	1968	196	65	2
L. plantarum	6	2/5	Olives, silage, human, kimchi, cabbage	3.2-3.8	2755-3154	27-153	62-85	1-5
L. reuteri	23	9/15	Beets, plant material, human, kimchi, mouse, rat, silage, porcine	1.7-2.8	1051-2818	3-320	35-111	0-9
L. sakei	2	2/4	Sausage	1.8-1.9	1862-1885	145-173	56-84	1-7
L. salivarius	12	5/25	Apple juice, dairy, poultry, human, kimchi	1.9-2.7	1552-2642	17-440	29-120	1-8
L. caccinostercus	1	1/1	Apple mash	2.7	2534	368	58	1
Unclassified	1	1/1	Human	1.3	1181	20	76	1

de Guchte *et al.*, 2006). Finally, genomes for *L. crispatus*, *L. jensenii*, and *L. iners* isolates have provided novel perspectives on the genomic basis of urogenital lactobacilli. These genomes have, for example, revealed the lack of a complete bacteriocin synthesis apparatus and most known adhesion factors in *L. iners* AB-1 (Macklaim *et al.*, 2011) and the presence of three bacteriolysin and seven adhesion and colonisation-related protein encoding loci in the *L. crispatus* core-genome (Study V).

The genomes of the sequenced *L. casei* group members are approximately 2.8 Mb in size (Table 5) and appear to all harbour a repertoire of genes that are involved in sugar uptake, carbohydrate utilisation, and amino acid biosynthesis (Makarova *et al.*, 2006; Cai *et al.*, 2009; Morita *et al.*, 2009). Importantly, pilus gene clusters similar to those that were disclosed in the genome study of *L. rhamnosus* GG (Study III) are also widespread and have hitherto been identified in the genomes of all probiotic (Douillard *et al.*, 2013a) and various other (Douillard *et al.*, 2013b; Kant *et al.*, 2014) members of this *Lactobacillus* phylogenetic group. These genome investigations have also provided insights into the genetic complexity of these species and have been helpful in defining the scale and scope of biomedically important effector molecules in the *L. casei* group that have previously been associated with a cytokine production (Péant *et al.*, 2005), promotion of *in vitro* intestinal epithelial homeostasis (Yan *et al.*, 2007; Lebeer *et al.*, 2008), and adaptation of strain GG to the host environment (Lebeer *et al.*, 2009). Other genome studies have on the other hand provided valuable insights into the genomic niche-associated evolution of *L. casei* and have determined the scale and scope of genetic variation of this versatile and important species (Cai *et al.*, 2009).

The *L. salivarius* group is the second most speciose group of the genus *Lactobacillus*, and includes representatives of both non-motile (13) and motile (12) *Lactobacillus* species (Salvetti *et al.*, 2012). The genomes associated with this phylogenetic group range from 1.9 to 2.7 Mb and can consist of multiple large replicons. For example, a megaplasmid comprises 11% of the genome of *L. salivarius* UCC118 (Claesson *et al.*, 2006). Although this megaplasmid contains no essential genes, it encodes functions that are beneficial to the host cell, such as the ability to use additional sugars, to hydrolyse bile salt, and to produce supplementary amino acids. The plasmid also includes a locus that encodes and is required for the synthesis of bacteriocin (Claesson *et al.*, 2006), which is active against *Listeria monocytogenes* (Corr *et al.*, 2007). Analysis of the genome of *Lactobacillus ruminis* ATCC 27782 has been highly useful in understanding the nature of flagellum-mediated motility in the genus *Lactobacillus* and has revealed a complete set of flagellum biogenesis genes (Forde *et al.*, 2011). The genome study has also described genes that share similarity with known pilin genes, suggesting that *L. ruminis* ATCC 27782 might contain a sortase-dependent pilus organelle.

Genomes in the *L. reuteri* group are on average 2.1 Mb and contain on average 2,055 CDSs (Table 5). They appear to be rich in genes encoding putative cell-surface-associated proteins (Båth *et al.*, 2005; Saulnier *et al.*, 2011), and some strains possess an EPS gene cluster that is involved in modulating host immune responses (Saulnier *et al.*, 2011). In addition, genomic island coding for the production of a broad-spectrum antimicrobial substance termed reuterin has been identified in selected *L. reuteri* strains. This reuterin island is particularly interesting because the production of reuterin appears to be linked to

the prevention of gastrointestinal infections (Saulnier *et al.*, 2011). Moreover, genomic analyses have revealed loci that are associated with the production of vitamins B_1 (Saulnier *et al.*, 2011) and B_{12} (Morita *et al.*, 2008), although it remains an open question whether *L. reuteri* produces an active (Mohammed *et al.*, 2014) or inactive (Santos *et al.*, 2007) form of vitamin B_{12}.

Other economically or biologically important lactobacilli are found in the *L. sakei* and *L. plantarum* groups (Table 5). These include the psychrotrophic bacterium *L. sakei* 23K (Chaillou *et al.*, 2005). Based on its genome, this organism has metabolic pathways for arginine catabolism and purine nucleoside scavenging. Other notable features include genes implicated in dealing with the harsh conditions associated with food processing and allowing growth on meat during refrigeration and in the presence of curing salts. Genome analysis have also revealed auxotrophy for all amino acids except aspartate and glutamic acid (Chaillou *et al.*, 2005). Another noteworthy strain is *L. plantarum* WCFS1, which is the first organism from the *Lactobacillus* genus to be sequenced (Kleerebezem *et al.*, 2003). The genome of *L. plantarum* WCFS1 is large, close to 3.3 Mb in size, and appears to encode a number of transporters and enzymes for the uptake and utilisation of sugars, thus providing an explanation for the widespread distribution of the *L. plantarum* species in nature (Kleerebezem *et al.*, 2003). The authors also described a number (>200) of extracellular proteins that could enable exchange signals with the environment and adherence to surfaces (Kleerebezem *et al.*, 2003).

1.3.1 Computational genomics of *Lactobacillus*

The current *Lactobacillus* genome data were largely obtained using Sanger and Roche 454 sequencing technologies (Appendix Table 3). The first 15 genomes were chiefly determined using Sanger technology (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004; Altermann *et al.*, 2005; Chaillou *et al.*, 2005; Makarova *et al.*, 2006; Claesson *et al.*, 2006; van de Guchte *et al.*, 2006; Frese *et al.*, 2011; Macklaim *et al.*, 2011; Morita *et al.*, 2008); after these genomes, sequencing has typically relied on the Roche 454 platform. For example, in 2009, most *Lactobacillus* genome projects were obtained using the Roche 454 platform (Appendix Table 3). Surprisingly, sequence data were generated using the Illumina sequencing platform alone in only three cases (part of the human microbiome project (HMP); Nelson *et al.*, 2010). As expected, Newbler represents the most popular assembler. In addition, Phrap, Jazz, Velvet, and CLC bio genome assemblers have been employed in more than two genome projects (Appendix Table 3). Among the 97 draft assemblies, the median contig number is 75. Given their high degree of sequence similarity and synteny with finished assemblies, draft assemblies are however presumed to offer a near-complete picture of the genome. The unresolved regions of these genomes most likely comprise primarily long, repetitive sequences, such as those encoding ribosomal genes, MGEs, and repetitive structures of some surface-protein genes (Seepersaud *et al.*, 2005; Edelman *et al.*, 2012). The most notable exception is the genome assembly of *L. rhamnosus* MTCC 5462, which comprises 2,543 contigs and might include a notable number of contigs that end in incomplete gene sequences (Prajapati *et al.*, 2012).

As expected for sequences with GC contents ranging from 32 to 59% (Marine *et al.*, 2011), the genomic GC content has not affected assembly quality, and no obvious correlation was detected between the contig numbers and the GC contents of different *Lactobacillus* genomes.

Regarding genome annotation, no single bioinformatic tool or approach has gained supremacy (Appendix Table 3). Structural annotation has relied on several gene callers, mainly Glimmer, GeneMark, tRNAscan-SE, and RNAmmer, as was the case in the 55 genome projects that were processed using the BCM, JCVI, or PGAAP genome annotation pipelines. Functional classification has primary been performed using BLAST and domain search tools. On occasion, protein functions have been amended manually (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004; Altermann *et al.*, 2005; Chaillou *et al.*, 2005; Makarova *et al.*, 2006; Claesson *et al.*, 2006; van de Guchte *et al.*, 2006; Frese *et al.*, 2011; Macklaim *et al.*, 2011; Morita *et al.*, 2008; Cai *et al.*, 2009; Study III; Morita *et al.*, 2009; Study IV; McNulty *et al.*, 2011; Forde *et al.*, 2011). Overall, the level of manual curation is higher for finished genomes and genomes that were processed before 2010, as reflected by their higher rates of start site consistency and exact DEs. In particular, more attention was given to the annotation of the first 15 *Lactobacillus* genomes. The initial annotations were in these projects generated using a variety of methods and curated manually. Many of the more recent genomes have on the other hand chiefly relied on the non-human assisted use of protein function prediction services. However, annotation processes are generally poorly documented, which complicated the comparison of annotation processes. For 14 genomes, no information was offered; for another 67 genomes, only the general protocol of the sequencing centre that had performed the sequencing was listed. Significantly, in various instances, neither genome annotations nor sequences have been updated since the data release by means other than the automatic annotation systems used by the databases. Annotations between different genomes can thus be inconsistent, even for strains of the same species. The most notable exception is the genome of *L. plantarum* WCFS1, which was re-sequenced and re-annotated by Siezen *et al.* recently (Siezen *et al.*, 2012).

1.3.2 Comparative genomics of *Lactobacillus*

Comparative analyses of *Lactobacillus* have expanded our understanding of the molecular evolution, diversity, and function of lactobacilli. Importantly, these studies have revealed significant functional and genomic variance between *Lactobacillus* genomes (Boekhorst *et al.*, 2004; Pridmore *et al.*, 2004; Canchaya *et al.*, 2006; Berger *et al.*, 2007; Ventura *et al.*, 2008; Canchaya *et al.*, 2006; Claesson *et al.*, 2008; Morita *et al.*, 2008; Azcarate-Peril *et al.*, 2008; O'Sullivan *et al.*, 2009; Kant *et al.*, 2011a; Lukjancenko *et al.*, 2012). For example, only 28 large regions of conserved gene order, ranging in size from 7 to 75 genes, were found in a comparative analysis of *L. plantarum* WCFS1 and *L. johnsonii* NCC 533 genomes (Boekhorst *et al.*, 2004). A lack of gene order synteny has also been shown for other distant lactobacilli, whereas the genomes of closely related lactobacilli tend to exhibit a high degree of gene order conservation across their entire genomes

(Canchaya *et al.*, 2006; Berger *et al.*, 2007; Ventura *et al.*, 2008). However, even the genomes of closely related strains of the same lactobacilli species can differ by one or more genomic islands (Pridmore *et al.*, 2004; Study III; Morita *et al.*, 2008; Azcarate-Peril *et al.*, 2008). These regions of diversity typically lack similarity with other lactobacilli regions and contain genes that might be relevant for environmental adaptation. For example, genomic islands contain genes coding for EPS biosynthesis in *L. gasseri* ATCC 33323 (Azcarate-Peril *et al.*, 2008), fimbrial components in *L. johnsonii* NCC 533 (Pridmore *et al.*, 2004), pilus fibres in *L. rhamnosus* GG (Study III), and B_{12} and reuterin biosynthesis in *L. reuteri* JCM 1112^T (Morita *et al.*, 2008). In addition, horizontal gene transfer has been linked to the ability of *L. plantarum* WCFS1 to adapt to a variety of niches, to the adaptation of *L. delbrueckii* to dairy environments (Liu *et al.*, 2009; van de Guchte *et al.*, 2006), and to the acquisition of amino acid metabolism, lipid biosynthesis, and restriction endonuclease genes during the evolution of *L. helveticus* DPC 4571 (Callanan *et al.*, 2008). However, predicting laterally acquired genome regions is not always straightforward. For example, the sequencing and analysis of the genomes of *L. ruminis* ATCC 27782 and ATCC 25644 identified a bacteriocin cluster in ATCC 27782, although it remained unclear whether ATCC 25644 also contained a complete bacteriocin locus; some genes associated with bacteriocin production were missing and the others did not assemble into a single contig in the latter strain (Forde *et al.*, 2011).

Comparative genomics studies have revealed interesting biological similarities and differences among lactobacilli. The lactobacilli genomes, for example, appear to contain a constant fraction of flavour-related genes, independently of their isolation. The approximately 2.0-Mb genomes of *L. delbrueckii* ATCC 11842 and *L. acidophilus* NCFM, for example, code for 15 and 14 flavour-related enzymes (Liu *et al.*, 2008), even though the first is a dairy isolate and the second is a human isolate. Analysis of *Lactobacillus* secretomes has revealed that on average, 8% of a lactobacilli proteome represents secreted or surface-associated proteins (Zhou *et al.*, 2010; Kant *et al.*, 2010). The largest predicted secretome (approximately 9% of the predicted proteome) is that of *L. acidophilus* NCFM, whereas the smallest predicted secretome (approximately 5% of the predicted proteome) was that of *L. reuteri* DSM 20016 (Kant *et al.*, 2010). Comparative genomics has been used to measure the diversity of CRISPR-cas systems, revealing that approximately two thirds of the analysed *Lactobacillus* strains have a CRISPR locus (Horvath *et al.*, 2009). Mucus-binding domain screens have been used to measure the prevalence of mucus-binding proteins in lactobacilli; the most abundant mucus-binding domain-containing proteins were found in GIT-associated lactobacilli, supporting the idea that mucus-binding proteins are involved in adherence to the intestinal mucus that covers intestinal epithelial cells (Boekhorst *et al.*, 2006).

1.3.3 Comparative core and pan-genomics of *Lactobacillus*

The microbial pan-genome comprises a core and an accessory gene pool (Tettelin *et al.*, 2005). Core genes are conserved across all isolates of a given group of organisms, whereas accessory genes are present in some but not all strains (Tettelin *et al.*, 2005). In general,

core genes are responsible for basic cellular processes and the main phenotypic traits of the group, whereas accessory genes contribute to diversity within the group and enable adaptation to specific environments (Medini *et al.*, 2005). When applied to *Lactobacillus* genomes, core genome investigations have defined differing sets of 593 (Canchaya *et al*., 2006), 141 (Claesson *et al.*, 2008), 383 (Kant *et al.*, 2011a), and 363 (Lukjancenko *et al.*, 2012) core genes that are shared by all lactobacilli. These apparently contradictory results are tentatively explained by methodological differences and by the number of genomes analysed in the individual studies; 5 by Canchaya *et al*., 2006, 12 by Claesson *et al.*, 2008, 20 by Kant *et al.*, 2011a, and 20 by Lukjancenko *et al.*, 2012. Surprisingly, the impact of genome number and the consequence of sequentially adding more genomes and the core genome size of an infinite number of *Lactobacillus* genomes were not estimated in any of the studies, as is typically performed in core and pan-genome studies (Medini *et al.*, 2005). Nevertheless, these and other comparative studies have provided key insights into the genome evolution of *Lactobacillus* and revealed that the gene complements are the results of extensive gene losses and gains during evolution (Claesson *et al.*, 2008; Kant *et al.*, 2011a) and are devoid of habitat-specific genes (Claesson *et al.*, 2008; O'Sullivan *et al.*, 2009), and that the core genome contains only a few genus-specific genes (Claesson *et al.*, 2008; Canchaya *et al*. 2006).

The core genomes of specific *Lactobacillus* species have also sparked interest. The establishment of the level of intraspecies diversity in seven *Lactobacillus* species using comparative genomic hybridisation and the mapping of short-read sequences to reference genomes revealed that individual strains lack 3-24% of the genes of the given reference genome and that the size of the core genome is correlated with the total size of a single reference genome (Siezen *et al.*, 2010; van Hemert *et al.*, 2010; Meijerink *et al.*, 2010; Berger *et al.*, 2007; Cai *et al.*, 2009; Raftis *et al.*, 2011; Nyquist *et al.*, 2011; Frese *et al.*, 2011; Douillard *et al.*, 2013b). Additionally, whole-genome assembly comparisons have been used for the investigation of genetic diversity in *L. paracasei* (Smokvina *et al.*, 2013), *L. rhamnosus* (Kant *et al.*, 2014), and *L. casei* (Broadbent *et al.*, 2012). This approach is more powerful than read-mapping and microarray-based approaches and can call genes that are present in genomes other than the reference genome. Using this approach, gene loss and gain was found to be the dominant force in genome evolution within both *Lactobacillus* species, although some variation took the form of differences in universally conserved genes. For example, *L. casei* strains exhibited >99% identity to the 16S rRNA sequence of *L. casei* ATCC 334. Nevertheless, on average, 119 strain-specific gene families are present in each genome, and only 61% of any individual genome is shared by all 17 strains. Mathematical modelling of the data indicates extensive genome diversity and that the *L. casei* pan-genome contains 9,072 gene families, of which 1,600 are common to all individuals (Broadbent *et al.*, 2012).

2 Aims of the study

The aim of this study was to develop algorithms for the automated function prediction of bacterial protein sequences and to advance our understanding of *Lactobacillus* physiology by annotating the genomes of *L. rhamnosus* GG and LC705 and *L. crispatus* ST1. Genes in the genomes that exhibited functions that are involved in interactions with the host and that responsible for strain-specific characteristics were of particular interest. To determine the scale and scope of genetic variation in *L. crispatus* and to infer physiological traits that are common for all *L. crispatus*, pan-genomic strategies were also considered. Overall, the goal was to develop efficient and accurate ways to annotate bacterial genomes.

3 Materials and methods

The bioinformatics methods developed are described in detail with appropriate references in original publications I and II. Detailed descriptions of materials and methods used in the genome investigations are in the original papers III-V.

3.1 Evaluation of bioinformatics methods

A comprehensive assessment of the newly developed bioinformatics methods is available in the original publications I and II. Briefly, the performance of the LOCP software tool (Study I) was evaluated based on 20 completely or partially sequenced genomes with known pilus operons. In this process, the genes in these genomes were assigned scores of their highest ranked LOCP predictions or were assigned a score of zero indicating that the gene was not reported by LOCP. The ability of LOCP to distinguish genuine pilus-related from other genes was then assessed using receiver operating characteristic analysis. In Study II, the data used in the development process of BLANNOTATOR were obtained from UniProt (Bairoch *et al.*, 2008). Protein function labels were restored to the state preceding the functional characterisation of the test sequences using scripts that were developed in-house. The details of the restoring process as well as the use of competing protein function classification methods and statistical tests are in the original paper II. In addition, a retrospective assessment of the classification accuracy of nine gene-calling systems was performed, and this assessment is described in this thesis. Specifically, CDSs in *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 were called using nine gene-calling tools, and the sensitivity, specificity, and F-score (*i.e.,* the harmonic mean of the sensitivity and specificity) were computed at the single-base level. The gold standard sets that were used in the comparison were the manually annotated gene models from the genomes as well as regions with and without transcriptional activity in *L. rhamnosus* GG. Transcriptome data were compiled from three previous *L. rhamnosus* GG gene expression studies (Laakso *et al.*, 2011; Koskenniemi *et al.*, 2011; Koponen *et al.*, 2012) and comprised 626,450 bases with evidence of transcription (defined as bases that are covered by probes with an intensity value ≥5 standard deviations above the mean intensity of negative control probes in ≥18 of the 360 RNA samples described in these three studies).

3.2 Strains and growth conditions

The *Lactobacillus* strains that were sequenced in Studies III-IV included *L. rhamnosus* GG (ATCC 53103), *L. rhamnosus* LC705 (DSM 7061), and *L. crispatus* ST1. To prepare genomic DNA, each strain was grown in the de Man, Rogosa, and Sharpe (de Man *et al.*, 1960) broth at 37°C. DNA was extracted as previously described (Pitcher *et al.*, 1989).

3.3 Sequencing and assembly

Whole-genome sequencing was done at the DNA sequencing and Genomics Laboratory at the Institute of Biotechnology, University of Helsinki. Briefly, DNA from *L. rhamnosus* GG and LC705 was processed according to publication III. Plasmid and fosmid libraries were sequenced using an ABI 3730 DNA sequencing instrument and Big Dye chemistry (Applied Biosystems, Foster City, CA, USA), whereas genomic fragment libraries were sequenced using Roche GS 20 pyrosequencing (454 Life Sciences/Roche Applied Biosystems, Branford, CT, USA). In Study IV, genomic DNA from *L. crispatus* ST1 was used for 454 library construction and sequenced using a Roche 454 instrument with GS FLX chemistry (454 Life Sciences/Roche Applied Biosystems, Branford, CT, USA). All DNA reads were processed and assembled using the Staden Package (Staden *et al.*, 1999) and/or Newbler (454 Life Sciences/Roche Applied Biosystems, Branford, CT, USA). For gap closure, the PCR-amplified fragments obtained using genomic DNA were sequenced using an ABI 3730 instrument and Big Dye chemistry (Applied Biosystems, Foster City, CA, USA).

3.4 Accession numbers for the submitted data

The genome sequences of *L. rhamnosus* GG and LC705 and the plasmid pLC1 have been deposited in the EMBL nucleotide sequence database under accession numbers FM179322, FM179323, and FM179324, respectively. The genome sequence of *L. crispatus* ST1 has been deposited in the EMBL nucleotide sequence database under accession number FN692037.

3.5 Publicly available genome sequences

The genome data used in Studies III-V were downloaded from the GenBank database (Benson *et al.*, 2013) and the PATRICK database (Gillespie *et al.*, 2011), as indicated in the original publications. The genome data referenced in this dissertation were downloaded in April 2012 from GenBank (Benson *et al.*, 2013) and the PATRICK database (Gillespie *et al.*, 2011). The NCBI database was preferred over the PATRICK database when a genome was available at both databases, provided the NCBI entry was properly annotated. Where possible, scaffold assemblies were preferred.

3.6 Structural and functional annotation

The genomes in Studies III and IV were scanned for CDSs, tRNAs, rRNAs, CRISPRs, genomic islands, prophage-like clusters, and rho-independent transcription terminators

using an array of computational methods (Table 6). Predicted protein sequences were then searched against a variety of databases and further processed using a set of bioinformatics tools with the aim of assigning function (Table 6). Automated computer annotations were verified, and discrepancies were resolved manually. Detailed descriptions of databases and software tools used are in the original papers III-IV. In Study V, genome sequences were mined for CRISPRs, genomic islands, plasmids, and prophage-like regions using a variety of methods (Table 6). A functional annotation update was also performed to ensure that protein function predictions were of identical quality for all of the investigated *L. crispatus* genomes. An overview of the computational methods used in Study V is presented in Table 6. The use of these tools is described in detail in the corresponding publication. Where feasible, the methods developed in Studies I and II were used to obtain details about pilus-like gene clusters and protein functions, respectively, in Studies III-V.

3.7 Metabolic pathway reconstruction

Using the KAAS tool (Moriya *et al.*, 2007), CDSs within the genomes of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 were annotated for EC numbers describing enzymatic activity. Metabolic pathway reconstructions were then realised by associating enzymatic activities with combined KEGG (Kanehisa *et al.*, 2004) and MetaCyc (Krieger *et al.*, 2004) reference reaction pathways and by manually investigating reaction maps. To resolve the precise metabolic activity of genes with partial EC codes, the enzymatic activity supported by the API 50 CH (BioMérieux, Marcy l'Etoile, France) carbohydrate-fermenting patterns was chosen where possible. In Study V, EC codes were determined using KAAS software (Moriya *et al.*, 2007) and the FMM server was used to assemble these EC codes to metabolic pathways (Chou *et al.*, 2009).

3.8 Comparative analyses

The methods used in Studies III-V for orthologue analysis, phylogenetic tree construction, and whole-genome alignment are summarised in Table 6. The *L. rhamnosus* whole-genome nucleotide alignments that are presented in this thesis were generated using BLASTN (Altschul *et al.*, 1997) and ACT (Carver *et al.*, 2005). The *L. rhamnosus* draft genomes were ordered and oriented with respect to the genome sequence of *L. rhamnosus* GG using progressive Mauve (Rissman *et al.*, 2009). To produce the *L. crispatus* whole-genome nucleotide alignments that are presented in this thesis, matching genome blocks between the genome of *L. crispatus* ST1 and the genomes of *L. crispatus* CTV-05, *L. helveticus* DPC 4571, *L. acidophilus* NCFM, *L. johnsonii* NCC 533, *L. gasseri* ATCC 33323, and *L. delbrueckii* ATCC 11842 were identified using PROmer and visualised using a MUMmer plot (Kurtz *et al.*, 2004). The draft genome of *L. crispatus* CTV-05 was ordered and oriented with respect to the genome sequence of *L. crispatus* ST1 using progressive Mauve (Rissman *et al.*, 2009).

3.9 *Core and pan-genome analyses*

Orthologue and paralogue groups among the *L. crispatus* genomes were in Study V identified using BLASTP (Altschul *et al.*, 1997) and OrthoMCL (Li *et al.*, 2003). To estimate the development of the size of the core and pan-genome as a function of the number of sequenced strains, orthologue and paralogue groups were determined iteratively for increasing numbers of sequenced genomes. At each sample size, the analysis was repeated 50 times with different random sets of *L. crispatus* genomes. The core genome trend was extrapolated by fitting an exponential decay (Tettelin *et al.*, 2005) to the medians of the core orthologue groups using a weighted least-squares regression. The number of pan-groups in an infinite number of *L. crispatus* genomes was predicted by fitting a power-law (Tettelin *et al.*, 2005) to the pan-group medians using a weighted least-squares regression. The regression analyses were performed using the *nls* function as implemented in the statistical software R (Ihaka & Gentleman, 1996). The methods used for identifying the orthologue and paralogue groups among the *Lactobacillaceae* genomes and for estimating the *Lactobacillaceae* core and pan-genome sizes were the same as those described in Study V with the exception of using a double exponential decay (Bottacini *et al.*, 2010) for the core genome data. The results of *Lactobacillaceae* analyses are presented in this thesis.

Table 6. Methods used for annotation of the *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* genomes.

	L. rhamnosus GG	*L. rhamnosus* LC705	*L. crispatus* ST1	*L. crispatus*
Structural annotation				
CDS	Glimmer, BLAST, ERGO	Glimmer, BLAST	Glimmer, BLAST, RAST	-
rRNA	RNAmmer, ERGO	RNAmmer	RNAmmer, RAST	-
tRNA	tRNAscan-SE, ERGO	tRNAscan-SE	tRNAscan-SE, RAST	-
CRISPR	PILER-CR	PILER-CR	PILER-CR	PILER-CR
Intrinsic terminators	TransTermHP	TransTermHP	TransTermHP	-
Plasmid related contigs	-	-	-	cBar
Genomic islands	PAI-IDA	PAI-IDA	PAI-IDA	IslandViewer
Functional classification				
General function prediction	BLAST, ERGO, InterProScan, COG	BLAST, PFAM, InterProScan, COG	RAST, COG InterProScan, BLAST, BLANNOTATOR	RAST, BLAST, COG, BLANNOTATOR
Gene Ontology terms	InterProScan to GO	InterProScan to GO	InterProScan to GO	-
Transporters	TransAAP, TCDB	TransAAP, TCDB	TCDB	TCDB
Enzymes	Merops, CaZY, KAAS	Merops, CaZY, KAAS	Merops, CaZY, KAAS	Merops, KAAS
cas genes	TIGRfam	TIGRfam	TIGRfam	TIGRfam
Bacteriocins	InterProScan	InterProScan	InterProScan, BLAST	BAGEL
Prophages	Prophinder	Prophinder	Prophinder	Prophinder
Adhesion factors	PFAM	PFAM	PFAM (revised)	PFAM (revised)
Secretome	SignalP, LipoP, TMHMM	SignalP, LipoP, TMHMM	LocateP	-
Other analysis				
Ortholog assignment	InParanoid	InParanoid	OrthoMCL, Cliquer	OrthoMCL
Phylogenetic analysis	BLASTP, PHYLIP	BLASTP, PHYLIP	Muscle, Gblock, PhyML	Mauve, PhyML
Whole-genome alignment	BLASTN, ACT, Gepard	BLASTN, ACT, Gepard	BLASTN, ACT, Gepard, Mauve	BLASTN, ACT, Mauve
Metabolic pathways	KEGG, MetaCyc	KEGG, MetaCyc	KEGG, MetaCyc	FMM

4 Results and discussion

The main objective of this study was to develop algorithms for protein classification that go beyond the present practice in several aspects and to advance our understanding of *Lactobacillus* physiology with the aid of genome sequencing and comparative genomics approaches. In particular, novel insights into the physiology, ecology, and biochemistry of lactobacilli were provided by comparing the genome of the widely studied (Bernardeau *et al.*, 2006) and commercially significant (Saxelin *et al.*, 2005; Giraffa *et al.* 2010) probiotic bacterium *L. rhamnosus* GG with the genome of the industrial dairy strain *L. rhamnosus* LC705 (Suomalainen & Mäyrä-Mäkinen, 1999). Furthermore, the comparative analysis of *L. crispatus* genomes unveiled cross-species conserved adhesion components that could protect the vagina from pathogen attack. The comparison also highlighted an array of other interesting differences and similarities between the vaginal *L. crispatus* isolates and *L. crispatus* ST1, which is known for its strong adherence to the chicken alimentary canal and to human vaginal and buccal cells (Edelman et al., 2012). In the following chapters, the main findings of these studies are presented in detail. First, the bioinformatics approaches used are described; then, the outcomes of the *Lactobacillus* genome studies are summarised.

4.1 Novel tools for predicting the function of bacterial proteins

Recently developed protein function prediction methods provide a comprehensive and efficient means for inferring protein function (see for example, Abascal & Valencia, 2003; Kunin & Ouzounis, 2005; Martin *et al.*, 2004; Vinayagam *et al.*, 2006; Hawkins *et al.*, 2006; Wass & Sternberg, 2008; Engelhardt *et al.*, 2005; Zdobnov & Apweiler, 2001) but can sometimes fail to characterise pilins (Scott & Zähner, 2006) and produce unreliable functional calls (Friedberg, 2006; Rost, 2002; Rost *et al.*, 2003). To overcome these difficulties, new computational approaches were developed for pilus operon (Study I) and protein function (Study II) prediction. As described in the following chapters, these newly developed bioinformatics tools provided many interesting insights into the physiology of *L. rhamnosus* and *L. crispatus*.

Pili are long filamentous protein assemblies that are located on the surface of bacteria and are often involved in the adhesion of bacteria to host cells (Madigan *et al.*, 2010). In gram-positive bacteria, these structures typically comprise one major pilin protein and two auxiliary pilin proteins that are cross-linked by a sortase enzyme (Telford *et al.*, 2006; Scott & Zähner, 2006). It appears that the pilin genes are usually located in an operon with a sortase gene (Scott & Zähner, 2006) and that their protein products display various features that are characteristic of gram-positive pilins, such as a positively charged tail and a membrane-spanning domain at the C terminus, an E box, a sortase recognition site (a LPXTG motif), a pilin motif, and a Sec-dependent secretion signal peptide (Telford *et al.*, 2006). In addition, sequence comparisons with gram-positive pilins have revealed conserved sequence motifs that stabilise the structure (Kang *et al.*, 2009). Although these

motifs have enabled the search for putative pilus operons (Ton-That & Schneewind, 2003), a novel bioinformatic algorithm was proposed in Study I for the systematic screening of pilus operons in gram-positive genomes. This tool, LOCP, which is written in Perl, scans sequences with five pilin-, five sortase recognition site-, and three sortase enzyme profile HMMs. Each sequence is labelled either as a hit (if at least one HMM model is matched) or as a miss, after which genome regions that are enriched with hits are identified based on a hypergeometric distribution. The use of a discrete distribution to locate gene runs that are statistically enriched in sequence features borrows from the concept of a previously described prophage-finding program (Lima-Mendez *et al.*, 2008) and was shown to distinguish genuine pilus clusters precisely from other genome regions. Specifically, LOCP identified all 28 genuine pilus clusters and made no false predictions in the given 20 evaluation genomes (the area under the curve was approximately 0.99). To date, no other bioinformatics tools other than LOCP have been developed for locating pilus operons in bacterial genomes.

To facilitate bacterial genome annotation, a new protein function classification system was created in Study II. Leveraging the advantages of DEs and GOs, the software BLANNOTATOR generates predictions based on database hits that are associated with consistent protein function information. This strategy is similar to those employed in the CLAN (Kunin & Ouzounis, 2005) and ConFunc (Wass & Sternberg, 2008) protein classification systems and is known to be less error-prone to annotation anomalies than methods that rely on a single or all database hits, such as the PFP (Hawkins *et al.*, 2006) and the ARGOT2 (Falda *et al.*, 2012) tools. Importantly, BLANNOTATOR was exceedingly helpful in systematising DEs and generated precise function predictions in evaluation tests. When applied to the predicted proteins of *L. crispatus* ST1, the algorithm assigned a biologically acceptable DE for 85% of the query sequences. In comparison, RAST- (Aziz *et al.*, 2008) or BLAST-based approaches provided a valid function prediction for approximately only 58 and 69% of the proteins in the test set. This method was particularly useful for predicting the function of proteins for which the top database hits were uninformative, as discussed in detail in Study II. The accuracy of BLANNOTATOR was further benchmarked by simulating the annotation process for more than 3,000 high-quality annotated bacterial protein entries and by assessing the ability of BLANNOTATOR to reproduce current annotations based on annotation information that predated the functional characterisation of the test entries. For this dataset, the method produced more precise predictions than any of the five other function classification approaches (*i.e.,* most significant BLAST match, the top informative BLAST match, the most common annotation among BLAST hits, the annotation associated with the highest cumulative BLAST bit score, and a word-based scoring scheme) that were tested. However, performance differences between the protein function prediction approaches were marginal, and even the worst performing approaches provided reasonable accuracy, indicating that major improvements in the field of homology-based function transfer are less likely to occur in future.

4.2 *L. rhamnosus* and *L. crispatus* genome sequencing

The objectives of Studies III (initiated in 2004) and IV (initiated in 2008) were to produce finished genomes for the previously undescribed isolates of *L. rhamnosus* and *L. crispatus* using the WGS sequencing and assembly strategy (Fleischmann *et al.*, 1995). To achieve this, the genomes of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 were sequenced using a combination Roche pyrosequencing and Sanger sequencing. These platforms provided approximately 17-19× read sequence coverage for each genome (Table 7) and were preferred in the process over Illumina sequencers because of their long read length (Shendure & Ji, 2008). A collection of PCR-assisted techniques was then applied to improve genome assemblies, providing to resolve all the gaps in the draft genomes of both *L. rhamnosus* GG and LC705 and all but one gap in the draft genome of *L. crispatus* ST1. Despite a single gap region remaining in the *L. crispatus* ST1 assembly, each assembly provided valuable information about the general genome features of the organism (Table 7). They also represented an important step forward in the biology of that species because, before their release, only two *L. rhamnosus* and five *L. crispatus* genome sequences had been deposited in public sequence databases (mainly generated by the HMP; Nelson *et al.*, 2010).

L. rhamnosus GG is one of the most extensively studied lactobacilli strains (Saxelin *et al.*, 2005; Bernardeau *et al.*, 2006) and has shown promising results for the treatment or prevention of respiratory tract infections (Hatakka *et al.*, 2001; Hojsak *et al.*, 2010), certain types of diarrhoea (Isolauri *et al.*, 1991; Guandalini *et al.*, 2000; Szajewska & Mrukowicz, 2001), and atopic diseases (Kalliomäki *et al.*, 2001; Kalliomäki *et al.*, 2003; Kalliomäki *et al.*, 2007). The identity of the specific effector molecules behind these beneficial effects was however chiefly lacking prior to the Study III, illustrating the value of knowing its genome sequence. Indeed, the determination and annotation of the genome of *L. rhamnosus* GG uncovered several genes of potential biomedical importance and expanded our knowledge of the *L. rhamnosus* bacterial components far beyond the few earlier *in vitro* verified instances (Vélez *et al.*, 2007; Yan *et al.*, 2007; Iliev *et al.*, 2008; Lebeer *et al.*, 2009). At the summary statistics level (Table 7), strain GG was identified to be comparable to other *L. casei* and *L. paracasei* and *L. rhamnosus* strains (reflecting the close phylogenetic relationships among these strains). It was observed to have one of the largest *Lactobacillus* genomes and it was predicted to contain only a slightly fewer CDSs and tRNA genes than the largest known lactobacilli genome from *L. plantarum* WCFS1 (Kleerebezem *et al*., 2003). A large majority (78%) of CDSs were predicted to start with an ATG and approximately 20% of all CDSs were preceded by a putative RBS (Figure 7), which is defined here as a DNA sequence located at a maximum of -20 bases from the start codon and showing strong similarity to the genome-specific RBS position weight matrix motif model. Genome mining of other genomic features revealed three prophage-like regions and a CRISPR locus that was notably similar to that of *L. salivarius* UCC118 and *L. casei* BL23 (Horvath *et al.*, 2009).

Table 7. Comparison of the genomic features of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1. Values presented for *L. rhamnosus* and *L. crispatus* are averages for the remaining seven *L. rhamnosus* and seven *L. crispatus* genomes, which are listed in Appendix Table 2.

	L. rhamnosus GG	*L. rhamnosus* LC705	*L. rhamnosus*	*L. crispatus* ST1	*L. crispatus*
ACC	FM179322	FM179323	NA	FN692037	NA
Rank (clade / all lactobacilli)	3 / 51	3 / 51	NA	6 / 63	NA
Read coverage	17	19	NA	18	NA
Genome size, Mbp	3010111	3033106	2892564	2043161	2234576
Scaffolds	1	1	419	1	114
Plasmids	0	1	NA	2	NA
GC content %	47	47	47	37	37
Coding efficiency %	85.00	85.00	80.14	88.00	83.57
CDS	2944	2992	2946	2024	2194
ATG/TTG/GTG %	77/11/12	78/11/11	74/10/11	86/9/5	83/10/7
Function Predicted %	73	72	69	77	67
Enzyme %	24	25	25	24	22
Transporter %	14	15	14	13	13
Transmembrane %	27	28	29	26	2
Average CDS size bp	870	865	800	892	852
rRNA operons	5	5	2-5	4	1-4
tRNA genes	57	61	51	64	57
CRISPR loci	1	0	1	2	1

The dairy-associated strain *L. rhamnosus* LC705 is widely used in the manufacture of cheese products (Saxelin *et al.*, 2011). LC705 also is one of the main components of a bacterial multispecies product that appears to alleviate irritable bowel syndrome symptoms (Kajander *et al.*, 2005; Kajander *et al.*, 2008) and appears to be immunologically active by inducing the expression of a diverse array of immune response genes in human macrophages (Miettinen *et al.*, 2012) and mast cells (Oksaharju *et al.*, 2011). Intriguingly, the dairy strain LC705 adheres poorly to human mucus (Tuomola *et al.*, 2000) and Caco-2 cells (Jacobsen *et al.*, 1999) and is markedly less effective in colonising humans than strain GG (Saxelin *et al.*, 2010), which originates from the stool specimen of a healthy human (Silva *et al.*, 1987). In Study III, the genome of *L. rhamnosus* LC705 was revealed to be slightly larger than that of strain GG. This strain was observed to contain a 2.97-Mb circular chromosome and a 64.5-kb circular plasmid that together included 2,992 CDSs, 61 tRNA genes, and 5 rRNA operons (Table 7). Approximately 77% of the CDSs in *L. rhamnosus* LC705 were found to begin with ATG and approximately 20% of them were preceded by an RBS (Figure 7), highlighting the similarity between the genomic compositions of *L. rhamnosus* LC705 and *L. rhamnosus* GG. Notably, LC705 was predicted to contain three prophage-like regions, indicating that transduction might have been an important mechanism for genome evolution in this species, as has been proposed for *L. casei* (Broadbent *et al.*, 2012).

The chicken isolate *L. crispatus* ST1 has been shown to colonise various areas of the chicken alimentary canal (Edelman *et al.*, 2002) and has in the literature been documented to strongly adhere to human vaginal epithelial cells, apparently through the high-molecular-mass *Lactobacillus* epithelium adhesin known as LEA (Edelman *et al.*, 2012). Other noteworthy traits associated with *L. crispatus* ST1 include its inhibition of the adhesion of avian pathogenic *E. coli* (Edelman *et al.*, 2003) and its ability to secrete proteins that enhance the cleavage of plasminogen into biologically active fragments (Hurmalainen *et al.*, 2007). In an effort to advance our understanding about the physiology of *L. crispatus* ST1 and to catalogue its adhesion factor potential, the genome of *L. crispatus* ST1 was determined and analysed in Study IV. The final genome assembly is delimited by a small gap of approximately 590 bp in the *lea* gene and was estimated to be approximately 2.04 Mb in size, representing the smallest *L. crispatus* genome reported to date. The genome was predicted to be devoid of plasmids and was found to have a low GC content (37%), which is similar to the GC contents of its closest relatives, *L. acidophilus* (Altermann *et al.*, 2005; 35%) and *L. helveticus* (Callanan *et al.*, 2008; 38%) but is different from those of *L. rhamnosus* GG and LC705 (47%). In terms of functional features, strain ST1 resembled the previously described *L. crispatus* strains and was predicted to contain 2,024 CDSs (Table 7), several of which might contribute to the maintenance of vaginal health (Study IV). An in-depth analysis of the CDS models revealed that ATG is the most common start codon in strain ST1 and that the use of alternative start codons is consistent with observations made in other *L. crispatus,* whereby TTG and GTG were associated with 10 and 9% of CDSs, respectively (Table 7). Only 6% of CDSs exhibited an RBS motif, similar in number to the observations for *L. crispatus* 125-2-CHN, JV-V01, and 214-1 but less than half of the number of observations for the remaining four *L. crispatus* strains. In Study V, a comparative genomics analysis of

L. crispatus ST1 and nine vaginal *L. crispatus* isolates was carried to investigate the scale and scope of the pan- and core genomic potential of and genomic diversity in *L. crispatus*.

Collectively, the results indicate that *L. rhamnosus* GG and LC705 contain a genome of approximately 3.0 Mb and that *L. crispatus* ST1 has a genome that is a bit over 2.0 Mb. Although these were not the first *L. rhamnosus* or *L. crispatus* strains to be determined at the genome level, they were the first for these two species to be assigned into a single scaffold, thereby extending the collection of high-quality genomes that are available for the *L. casei* clade beyond *L. casei* (Makarova *et al.*, 2006; Cai *et al.*, 2009) and the collection of high-quality genomes that are available for the *L. delbrueckii* clade beyond *L. johnsonii* (Pridmore *et al.*, 2004; Wegmann *et al.*, 2009), *L. acidophilus* (Altermann *et al.*, 2005), *L. delbrueckii* (van de Guchte *et al.*, 2006), *L. delbrueckii* (Makarova *et al.*, 2006), *L. helveticus* (Callanan *et al.*, 2008), and *L. gasseri* (Azcarate-Peril *et al.*, 2008); thus, this research opened new avenues for the comparative genomics of *Lactobacillus*.

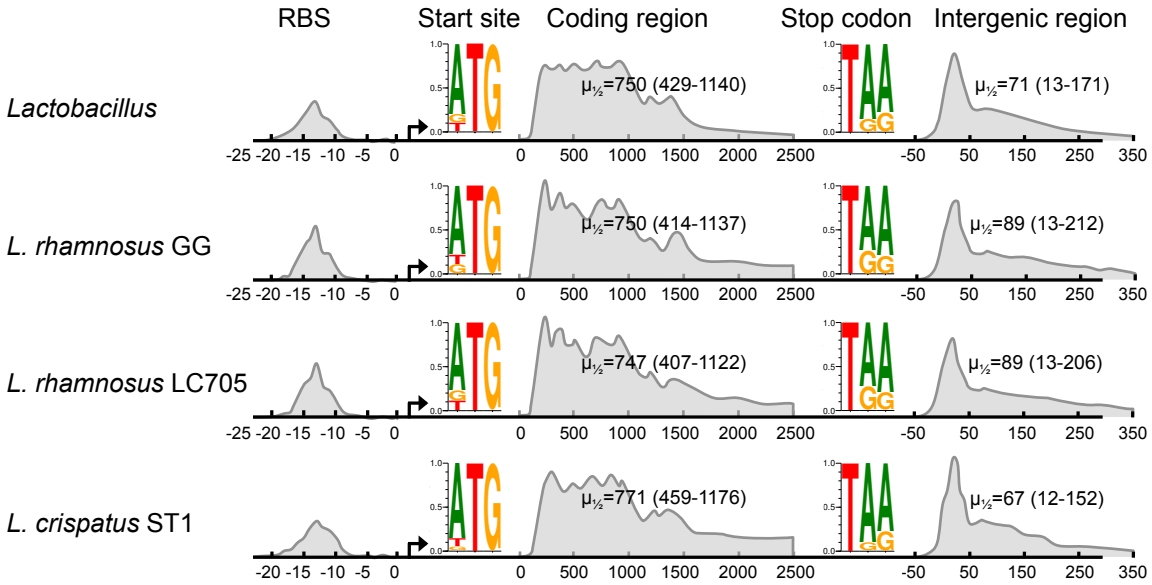


Figure 7. Comparison of consensus CDS models in the *Lactobacillus*, *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 genomes. From left to right: the spacing distribution between the 5′ A residue of the RBS and the translational start point, the sequence logo of the translation start point, the length distribution of CDSs, the sequence logo of the translational stop point, and the length distribution of intergenic regions. The median (and interquartile ranges) of the CDS and intergenic region lengths is indicated.

4.3 Gene calling in *L. rhamnosus and L. crispatus* sequences

Identifying CDSs in genomes is one of the first and most crucial steps in any genome sequence analysis (Angelova *et al.*, 2010) and should receive a great deal of care because annotation errors made at this step of the analysis can have detrimental consequences on

our understanding of the functional capacity of the organism. To ensure that as few annotation errors as possible had occurred at the initial automated gene calling step in Studies III and IV, an extensive set of gene callers was applied to the *L. rhamnosu*s GG and LC705 and *L. crispatus* ST1 genomes to evaluate their performance at locating genes within their genomes. Of the nine tested gene-calling systems, Glimmer (Delcher *et al.*, 2007) was among the top performers in terms of sensitivity and specificity (Table 8). However, the performance of Glimmer was comparable to those of GeneMark (Besemer *et al.*, 2001) and Prodigal (Hyatt *et al.*, 2010), and only marginal differences existed between these three methods, as has been observed in previous gene-calling performance tests (Hyatt *et al.*, 2010; Angelova *et al.*, 2010). The accuracy of the remaining methods was slightly lower; however, their gene models also matched fairly well with the manually reviewed CDS models and regions in the GG genome with transcriptional activity. An exception to the above was ERGO's gene-calling software (Overbeek *et al.*, 2003), which produced gene calls that did not match those obtained using the other methods and those observed in other lactobacilli. Specifically, the ERGO system preferred CDSs that started from codons other than ATG and produced CDS models that were aligned only partially with those found in sequence databases, suggesting that this service might not provide optimal results for lactobacilli. This result is consistent with the results of an earlier genome investigation, wherein the re-annotation of the genome of *Caulobacter crescentus* strain NA1000 resulted in a reduction in the use of rare codons and led to the improved annotation of 7% of the original 3,879 ERGO gene models (Ely & Scott, 2014). Another interesting finding was that low sensitivity values were evident in the gene expression test (Table 8). This is partially explained by the presence of polycistronic transcripts that can include intergenic regions (and are thus expressed) but also relates to the finding that various probes targeting an antisense region produced an expression signal that was higher than the chosen threshold.

Table 8. Comparison of CDS calls of nine gene callers. CDS calls were compared against manually refined annotation data and regions with and without transcriptional activity in *L. rhamnosus* GG. The sensitivity (Sn) is the fraction of gold standard bases that was captured by each prediction, and specificity (Sp) is the fraction of non-gold standard bases that was captured by each prediction.

	GG CDSs		LC705 CDSs		ST1 CDSs		GG microarray	
Method	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Glimmer	98.2 %	99.4 %	98.5 %	99.3 %	99.1 %	99.8 %	61.7 %	95.5 %
Prodigal	98.3 %	99.2 %	98.4 %	99.3 %	99.0 %	99.6 %	61.8 %	95.5 %
GeneMark	98.1 %	99.2 %	98.3 %	99.2 %	99.3 %	99.3 %	61.8 %	95.5 %
EasyGene	96.8 %	99.7 %	97.0 %	99.7 %	98.6 %	99.7 %	58.3 %	96.0 %
YACOP	96.2 %	99.6 %	96.3 %	99.5 %	97.4 %	99.6 %	58.4 %	96.2 %
RAST	98.3 %	98.4 %	96.7 %	97.5 %	98.8 %	98.6 %	64.1 %	95.3 %
ERGO	59.0 %	95.5 %	-	-	-	-	58.1 %	80.4 %
Critica	95.6 %	99.7 %	95.4 %	99.6 %	97.4 %	99.7 %	56.8 %	96.5 %
Zcurve	98.1 %	96.9 %	98.0 %	97.1 %	99.2 %	98.5 %	63.5 %	92.0 %

Overall, the three top-performing gene-calling systems were almost equal in quality. However, the rank of Glimmer as the performing method in five out of the eight performance tests provide a strong justification for the choice of Glimmer for initial gene calling. It is, however, possible that Glimmer might have missed some CDSs residing in genome regions with an abnormal base composition and failed to characterise the correct start site for some CDSs. However, the manual annotation phase, which constituted the manual editing of start sites and the screening of intergenic regions for missed CDSs, should have rectified these problems in the *L. rhamnosu*s GG and LC705 and *L. crispatus* ST1 genomes.

4.4 Functional annotation of *L. rhamnosus* and *L. crispatus*

Protein functional prediction consisted of running a battery of automatic, mostly homology-based function prediction tools, followed by manual curation of the results. The tools developed in Studies I and II were also used. A specific focus was placed on CDSs that code for EPS biosynthesis, extracellular proteins, antimicrobial peptides, proteinaceous adhesion factors, enzymes, prophage-like proteins, and CRISPR-Cas system components. Figure 8 outlines the main findings of these analyses.

4.4.1 General functional prediction of *L. rhamnosus* and *L. crispatus* genes

The application of bioinformatic methods constituted a central component of the functional classification of the predicted proteins. In general, the bioinformatic analysis provided important insights into the physiology of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 and enabled the assignment of initial functions to 72-77% of the predicted proteins; 10-13% of the predicted proteins represented conserved proteins with unknown functional roles, and 13-15% of the predicted proteins lacked homology information and remained classified as hypothetical. The success rate of the annotation processes was similar to those described in various earlier (Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004; Altermann *et al.*, 2005) and more recent (Forde *et al.*, 2011; Macklaim *et al.*, 2011) *Lactobacillus* genome studies, indicating the presence of a rather constant fraction of classifiable CDSs in any *Lactobacillus* genome regardless of publication date. Automatic protein function prediction was especially successful in defining genes that encode transcription factors, two-component regulatory systems, and enzymes, suggesting that the functional assignment of these gene types did not require much manual input. In contrast, the automated protein function prediction was less useful for proteinaceous adhesins, host-interaction factors, CRISPR-Cas systems, and transporters. This situation appears to be reminiscent of the results obtained in other *Lactobacillus* genome sequence studies, given the number of studies focusing on re-discovering these types of proteins in lactobacilli (Boekhorst *et al.*, 2006; Zhou *et al.*, 2008; Horvath *et al.*, 2009; Kleerebezem *et al*., 2010; Kant *et al.*, 2010;). Moreover, the concordance between the results generated by different protein function prediction

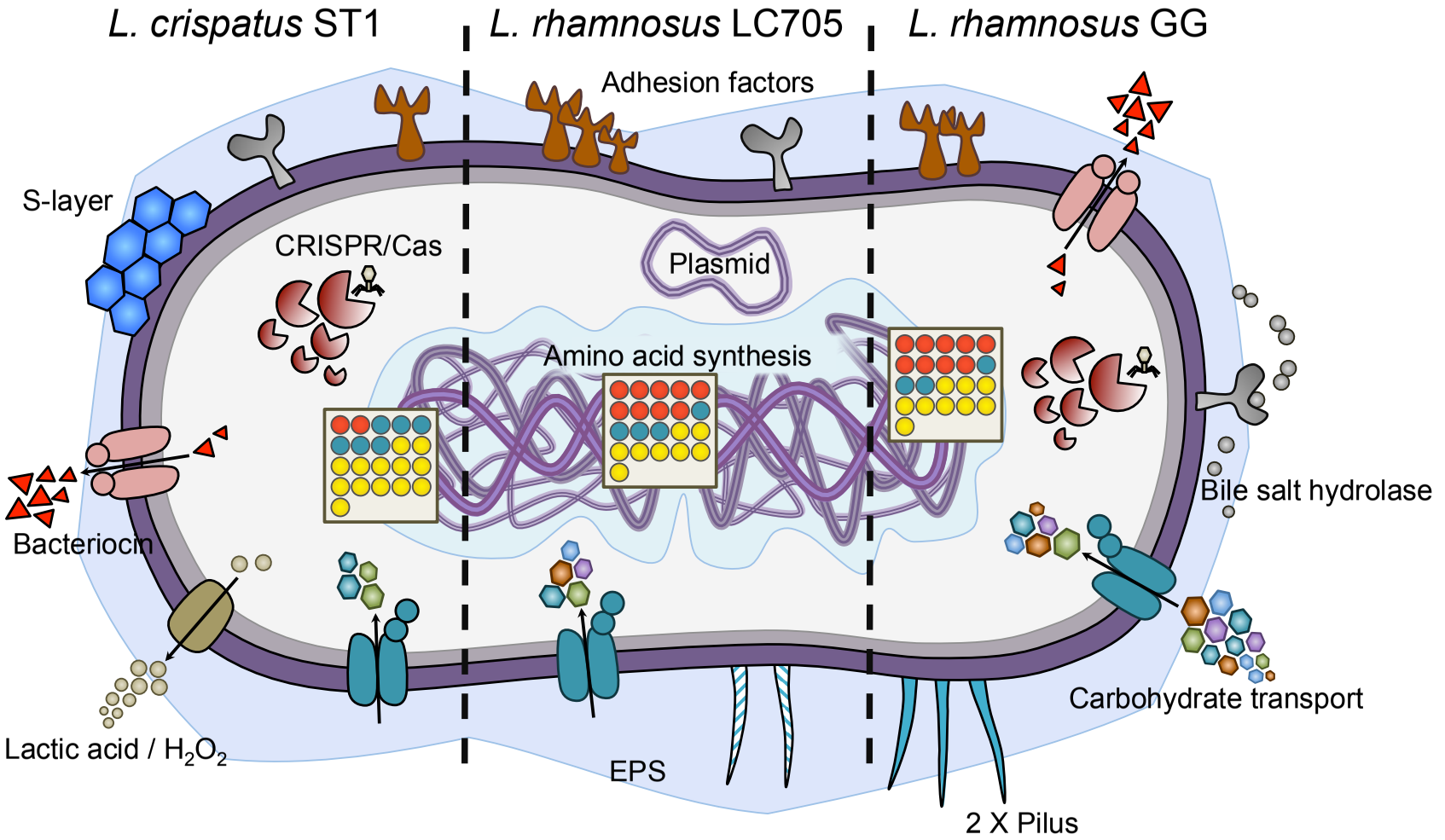


Figure 8. The main findings of Studies III-V. Squares in the middle show the number of amino acids that *L. crispatus* ST1 (left), *L. rhamnosus* LC705 (middle), or *L. rhamnosus* GG (right) produce *de novo* (red circles), obtain through inter-conversion (blue circles), or extract from the medium (yellow circles). The notations used are the same as those used in Figure 5.

procedures was low. In Study II, BLANNOTATOR generated a correct prediction (an annotation that was accepted by the human operator) for approximately 85% of the *L. crispatus* gene products for which it assigned a function. In contrast, the two other homology-based protein function prediction tools that were applied to this dataset generated supported annotations for approximately 58% (RAST) and approximately 69% (the best BLAST approach) of the gene products with approved functional annotation, indicating the importance of choosing the correct annotation method. However, it should be noted that different methods failed with regard to different proteins and that pooling the results was necessary in several instances. For example, the best-BLAST approach was poor at producing component composition annotations for phosphotransferase system (PTS) transporter genes, unlike the InterProScan; RAST failed at calling CRISRP-cas components, unlike BLANNOTATOR; and GO term predictions were overly general when calling adhesins but were highly useful in classifying proteases.

4.4.2 Host-interaction molecules in *L. rhamnosus* and *L. crispatus* strains

The computational prediction of localisation sites of *L. rhamnosus* GG and LC705 and *L. crispatus* ST1 proteins revealed that close to 10% of these predicted proteins are secreted and/or membrane-associated and are therefore potentially involved in processes such as nutrient acquisition, cell communication, microbe-host interaction, and adhesion (Lebeer *et al.*, 2008; Kleerebezem *et al.*, 2010; Segers & Lebeer, 2014). Although the values are in good agreement with those seen for closely related lactobacilli (Kleerebezem *et al.*, 2010), these predictions provide only an initial guess of the secretomes. Particularly, a relatively limited number of concordances appeared between the predictions and experimentally determined cell envelope (Koskenniemi *et al.*, 2011) or secreted (Sanchez *et al.*, 2009) proteins in *L. rhamnosus* GG. There also is a noted disagreement between the results from different studies. For example, two subsequent studies have defined that the secretomes of strains GG and LC705 could contain over 80 (Zhou *et al.*, 2008) and 700 (Kant *et al.*, 2010) more proteins than observed in Study III. In the first case, most disagreements were related to proteins that are N-terminally (Study III defining 149-159 fewer predictions) or C-terminally (Study III describing 18-19 more predictions) anchored, despite of the obvious similarities in the underlying computational procedures. Which of the approaches is preferable remained unclear; however, the high rate of false-positives (84%) that were reported according to the N-terminally anchored sorting predictions of LocateP (Berlec *et al.*, 2011) argues against the other study. Despite the uncertainty of the sorting predictions, subcellular-location predictors were used in Studies III and IV because homology-based functional inference failed to provide clues about protein sorting. GO annotation data from the cellular component ontology were assigned to approximately only 30% of the CDSs, whereas the ambiguous naming of known extracellular proteins excluded the use of gene names.

Bacterial surface polysaccharides are ubiquitous components of the cell envelope of lactobacilli and are purportedly involved in determining host-microbe interactions (Lebeer *et al.*, 2008). In particular, a long galactose-rich EPS molecule from *L. rhamnosus* GG is

important for required for optimal survival of strain GG inside the murine GIT (Lebeer *et al.*, 2011) and an EPS with high rhamnose content from another *L. rhamnosus* strain has been shown to stimulate various cytokines in human cell-line experiments (Chabot *et al.*, 2001; Péant *et al.*, 2005). The EPS gene cluster of LC705 that was discovered in Study III exhibited high similarity to EPS loci present in four other *L. rhamnosus* strains and shown to produce an EPS with high rhamnose content (Péant *et al.*, 2005). In contrast, the EPS locus in *L. rhamnosus* GG was genetically different from that of the *L. rhamnosus* LC705 and verified the presence of a previously identified EPS gene cluster in strain GG (Lebeer *et al.*, 2008), providing groundings for the comparison of the genomic neighbourhoods of these two EPS gene clusters. Regarding *L. crispatus*, some strains have been observed to produce EPS (Donnarumma *et al.*, 2014). However, the genetic composition of EPS loci in *L. crispatus* has remained uncharacterised, despite of the vast amount of literature on EPS clusters in closely related *Lactobacillus* genomes (Pridmore *et al.*, 2004; Altermann *et al.*, 2005; Callanan *et al.*, 2008; van de Guchte *et al.*, 2006; Azcarate-Peril *et al.*, 2008). In Study V, eight *L. crispatus* strains were identified to include a gene cluster associated with EPS biosynthesis. Each of these eight clusters was predicted to comprise a set of five highly conserved genes encoding a transcriptional regulator, a polymerisation and chain length determination protein, a tyrosine protein kinase, a protein-tyrosine phosphatase, and a priming glycosyltransferase. In contrast, differences in the glycosyltransferase genes situated at the 5' end of the *L. crispatus* EPS clusters suggested that the EPSs might contain different sugar monomers and glycosidic linkages.

Adhesion to host tissues has long been considered a central factor and a prerequisite for the long-term colonisation by and realisation of the health benefits of probiotic bacteria (Lee & Salminen 1995; Pfeiler & Klaenhammer, 2007; Siezen & Wilson, 2010; Segers & Lebeer, 2014; Lebeer *et al.,* 2008;). However, prior to Studies III-IV, few adhesins been identified in *L. rhamnosus* (Chan *et al.*, 1985) or in *L. crispatus* (Antikainen *et al.*, 2002; Hurmalainen *et al.*, 2007; Edelman *et al.*, 2012), suggesting that new insights into *L. rhamnosus* and *L. crispatus* adhesion factors were enabled by the analysis of their genome sequences. In Studies III and V, proteins were classified into adhesion- or colonisation-related protein domain families. The domains were collected from PFAM, and their potential adhesion or colonisation associations were determined by manual examination of the corresponding literature. In Study III, this approach revealed over 30 putative proteinaceous adhesion and colonisation factors in both *L. rhamnosus* GG and LC705, including many that have later been verified experimentally (Vélez *et al.*, 2010; Lebeer *et al.*, 2012; von Ossowski *et al.*, 2011) and some others with physiological roles that have yet to be verified, such as the 382.1 KDa protein in *L. rhamnosus* LC705 that contains several collagen-binding domains. The search also revealed SpaCBA pilin subunits in *L. rhamnosus* GG that has been proven to be critical for its efficient adherence to human cells (Study III; von Ossowski 2010; Lebeer *et al.*, 2012). Repeating the analysis for *L. crispatus* using a revised domain list revealed nine to 13 adhesion- and colonisation-related proteins in each *L. crispatus* strain, many of which are part of the *L. crispatus* core genome. Notably, this search failed to characterise the LEA-protein, which is critical for the adhesion of strain ST1 to vaginal epithelial cells (Edelman *et al.*, 2012) and that was in Study IV shown to displace *Gardnerella vaginalis* from vaginal cells, indicating that

LEA-mediated adhesion might involve some yet undisclosed bacterial adhesion domain. Intriguingly, application of the revised domain set to *L. rhamnosus* GG and LC705 returned only eight and ten proteins, respectively. A detailed investigation of the functional annotations of the two sequence sets revealed differences in carbohydrate-active enzymes, indicating that the first domain set might have included erroneous, non-adhesion-related PFAM models. In addition to the PFAM search, protein functional information was used in the search for adhesins. However, the scarcity of adhesion-related GO annotations and the inconsistent naming of known bacterial adhesins in public databases precluded the use of these search strategies at full power.

Although pilins were successfully reported by searching for adhesion- or colonisation-related protein domain families, the approach used in Study III was found to be laborious for the genome-scale mining of hundreds of genomes. Thus, a new tool was developed in Study I for the systematic screening of pilus operons in bacterial genomes and then used to investigate the distribution of pilus clusters in all complete gram-positive prokaryotic genomes that had been deposited in NCBI database. Interestingly, putative pilus operons were found in 67 out of the 181 genomes analysed, including four lactobacilli genomes. Further analysis with all available *Lactobacillus* genomes that are listed in Appendix Table 2 demonstrated the presence of putative pilus genes clusters in 29 strains. Putative pilus gene clusters were identified in selected *L. gasseri*, *L. reuteri*, and *L. ruminis* strains. In addition, pilus operons were found to be an almost universal feature of the *L. casei* group, with the exception of *L. rhamnosus* MTCC 5462. The wide distribution of pilus gene clusters within the *L. casei* group has recently been described in other studies and for a greater number of *L. rhamnosus* genomes, revealing (i) the occurrence of SpaCBA pilin subunits in 4 out of 13 (Kant *et al.*, 2014) and in 34 out of 100 (Douillard *et al.*, 2013b) *L. rhamnosus* strains and (ii) the presence of SpaFED pilin subunits in all strains that were investigated (Douillard *et al.*, 2013b; Kant *et al.*, 2014). Of note, LOCP returned some false predictions based on the gene annotations. The fraction of false positives in the analysis was however tolerable and substantially less than the number of sequences in these protein collections that contained an LPXTG-motif or an E-box (typically used to search pilins).

4.4.3 Bacteriocins in *L. rhamnosus* and *L. crispatus* strains

Lactobacilli produce an extensive set of antimicrobial substances including metabolic by-products (Daeschel, 1989; Leroy & Vuyst, 2004; Nes & Johnsborg, 2004) such as lactic acid, acetic acid, ethanol, diacetyl, and hydrogen peroxide, as well as bacteriocins (peptides or small proteins that exhibit antimicrobial activity; Riley & Wertz, 2002; Jack *et al.*, 2005). Notably, bacteriocin-producing lactobacilli are of great importance for use in food preservation (Leroy & Vuyst, 2004; Nes & Johnsborg, 2004; Cotter *et al.*, 2005; Mills *et al.*, 2011) because they can protect food against contamination with specific pathogenic and spoilage organisms, such as *L. monocytogenes* (Corr *et al.*, 2007) and *Clostridium tyrobutyricum* (Jiménez-Díaz *et al.*, 1993) without affecting harmless LAB. In Study III, an 8.7-kb putative type IIb bacteriocin locus was identified in the GG genome.

In addition to two short bacteriocin structural genes, the locus appeared to encode five genes implicated in immunity, bacteriocin production, and bacteriocin processing, indicating that strain GG (like several other *L. rhamnosus* strains, Jacobsen *et al.*, 1999) produces a bacteriocin. However, contradictory evidence exists regarding the role of bacteriocin in the antimicrobial activity of strain GG and whether the inhibition of *Salmonella typhimurium* by strain GG is due to bacteriocin activity (Silva *et al.* 1987; Jacobsen *et al.*, 1999) or to lactic acid accumulation (De Keersmaecker *et al.*, 2006). A similar locus was identified in the LC705 genome. However, this region appeared to be non-functional because two of its genes were truncated and were probably pseudogenes. The *in silico* analysis of the *L. crispatus* genomes described in Study V revealed loci in each strain, which encoded bacteriolysins that are similar to the previously described enterolysin A (Nilsen *et al.* 2003) and helveticin J (Joerger & Klaenhammer, 1990), which lyse sensitive cells by catalysing cell wall hydrolysis. In addition, vaginal *L. crispatus* isolates contained genes that are implicated in the production of class II bacteriocins, indicating that these strains might produce an active bacteriocin, as previously reported for *L. crispatus* ATCC 33820 (Kim & Rajagopal, 2001) and *L. crispatus* JCM 2009 (Tahara & Kanatani, 1997).

Although bioinformatics services exist for mining bacteriocin loci from genomes (van Heel *et al.*, 2013), the annotation of bacterial genomes for bacteriocins is not an easy task due to the small size and low degree of conservation of bacteriocins and the fact that they are often omitted from genome annotations and/or lack descriptive functional descriptions. Furthermore, genome regions surrounding bacteriocin genes and genes that are implicated in both the production and processing of bacteriocins are often enriched in pseudogenes. During manual annotation, these problems can be tackled by searching the surroundings of bacteriocin-related genes for small ORFs and by comparing these ORFs to HMMs that correspond to bacteriocin-related sequences, as was performed in Studies III and IV. Alternatively, bacteriocin calling can be based on bacteriocin prediction systems, such as BAGEL (van Heel *et al.*, 2013), as was done in Study V. In general, these two approaches should be very similar and result in comparable protein sets; however, differences are possible. For example, BAGEL revealed four class III bacteriocin loci for ST1, two of which were also identified in a manual search. However, analysis of the GG and LC705 genomes using BAGEL revealed seven bacteriocins, including only half of the bacteriocin type II leader motif proteins that were described in Study III; these findings indicate the difficulty of the prediction task involved, at least for these genomes.

4.4.4 Prophage elements and CRISPR loci

Based on the *in silico* identification of prophage-like regions (Lima-Mendez *et al.*, 2008), the genome of *L. rhamnosus* GG had two large and one short putative prophage element, whereas the genome of LC705 had one large and two short putative prophage elements. The prophage-like regions in the LC705 genome shared a low similarity to those in GG and resided at different locations, suggesting an important role for bacteriophage in *L. rhamnosus* evolution, as has previously been proposed for *L. casei* (Cai *et al.*, 2009) and is

reminiscent of the high degree of diversity seen among prophages that have been identified in the genomes of *L. gasseri*, *L. salivarius*, and *L. casei* previously (Ventura *et al.*, 2006). Consistent with the high level of lysogeny (77%) in vaginal *L. crispatus* strains (Damelin *et al.*, 2011), each vaginal *L. crispatus* isolate was predicted to have at least one putative prophage element in Study V. Conversely, the genome of *L. crispatus* ST1 revealed no putative prophage elements, possibly due to the fact that it was predicted to have a different type of CRISPR locus than the vaginal isolates. Overall, computational tools were successful in identifying prophage-like clusters, and the methods used appeared to capture all relevant prophage elements and resulted in only one false prediction: a prophage like-region in *L. crispatus* ST1 that harboured housekeeping genes. Moreover, the phage-finder services helped in the annotation of phage-related genes.

CRISPR-Cas systems constitute a widespread class of RNA-based immunity systems that control invasions of bacteriophages and plasmids in prokaryotes (Deveau *et al.*, 2010; Marraffini & Sontheimer, 2010). These systems are present in approximately two-thirds of *Lactobacillus* strains (Horvath *et al.*, 2009) and provide an exceptional tool for the control of phage infections (Marraffini & Sontheimer, 2010); a significant and prevalent threat that disrupts dairy fermentation cycles, thus stalling the manufacturing chain and lowering the quality of the end product (Mc Grath *et al.*, 2007). In Study III, a genomic screen for CRISPR repeats identified the presence of one CRISPR array in the genome of strain GG. This region comprised of 24 perfect repeats, and the spacers showed substantial sequence identity with various *L. rhamnosus*-specific phages, indicating that they might be its phage targets. Further, the CRISPR array located next to four Type I *cas* genes, corroborating the possibility of a functional system. Intriguingly, the Cas-proteins showed notable similarity to those described in *L. salivarius* UCC188 and *L. casei* BL23 but not to those described in *L. casei* ATCC 344 (Horvath *et al.*, 2009), indicating that these Cas-proteins have not followed an evolutionary development similar to that of their bacterial hosts. No CRISPR-Cas systems were detected in strain LC705. In Study V, full or partial CRISPR-Cas systems were identified in each of the studied *L. crispatus* strains. All vaginal *L. crispatus* isolates contained at least one Type II *cas* gene and a CRISPR array comprising 36-bp direct repeats and at least two to six spacer sequences each. Homology searches among the spacers, and public virus and plasmid sequences did not reveal the putative targets of the crRNAs, suggesting a pool of undisclosed vaginal bacteriophages and plasmids. However, many of the spacers were shared by different vaginal strains, indicating that these isolates might have encountered common invaders in the past. Unlike the vaginal *L. crispatus* isolates, *L. crispatus* ST1 was predicted to carry eight Type I *cas* genes and two CRISPR arrays comprising 15 and 16 repeats. The repeats were highly similar and resembled a repeat that was recently described as present in vaginal metagenome samples (Rho *et al.*, 2012). In contrast, the spacers of these systems did not match known plasmid or virus sequences. Similar to the prophage search tools described above, the CRISPR array scanners were accurate and reliable. The tools uncovered all CRISPR-appearing genomic regions, accurately predicted CRISPR array boundaries, and resulted in only one false prediction; namely, a CRISPR array situated within the *LCRIS_01228* gene in *L. crispatus* ST1. Thus, genome sequencing and computational biology provide a powerful means for annotating prophage elements and CRISPR arrays in *Lactobacillus*.

4.5 Genomics of *L. rhamnosus* and *L. crispatus* metabolism

Metabolic pathway reconstructions were used to investigate carbohydrate metabolism and amino acid biosynthesis pathways in *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1. Overall, substantial percentages of genes were identified as involved in transport (13-15%) or enzymatic reactions (22-25%; Studies III and IV), percentages that are consistent with those described for other *L. rhamnosus* and *L. crispatus* genomes (Table 7). The assignment of these genes into reference pathways provided some of the first insights into the biosynthetic capabilities of these organisms and indicated that strains GG and LC705 can synthesise nine amino acids *de novo* and synthesise three through inter-conversion. Additionally, genes for the conversion of serine to cysteine were annotated for plasmid pLC1, extending the biosynthetic potential of LC705 beyond that of *L. rhamnosus* GG, which appeared to lack the genes that are implicated in this conversion. These biosynthetic capabilities are reminiscent of those of *L. casei* ATCC 344, which was annotated to have nine amino acid biosynthesis routes (Makarova *et al.*, 2006), and are almost comparable to the collection of metabolic pathways that was described for the versatile *L. plantarum*, which was predicted to contain complete pathways for the biosynthesis of most amino acids (Kleerebezem *et al.*, 2003). Through *de novo* synthesis and amino acid inter-conversions, *L. crispatus* ST1 was predicted to be able to produce eight amino acids (Study IV). Bioinformatic analysis of vaginal *L. crispatus* genomes suggested that these and the strain ST1 share the same biosynthetic potential, except for CTV-05, which was predicted to be auxotrophic for aspartate (Study V). These biosynthetic capabilities are similar to those found in *L. johnsonii* NCC 533, *L. acidophilus* NCFM, and *L. helveticus* DPC 4571, which have been reported to have the ability to produce 4 (Pridmore *et al.*, 2004), 10 (Altermann *et al.*, 2005), and 4 (Callanan *et al.*, 2008) amino acids (either *de novo* or as derivatives), respectively.

To compensate for their limited amino acid biosynthetic capabilities, many strains of *Lactobacillus* have developed sophisticated proteolytic and transport systems to obtain amino acids from their habitats (Pridmore *et al.*, 2004; Altermann *et al.*, 2005). Indeed, based on *in silico* analyses, *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 contain arrays of extracellular proteinases, cytoplasmic peptidases, and peptide and amino acid transporters. This result suggests that these bacteria exhibit enhanced abilities to utilise exogenous amino acids and peptides. Notably, *L. rhamnosus* GG and LC705 were predicted to have nearly identical casein degradation systems, although only LC705 has the ability to degrade milk protein (Study III). Regarding pyrimidine and purine biosynthesis, *L. crispatus* ST1, *L. rhamnosus* GG, and *L. rhamnosus* LC705 resembled their close relatives. Similar to *L. johnsonii* NCC 533 (Makarova *et al.*, 2006), *L. crispatus* ST1 needs to obtain pyrimidines from its environment, whereas *L. rhamnosus* GG and LC705 were predicted to synthesise both types of nucleotide bases, as has previously been described for *L. casei* ATCC 334 (Makarova *et al.*, 2006).

Genomic analysis revealed that both *L. rhamnosus* GG and LC705 include a series of genes that code for sugar metabolism and that these strains exhibit rather similar carbohydrate metabolism. Two noteworthy exceptions were the maltose and rhamnose pathways, which were found to be intact and functional only in LC705. In addition,

analyses presented in Study III revealed GG-specific frameshifts in two genes that act in lactose utilisation, providing a plausible explanation for the results obtained from a sugar utilisation assay, which indicated that GG cannot utilise lactose, unlike its dairy counterpart, LC705 (Study III). Moreover, similar sets of glycosidase genes were predicted to be present in the genomes of *L. rhamnosus* GG and LC705. Based on functional annotations and protein-sorting analyses, approximately ten of these genes contribute to peptidoglycan hydrolysis and the decomposition of complex polysaccharides.

The *in silico* reconstruction of *L. crispatus* sugar utilisation pathways suggested that *L. crispatus* can use a range of carbohydrates (Study V), as has previously been reported for several other members of the *L. delbrueckii* group (Pridmore *et al.*, 2004; Altermann *et al.*, 2005; Azcarate-Peril *et al.*, 2008). As with amino acid biosynthesis, CTV-05 differed the most and was predicted to lack various sugar utilisation pathways that were present in the other strains, most likely because of the sequencing gaps that are present in the corresponding genomic loci. Interestingly, *L. crispatus* pathway data argue against the classical grouping of *L. crispatus* as a homofermentative species (Salvetti *et al.*, 2012). Instead, pathways for both homofermentation (the Embden-Meyerhof-Parnas pathway) and heterofermentation (the pentose phosphoketolase pathway) were observed in all *L. crispatus* strains investigated in Study V, as is typical of facultatively heterofermentative species (Pot *et al.*, 1994; Hammes & Vogel, 1995; Felis & Dellaglio, 2007; Salvetti *et al.*, 2012; Salvetti *et al.*, 2013).

Practically, the reconstruction of metabolic pathways involved the use of a battery of bioinformatics tools and resources (Table 6). In Study III, KEGG reference pathways were noted to be useful for obtaining an overall view of metabolism but were often insufficient for explaining specific metabolic capabilities due to their complexity. In contrast, the reaction maps in the MetaCyc database contain on average only 4.4 reactions (Altman *et al.*, 2013) and allowed the efficient examination of whether genes for a particular bioconversion were present in the organism (Studies III and IV). Overall, the data in these reaction pathway collections agreed well, which is consistent with the reported high degree of overlap (63%) between the KEGG and MetaCyc reaction spaces (Altman *et al.*, 2013). However, inconsistencies were also observed. According to the KEGG, reaction 4.4.1.8 transforms pyruvate into cysteine. In contrast, MetaCyc found that this transformation involved another reaction and an enzyme that was not found in *L. crispatus* ST1, indicating a lack of cystathionine beta-lyase reaction in ST1. Other discrepancies involved reactions 1.1.1.351 (which catalyses the reduction of 6-phosphogluconate to ribulose 5-phosphate) and 1.2.1.59 (which catalyses the sixth step of glycolysis) that were associated with the phosphoketolase and Embden-Meyerhof-Parnas pathways only in MetaCyc and KEGG maps, respectively. In Study V, the presence of metabolic routes was tested by matching each strain's EC complement against the EC sets that are annotated to enable the conversion of a given starting compound to a particular end product. Metabolic routes between two given compounds were retrieved using the FMM web-server (Chou *et al.*, 2009), which connects different KEGG maps and reconstructs metabolic pathways between metabolites. This approach greatly reduced the amount of work involved in understanding metabolic activities, although the test was found to yield some false

positive calls due to an unrealistic linking between some reactions. The FMM service also ignored some metabolic conversions that were recognised by the KEGG web service, possibly causing some pathways to remain undetected. Overall, use of the KEGG and MetaCyc databases and the reference-pathway approach were crucial for understanding the metabolic capabilities of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1. The FMM server aided the process but at the cost of yielding some errors. However, the level of errors is acceptable considering the workload involved with reconstructing pathways manually, especially if there is a need to investigate tens or hundreds of genomes.

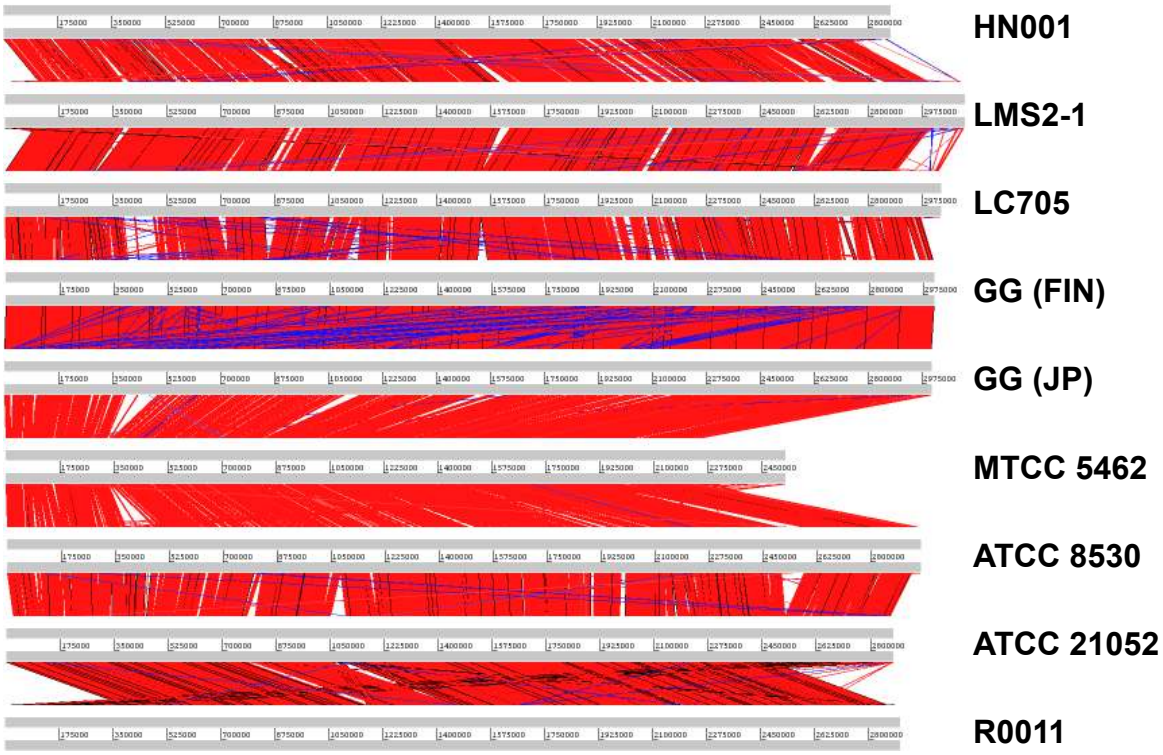


Figure 9. Alignment of nine *L. rhamnosus* genomes using ACT. The vertical bands between the genomes represent BLASTN matches (bit score ≥ 500) between the two sequences. Forward and reverse matches are indicated by red and blue, respectively. The draft genome sequences were ordered and oriented according to the genome sequence of *L. rhamnosus* GG (FIN) using progressive Mauve.

4.6 Comparative genomics of *L. rhamnosus* and *L. crispatus*

Whole-genome alignment is a powerful tool for understanding the genetic forces that have shaped genomes (Ali *et al.*, 2013; Darling *et al.*, 2004). It has promoted the discovery of key lactobacilli effector molecules (Morita *et al.*, 2008) and has vastly expanded our

knowledge of the diversity and complexity that exists among *Lactobacillus* species (Forde *et al.*, 2011; Boekhorst *et al.*, 2004; Canchaya *et al.*, 2006; Berger *et al.*, 2007; Ventura *et al.*, 2008; van de Guchte *et al.*, 2006; Broadbent *et al.*, 2012). In Study III, a whole-genome comparison revealed a high degree of stability among the genomes of *L. casei* group bacteria, a finding that has recently been confirmed in several other studies (Cai *et al.*, 2009; Broadbent *et al.*, 2012; Douillard *et al.*, 2013a). The high level of genome relatedness and synteny is further underscored by a genome comparison of nine *L. rhamnosus* isolates, which revealed notable whole-genome synteny and no evidence of chromosomal rearrangements between these nine *L. rhamnosus* strains (Figure 9). Specifically, the nine *L. rhamnosus* isolates shared approximately 88% of their DNA. The colinearity was however noted to be punctuated by 1-182 genomic islands, which were generally consistent with genomic island predictions and exhibited abnormal sequence compositions, further attesting to their foreign origin and suggesting a key role for lateral gene transfer during *L. rhamnosus* evolution. In particular, the genomes of *L. rhamnosus* GG and *L. rhamnosus* LC705 shared extensive synteny, which was found to be punctuated by four (strain LC705) to five (strain GG) genome regions that displayed nucleotide composition deviations relative to the remainder of the genome. An examination of these genome regions in Study III revealed many genes of biomedical importance and implicated in EPS biosynthesis, host colonisation (SpaCBA pilins), prophage, and metabolism-related functions. It was further concluded that the GG-specific EPS and prophage and LC705-specific sugar utilisation islands were most likely acquired by horizontal gene transfer at a late point of divergence. In contrast, the phylogenetic distribution of the SpaCBA cluster in the lineage was best explained by a horizontal gene transfer by the common ancestor of *L. casei* and *L. rhamnosus*, as was apparent in more recent and broader comparative genome analyses that confirmed the presence of *SpaCBA* genes in the genomes of many *L. casei* strains (Broadbent *et al.*, 2012) but only in a few *L. rhamnosus* strains (Douillard *et al.*, 2013a; Douillard *et al.*, 2013b; Kant *et al.*, 2014).

As expected, genome alignments between *L. crispatus* strains revealed a high level of similarity and synteny (Study V). The genomes of 214-1 and SJ-3C-US were the most conserved; approximately 97% of their sequences were conserved in at least one other strain. The least related strain was the *L. crispatus* ST1; only approximately 82% of the genome of this isolate being alignable with the genomes of the nine vaginal isolates, underscoring its non-human origin. The data also indicated that the genome size differences observed in *L. crispatus* are not due to chromosomal insertions, inversions, deletions, or re-arrangements. Instead, horizontally acquired genomic islands and putative prophage elements explain a notable portion of the genomic differences in *L. crispatus* (Study V). Interestingly, the conservation of gene order in *L. crispatus* genomes has survived over a long evolutionary timescale and is also present in other lactobacilli of the *L. delbrueckii* group. Specifically, comparisons of the genomes of selected lactobacilli from the *L. delbrueckii* group revealed extensive sequence similarity and genome synteny between *L. crispatus* ST1 and other strains from this group (Figure 10) and indicated that the same overall gene order known to exist between most other members of the subgroup (Canchaya *et al.*, 2005; Berger *et al.*, 2007; Callanan *et al.*, 2008) is also valid for strain ST1. The most noteworthy exception was observed by comparing the *L. crispatus* ST1 and

L. gasseri ATCC 33323 genomes; this comparison revealed sequence scrambling around the replication terminus, which is best explained using the fork replication theory (Tillier & Collins, 2000; Canchaya *et al.*, 2005).

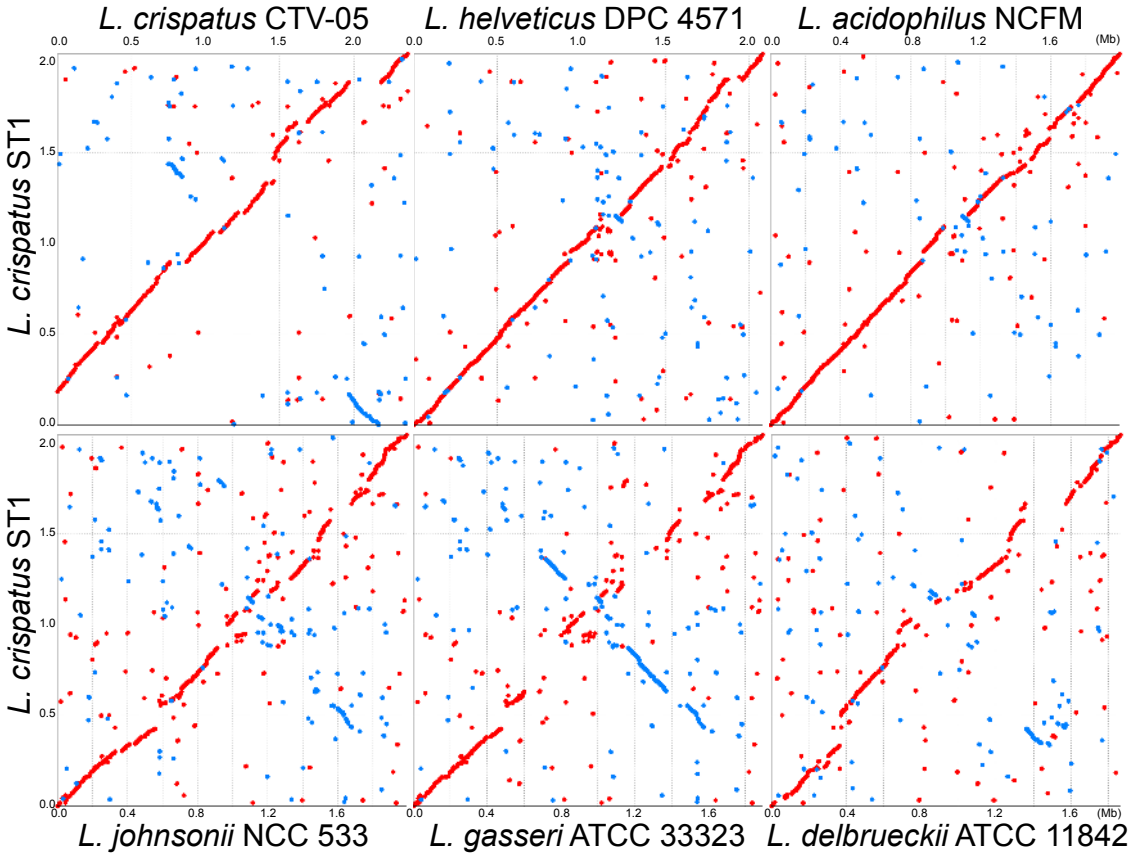


Figure 10. Comparison of the genome of *L. crispatus* ST1 with the genomes of selected lactobacilli of the *L. delbrueckii* group. PROmer alignments of the genome of *L. crispatus* ST1 with the genomes of *L. crispatus* CTV-05, *L. helveticus* DPC 4571, *L. acidophilus* NCFM, *L. johnsonii* NCC 533, *L. gasseri* ATCC 33323, and *L. delbrueckii* ATCC 11842 are shown. Red dots indicate conserved DNA sequences in the same orientation. Blue dots indicate conserved DNA sequences in the reverse orientation.

In addition to pairwise alignment approaches, sophisticated genome alignment tools were tested to address the problem of whole-genome alignment. The *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. casei* ATCC 334 genomes were aligned using Mauve (Darling *et al.*, 2004; Rissman *et al.*, 2009), TBA (Blanchette *et al.*, 2004), and MUMmer (Kurtz *et al.*, 2004). Of these sophisticated genome alignment tools tested, the Mauve alignment tool (Darling *et al.*, 2004; Darling *et al.*, 2010) was considered the most suitable alignment package for studying these lactobacilli genomes. However, even Mauve did not provide an additional level of detail over simpler alignment approaches. Moreover, because the genomes in question were small (Table 7) and alignments were largely collinear, manual

interpretation of the results was not laborious. Thus, it appeared that simple alignment techniques were sufficient for collinear and closely related bacterial genomes, such as those that were investigated in Studies III-V. Among the pairwise alignment approaches examined, ACT (Carver *et al.*, 2005) was most practical due to its ability to visualise several genomes simultaneously (see Figure 9). An obvious defect of this approach was that the results depend on the piling order and that an incorrect piling order can hide interesting genome differences. In contrast, summarisation of the alignment information over several dotplots was found laborious (see Figure 10), even though figures generated using dotplot visualisation utilities, such as Gepard (Krumsiek *et al.*, 2007) and MUMmer (Kurtz *et al.*, 2004), provided more information. Although not considered comparative genomics approaches, methods for sequence composition based genomic island prediction were noted to be useful for whole-genome alignment validation and for explaining the emergence of alignment-free genomic regions. However, sequence composition based genome island predictors were found to produce inconsistent results. For example, the IslandViewer resource (Langille & Brinkman, 2009) identified 102 genomic islands that constituted approximately 7% of the *L. crispatus* pan-genome. Shockingly, no genomic islands were recovered by all three methods included in the IslandViewer resource, and only six were supported by two methods (Study V).

4.7 Orthologue grouping of *L. rhamnosus* and *L. crispatus* genes

Three orthologue grouping tools were evaluated in Studies III-V. In Study III, orthologue groups were identified using a search strategy similar to that used in the InParanoid tool (Remm *et al.*, 2001). Although accurate, conversion of the multiple pairwise orthologue groups into multi-species clusters was found to be a laborious process; thus, this approach was abandoned. Instead, orthologue groups were identified in Studies IV and V using the OrthoMCL tool (Li *et al.*, 2003), which was noted to exhibit a good balance of sensitivity and specificity. In addition, a maximal clique approach based on RBHs was tested but not further used because long run time requirements (Study IV). Similar to previous ortholog assessment studies (Salichos & Rokas, 2011; Trachana *et al.*, 2011), the three orthologue grouping tools were noted to produce comparable results. For example, *L. rhamnosus* GG and LC705 were predicted to have 364 and 430, respectively, CDSs lacking orthologous counterparts in the other *L. rhamnosus* strain or in *L. casei* ATCC 344 according to the InParanoid search strategy (Study III). In contrast, OrthoMCL identified 273 and 292 such CDSs, respectively. More recently, an RBH comparison was conducted on the protein collections of 13 *L. rhamnosus* strains (Kant *et al.*, 2014). However, this analysis revealed only 94 and 26 unique elements for GG and LC705, partly because of methodological differences and partly because of the inclusion of new *L. rhamnosus* genomes, such as that of another strain GG isolate (Morita *et al.*, 2013).

In Study V, the extent of the core and pan genomic potential of ten *L. crispatus* isolates was calculated. It was predicted that pan-genome of these ten strains comprised 3,929 orthologue groups, 1,224 of which were present in each strain (Study V). This set of core groups captured approximately 31% of the given collection of orthologous groups and was

comparable to those observed previously for *L. casei* (approximately 29%; Broadbent *et al.*, 2012), *L. paracasei* (approximately 43%; Smokvina *et al.*, 2013), and *L. rhamnosus* (approximately 43%; Kant *et al.*, 2014). Based on the regression analysis, the core was also considered a good estimate of the final orthologue group repertoire of an unlimited number of *L. crispatus* strains. Regarding strain-specific functions, on average each *L. crispatus* strain was predicted to contain 131 orphan orthologous groups. However, the number of orphan groups ranged widely from 51 for strain MV-1A-US to 287 for strain FB077-07. Surprisingly, even the more phylogenetically distant *L. casei* and *L. rhamnosus* species have been reported contained approximately a 100 strain-specific elements (Broadbent *et al.*, 2012; Kant *et al.*, 2014), which might indicate that strain-specific gene pools of different *Lactobacillus* species are relatively same in size irrespective of the phylogenetic distance, life habitat, and genome size of the given species. Methodologically, the current *Lactobacillus* pan and core-genome investigations were comparable. However, some studies involved the use of relatively simple orthologue grouping approaches (Kant *et al.*, 2014), whereas others were based on more sophisticated procedures (Broadbent *et al.*, 2012; Smokvina *et al.*, 2013; Study V). Some studies also failed to address functional annotation anomalies and inconsistencies that resulted from differences in the original protein function predictions. Given the relative high number of orthologue groups that contained sequences with inconsistent original protein function annotations (approximately 36% for *L. crispatus*; Study V), studies omitting the re-annotation phase might have misidentified the correct biological role for some orthologues. Resolving the annotation anomalies among the *L. crispatus* using the BLANNOTOATOR tool at least improved the annotation consistency and allowed the assignment of more similar functional descriptions to orthologous sequences in Study V.

Comparative genomics data were also used to resolve gene-phenotype relationships. For example, this approach was applied in Study III to proteinaceous adhesion factors of *L. rhamnosus* GG and LC705 to aid in understanding the role of these adhesins in microbe-host interactions. Specifically, the initial screen for adhesion factors identified several types of proteins in *L. rhamnosus* GG (31 adhesins) and LC705 (37 adhesins) with adhesion- or colonisation-related protein domain families. Using comparative genomics, this protein set was narrowed to eight GG-specific candidates purportedly involved in determining its host-microbe interactions. Importantly, the gene-phenotype correlation analysis exhibited high specificity and sensitivity and was able to identify SpaC, which has recently been shown to be the single most important adhesin of GG (Lebeer *et al.*, 2012), as well as other proteins for which roles in adhesion and biofilm formation (Vélez *et al.*, 2010) or mucus binding (Ossowski *et al.*, 2011) have been verified. Interestingly, the current *L. rhamnosus* genome data would have resulted in a stronger association between *SpaCBA* genes and host cell binding if *L. rhamnosus* strains LMS2-1 and E800 had shown levels of adherence that were comparable to that of GG, based on the notion that the *spaCBA* genes are present only in the genomes of *L. rhamnosus* strains GG, LMS2-1, and E800 (Kant *et al.*, 2014). All other GG-adhesins found in Study III have a counterpart in at least one other *L. rhamnosus* strain.

At the time of Study V, ten *L. crispatus* genomes were available in public databases. However, the full exploitation of these data was limited by the lack of commensurable

phenotype information for the studied bacteria. Nevertheless, the genome comparisons that were conducted in Study V revealed interesting relationships between the isolation source and putative prophage elements and between the natural habitat of these strains and adaptive immunity systems, suggesting that different types of CRISPR-Cas systems are beneficial in different niches (Study V). In addition, ortholog assignments of the *L. crispatus* and *G. vaginalis* protein complements and projection of the adhesion factor information across species was able in Study V used to identify *L. crispatus* core-genome encoded proteins that are implicated in the competitive exclusion of *G. vaginalis*, providing an explanation for the inverse association between *L. crispatus* and *G. vaginalis* colonization in the human vagina (Fredricks *et al.*, 2007; Srinivasan *et al.*, 2012; Shipitsyna *et al.*, 2013).

4.8 Phylogenetic reconstructions

Similar to molecular phylogenetic approaches (Felis & Dellaglio, 2007; Salvetti *et al.*, 2012), the phylogenomic strategy adopted in Study III revealed that *L. rhamnosus* is closely related to *L. casei*. In particular, the study revealed a close phylogenetic relationship between *L. rhamnosus* GG and LC705 and demonstrated that among the 25 LABs that were tested, these strains were particularly phylogenetically close to *L. casei* ATCC 334. The relationship between *L. casei* and *L. rhamnosus* was supported with 100% confidence, corroborating the known assignment of *L. rhamnosus* to the *L. casei* subgroup (Felis & Dellaglio, 2007; Salvetti *et al.*, 2012), together with *L. casei* and *L. paracasei* as well as *Lactobacillus zeae*, which was recently reclassified as *L. casei* by Salvetti *et al.*, 2012.

In Study V, a *Lactobacillus* phylogeny was derived from the concatenated alignment of 72,019 single-nucleotide polymorphisms from the core regions of three *L. acidophilus*, ten *L. crispatus,* five *L. helveticus*, and one *Bacillus subtilis* strains. The analysis yielded a tree with a high confidence and indicated that *L. crispatus* and *L. helveticus* were sister species to *L. acidophilus*. This finding contradicted some previous reports indicating that *L. crispatus* and *L acidophilus* cluster together first (Canchaya *et al.*, 2006; Felis & Dellaglio, 2007; Salvetti *et al.*, 2012) but agreed with some other phylogenomic studies (Kant *et al.*, 2011a). Among the *L. crispatus* cluster, the chicken isolate ST1 was the first to branch off from the others.

4.9 Intrafamily variation in *Lactobacillaceae*

In addition to the genome analyses presented in Studies III-V, a comparative analysis of the available *Lactobacillus* and *Pediococcus* genomes is presented here to better describe this family and to determine the scale and scope of the pan and core genomic potentials of these bacteria. Using BLAST (Altschul *et al.*, 1997) and OrthoMCL (Li *et al.*, 2003), the full complement of *Lactobacillaceae* protein sequences was assigned to 30,693 orthologue

groups. The pan-genome was found to be approximately 15-fold the size of a single genome and was predicted to be open, based on a power-law regression (positive exponent β=0.526±0.003, Figure 11). Notably, the pan-genome was found to grow by at least one orthologue group per additional genome until 3.18 million isolates have been sequenced, emphasising the need to sequence more *Lactobacillaceae* genomes. A relatively large fraction of the orthologue groups of any given isolate was conserved and shared by at least one other organism in this group. Genomes that were the only representative of their species, such as *Lactobacillus kisonensis* F0435, had the largest strain-specific gene pools (over 20% of their orthologue groups are orphans), whereas strains such as *L. reuteri* DSM 20016^T and JCM 1112^T, which originated from the same isolate, included only a few strain-specific elements. The average number of orphan orthologue groups in each *Lactobacillaceae* was 114, which is approximately the same average value as that reported for individual *Lactobacillus* species (Broadbent *et al.*, 2012; Kant *et al.*, 2014; Study V). As expected, the size of the core genome decreased with the addition of genomes: on average, ten genomes produced a core genome of 530 orthologue groups, whereas a set of twenty genomes produced a collection of 389 core orthologue groups. This trend is reminiscent of that observed in previous core genome analyses (Tettelin *et al.*, 2005; Bottacini *et al.*, 2010; Broadbent *et al.*, 2012). It is also interesting that the *Lactobacillaceae* core genome curve was consistent with two previous total core estimates, which suggested 383 and 363 orthologue groups for a set of 20 and 21 *Lactobacillus* genomes, respectively (Kant *et al.*, 2011a; Lukjancenko *et al.*, 2012). Notably, extrapolation of the core genome curve showed that the core genome reaches a plateau at 59±10 orthologue groups for an infinite number of *Lactobacillaceae* (Figure 11), whereas the current *Lactobacillaceae* core genome included 66 orthologue groups, a value that approximates well the predicted core genome. Therefore, the strains included in the analysis can be judged to have represented the common features of the family *Lactobacillaceae* quite well. However, because some draft genomes contain up to hundreds of sequence gaps, genuine gene products might have been ignored in this analysis, thereby causing an under-estimation of the size of the core genome and a slight over-estimation of the size of the pan-genome. The extent of this error is unknown and remains to be elucidated. Moreover, because this extrapolation curve was the first to be fitted to *Lactobacillaceae* core genome data, the validity of the goodness of estimation remains open. Nevertheless, the data undoubtedly indicate the presence of large repertoires of undiscovered genes in *Lactobacillaceae* genomes that are yet to be sequenced and suggest that the *Lactobacillaceae* core is perhaps smaller than previously anticipated (Canchaya *et al.*, 2006; Kant *et al.*, 2011a; Lukjancenko *et al.*, 2012).

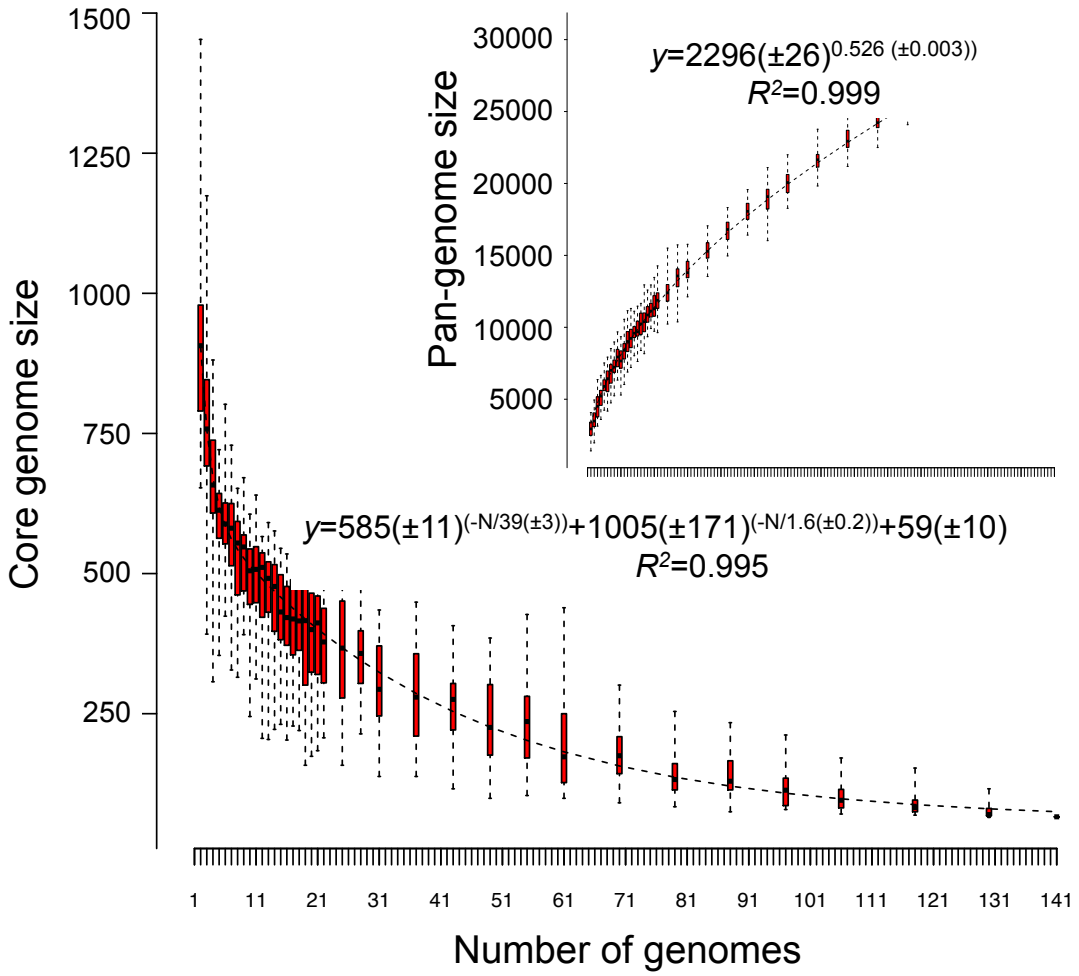


Figure 11. Pan- and core genomes of *Lactobacillaceae* according to the number of considered genomes. At each step, N genomes were chosen 50 times randomly, and the pan-genome and the core genome were recalculated. The dashed lines represent least squares fits to the medians, and R^2 describes the suitability of the fit. The boxes represent the 25–75 percentiles, the horizontal lines represent the medians, and the whisker data represent the extremes estimated from 50 iterations. The pan-genome curve is a least squares fit of the power law $y = k\,N^{\beta}$ to medians. An exponent $\beta > 0$ indicates an open pan-genome. The regression analysis for the core genome was performed by fitting a double exponential decay $y = \kappa_1^{(-N/\tau 1)} + \kappa_2^{(-N/\tau 2)} + \Omega$ to the medians with a least square regression. Ω is the asymptotic core genome size.

5 Conclusions

Whole-genome sequencing of bacteria and genome mining is a powerful approach for addressing microbiological questions (Fleischmann *et al.*, 1995; Mardis, 2008; Loman *et al*., 2012). However, the extraction of biological knowledge involves the handling of vast amounts of sequence data and is largely beyond human limits; thus, researchers rely on sophisticated computational procedures (Edwards & Holt, 2013; Richardson & Watson, 2013; Ali *et al.*, 2013). In this thesis, two new algorithms were developed for laborious genome annotation tasks: LOCP and BLANNOTATOR. The LOCP service is based on the detection of genomic segments that are enriched in sortase and pilin-resembling genes and provides a major step forward in understanding the distribution of pilus appendixes in gram-positive bacteria. The other novel tool was designed for the functional annotation of bacterial protein sequences. Based on sequence similarity searches and clustering the annotation space, BLANNOTATOR yields DEs that are accurate and that do not suffer from annotation anomalies. While developing the algorithm, DEs were preferred over GO annotations, because of their great value in inferring host-interaction factors. In addition to automated protein function prediction methods, this study focused on the genomics of *Lactobacillus* and demonstrated the power of computer-assisted genome annotation for improving our understanding of the biochemistry, niche-adaptation, and host-interactions of lactobacilli. In particular, three *Lactobacillus* genomes were sequenced and annotated. Genome analysis identified genes that are instrumental for survival and of industrial value. Importantly, for each of the investigated species, comparative genomics identified host-interaction factors, the roles of which were validated using experimental methodologies.

First, two software tools were designed for laborious bacterial genome annotation processes for which bioinformatics solutions did not exist prior to these studies: one for locating pilus operons and another for DE prediction. In Study I, a new algorithm was presented to manage the cumbersome and slow process of locating pilus operons in gram-positive genomes. Unlike the error-prone and arduous manual curation of pilus gene regions, LOCP was designed for ease of use and to accurately locate pilus operons from any set of sequences. For example, LOCP analysis of the 135 *Lactobacillus* genomes that are listed in Appendix Table 2 revealed 29 potential pilus carriers among these organisms, many of which have been recently verified using other approaches (Broadbent *et al.*, 2012; Kant *et al.*, 2014). In Study II, an effort was made to explore the possibilities for improving annotation anomalies and inconsistencies, which are rather common in sequence databases (Andorf *et al.*, 2007; Schnoes *et al.*, 2009). During the manual curation processes, existing protein function prediction tools were noted to inconsistently classify orthologues and genes that were judged to share the same protein functions. Because manual correction of these annotation errors was laborious, BLANNOTATOR was developed in Study II. Comparisons based on simulated data indicated that BLANNOTATOR was better than the five other protein function prediction methods that were tested in Study II and that this method made function calls with a highly consistent nomenclature. However, even the simplest protein function prediction strategies, such as the best BLAST approach, performed well in Study II; this illustrates that substantial improvements using homology-based annotation alone are unlikely to be made in the

future. However, the most important benefit of the method developed in Study II resulted from its ability to systematise DE calls, thereby influencing the manual curation process and the comparative analysis of protein function. The ability to classify proteins into consistent classes was especially useful for determining host-interaction factors and when dealing with multiple organisms, such as in Study V, where the functional catalogues of ten *L. crispatus* strains needed to be commensurable.

Second, the genomes of *L. rhamnosus* GG and LC705 were sequenced and annotated in Study III. These organisms were chosen for genomic investigation to map molecules that might explain their successful use in preventing and treating various diseases (Hojsak *et al.*, 2010; Hatakka *et al.*, 2001; Isolauri *et al.*, 1991; Guandalini *et al.*, 2000; Szajewska & Mrukowicz, 2001; Kalliomäki *et al.*, 2001; Kalliomäki *et al.*, 2003; Kalliomäki *et al.*, 2007; Kajander *et al.*, 2005; Kajander *et al.*, 2008) and to identify the bacterial components that are responsible for their adhesion to human tissues (Jacobsen *et al.*, 1999; Tuomola *et al.*, 2000; Saxelin *et al.*, 2010). Among the approximately 3,000 genes in each strain, Study III disclosed those that are implicated in EPS biosynthesis, sugar and peptide usage, and the production of a putative bacteriocin. The study also revealed the presence of two gram-positive pilus gene clusters in *L. rhamnosus* genomes. Using immunoblotting and immunogold electron microscopy, the expression of a *spaCBA* gene cluster in strain GG was in Study III shown, and the presence of this pilus structure on the cell surface was confirmed. This finding represented the first description of such a pilus structure in lactobacilli and established the presence of these adhesion structures in non-pathogenic bacteria. Furthermore, the mucus-binding capacity of the SpaC subunit and its importance for the adhesion between strain GG and human intestinal mucus was demonstrated in Study III. Recently, further evidence has accumulated that describes the role of SpaC pilin in adhesion (von Ossowski e*t al.*, 2010; Lebeer *et al.*, 2012), confirming that this pilin plays a crucial role in the adhesion of strain GG to human tissues. Although recent studies have demonstrated the absence of SpaFED pili on the cell surface of *L. rhamnosus* GG (Reunanen e*t al.*, 2010), the fact that the *SpaFED* genes are present in the genomes of many *L. rhamnosus* (Douillard *et al.*, 2013b; Kant *et al.*, 2014) and *L. casei* (Broadbent *et al.*, 2012) strains corroborates their potential importance in the adhesion processes of *L. casei* group bacteria. Notably, the discovery that the SpaC pilin is essential for the efficient adherence of strain GG to human tissues was based on ortholog analysis and represents an impressive demonstration of the power of comparative genomics in predicting host-interaction factors in lactobacilli. Specifically, 31 proteins purportedly involved in adhesion or colonisation were detected in the GG. Only eight of these proteins were however predicted to be GG-specific, thus providing a plausible explanation for the differing adhesion characteristics of the two *L. rhamnosus* strains to human mucus (Jacobsen *et al.*, 1999; Tuomola *et al.*, 2000; Saxelin *et al.*, 2010). It is clear that without this type of analysis, all 31 of the genes identified in Study III would have had to been validated experimentally. In contrast to the successful detection of adhesion factor components, the search for molecules mediating immune responses failed, because not much known about the role of LC705 in disease alleviation. This failure indicates the role of commensurable phenotype information in comparative genomics. Overall, the *L. rhamnosus* GG and LC705 genomes are the first two genomes of free-living organisms to

have been sequenced in Finland and have provided a notable framework for various *Lactobacillus* comparative genomics studies and other studies that attempt to understand the mechanism underlying the interaction of probiotics with host tissues.

Finally, the genomic potential of *L. crispatus* ST1 was determined in Study IV and compared to those of nine vaginal *L. crispatus* isolates in Study V. Chicken-isolated ST1 was subjected to whole-genome sequencing to reveal the coding regions of genes encoding LEA and other adhesins. Although a complete genome could not be produced in the study, the project generated a high-quality reference genome with only one unresolved region (within the *lea* gene) and provided a valuable insight to the genomic foundations of an important urogenital species. Bioinformatic analysis of the nearly complete genome of ST1 identified approximately 2,100 genes, including genes that are implicated in EPS biosynthesis, antimicrobial activity, acquired resistance against bacteriophages, and adhesion. The metabolic pathways constructed in Study V revealed auxotrophy for 12 amino acids. The analysis also highlighted the presence of both pentose phosphate and glycolytic pathways in *L. crispatus*, defining the species as facultatively heterofermentative; this finding contradicts earlier assumptions (Salvetti *et al.*, 2012; Salvetti *et al.*, 2013). Extensive comparative genomic analysis provided evidence for considerable sequence identity and synteny among the genomes of the ten *L. crispatus* isolates that were investigated in Study V. Moreover, the study revealed genes in the *L. crispatus* core genome coding for adhesins that are involved in the competitive exclusion of *G. vaginalis* from vaginal cells, thus providing an attractive explanation for the proven inverse association between *L. crispatus* and *G. vaginalis* colonisation in the human vagina (Fredricks *et al.*, 2007; Srinivasan *et al.*, 2012; Shipitsyna *et al.*, 2013). Importantly, the competitive exclusion process was demonstrated to result from the core-encode LEA protein (Edelman *et al.*, 2012) by measuring the adhesion capacity of vaginal *L. crispatus* and *G. vaginalis* isolates to vaginal cells in the presence and absence of pretreatment with LEA-specific Fab fragments. Collectively, Studies IV and V established the genetic landscape of key urogenital lactobacilli. They suggested significant relationships between the life environments of different *L. crispatus* isolates and their adaptive immunity systems and revealed proteins that might protect the vagina from *G. vaginalis* and bacterial vaginosis. These studies also demonstrated the benefit of including phenotypically characterised strains in sequencing projects and indicated that the linking of bacterial traits to genes is easier when using well-characterised strains (as in Study III).

In conclusion, Studies III-V highlighted the power of whole-genome sequencing for generating new hypotheses about lactobacilli and their host-interaction factors. However, the success of sequencing studies is affected by several choices. First, one needs to choose an appropriate sequencing method. At present, lactobacilli have been sequenced largely using the Roche 454 platform (Appendix Table 2), which provides a long read length and good accuracy (Liu *et al.*, 2012; Quail *et al.*, 2012). This sequencing approach is feasible, given that longer reads appear to result in assemblies that are more continuous than shorter reads. However, sequencing costs might be substantially reduced without sacrificing too much quality using other sequencing machines that yield comparable read lengths at a lower price per sample (Table 1). Second, if the assembly phase fails to assign reads into single contigs, the desired genome-finishing level needs to be chosen. Although draft

genomes are cheap and fast to produce, they miss data and can suffer from truncated genes (Klassen & Currie, 2012). The problems are mainly caused during the assembly of low-coverage areas, ribosomal gene clusters, and repetitive elements; however, there is no guarantee that important genes will not be missed in draft genomes. Nevertheless, the risk of missing an important gene is rather small, as was evident from a high level of sequence relatedness and colinearity that was in Study V found between the draft and high-quality *L. crispatus* genomes.

The third choice to make is to decide the level of manual curation of the genome data. In Studies III and IV, a semi-automated annotation strategy was used, whereby the gene start sites and functional annotations were subjected to manual curation. In this process, the manual curation of gene start sites was based on comparing the gene models among lactobacilli and affected only a small fraction of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 CDSs, suggesting that the current gene-calling systems are accurate (as shown in Table 8) or, at least, call genes consistently across lactobacilli. The manual curation of function calls involved comparing protein function calls, reviewing the protein functions of orthologous sequences, and choosing the best alternatives. This step made the data more interpretable and enabled the description of gene functions with nomenclature that was more consistent; however, this step affected only some hundreds of genes in each genome. Moreover, the refinements made were mostly cosmetic and rarely changed the predicted protein function completely, indicating that the manual curation did not provide any additional advantage for a large majority of *L. rhamnosus* GG, *L. rhamnosus* LC705, and *L. crispatus* ST1 proteins. Specifically, the bioinformatic tools that were used in Studies III-V (Table 6) performed remarkably well for enzymes, phage-related proteins, and transcription factors in the given bacteria. In contrast, the classification of genes that were implicated in EPS production, adhesion, and antimicrobial activity was less satisfactory (Studies III-V). Thus, it was necessary to generate annotation data for these proteins using specific tools and to manually refine the results provided by the automated methods. It was also noted that in various instances, DEs were more informative than the GO terms of the corresponding gene products. In particular, GO terms were of limited value in the quest for genes that are involved in immunomodulatory processes, bacterial-host interaction, and the production of bacteriocins. Unfortunately, the large vocabulary that is present in DEs precludes their use for the automated cataloguing of protein genes into categories. Surprisingly, the various gene callers and protein function prediction approaches examined provided results of almost equal quality, indicating that the choice of gene caller or automated protein function prediction method might not be as critical as stated previously (Bakke *et al.*, 2009). It remains to be determined whether these findings are specific to *Lactobacillus* or can be generalised to a wider array of bacteria.

The power of comparative genomics and genome island prediction in identifying genes determining host-microbe interactions was also demonstrated. In Study III, comparative analyses suggested key roles for the SpaCBA pilins in the adhesion of *L. rhamnosus* GG to human tissues. In Study V, protein comparison data revealed a role for the LEA protein in the competitive exclusion of *G. vaginalis* from vaginal cells. Notably, both discoveries were confirmed experimentally. Based on the successful analyses, it can be concluded that comparative genomics provides an appealing starting point to call host-interaction factors

and gene-phenotype associations in *Lactobacillus*. However, the success of comparative approaches in resolving correlations between genotypes and phenotypes appears to depend more on the amount of commensurate biological information available for each organism (Study III) than on the number of genomes under study (Study V). Comparative analyses should therefore be based on organisms for which commensurate empirical information is available rather than on the genome assemblies of poorly characterised bacteria, such as those generated as a part of the HMP (Nelson *et al.*, 2010). The HMP is however useful for studies aiming to find functions universally conserved in the given set of organisms, as in Study V, where comparative genomics was used to investigate the genetic mechanisms underlying the inverse association between *L. crispatus* and *G. vaginalis* colonisation in the human vagina (Fredricks *et al.*, 2007; Srinivasan *et al.*, 2012; Shipitsyna *et al.*, 2013). Genomic island calling represents another useful computational method for the annotation of host-microbe interaction, as was exemplified in Study III by the finding of an island in *L. rhamnosus* GG that contains genes for 3 secreted pilins and a pilin-dedicated sortase. The approach was also useful in the annotation of putative prophage elements in vaginal *L. crispatus* strains. Based on Studies III-V, it can be argued that the use of genomic island predictors for *Lactobacillus* genomes is advisable and these predictors can reveal genomic regions that code for functions that differentiate the given strain from others. Importantly, this approach can even be fruitful for identifying genes that underlie strain-specific traits in the absence of evolutionary counterparts. Nevertheless, orthology grouping and genome island prediction are simply methods that allow finding patterns of sequence conservation. To understand the biological relevance of the conservation patterns and to filter patterns, sequences underlying these patterns need to be associated with function information, using tools such as LOCP (Study I) and BLANNOTATOR (Study II). For example, application of the LOCP tool to the Study III data helped to understand which of the 364 GG-specific proteins are relevant for adhesion and provided additional information for the support of pilus-encoding gene clusters in strains GG and LC705. BLANNOTATOR on the hand played a pivotal role in Study V and provided to classify proteins into consistent classes. It is clear that without this analysis, functions of many host-interaction factors would have remained undetected.

To conclude, this study has described two bioinformatics algorithms for cumbersome genome annotation tasks and has disclosed the genomes of two also human-associated *Lactobacillus* species: *L. rhamnosus* and *L. crispatus*. The algorithms yielded impressive accuracy and were of great value in improving our understanding of *Lactobacillus* host-interaction factors. Annotation of the *L. rhamnosus* and *L. crispatus* genomes has provided new insights into the physiological, genetic, biochemical, and fermentative properties of two biomedically important *Lactobacillus* species. Markedly, analyses revealed molecules involved in host-interaction and those that might protect the vagina from pathogen attack, thereby representing a major advance in understanding the host-interaction mechanisms of lactobacilli.

Acknowledgements

The study was mainly carried out at the Institute of Biotechnology and Faculty of Biomedicine, University of Helsinki during 2007–2010. Some additional studies were carried out during 2010–2014 at Valio Ltd. and Faculty Biomedicine, University of Helsinki. The work was financially supported by Valio Ltd and the Viikki Doctoral Programme in Molecular Biosciences, Biocenter Finland.

I am most grateful to my supervisor, professor Liisa Holm, for her guidance and provision of splendid facilities. It has been a pleasure to be part of her research group. My co-authors in Valio Ltd., Faculty of Veterinary Medicine and Faculty of Microbiology are also acknowledged for their valuable contribution. Everything that I know about bacteria I have learned from you. I am especially grateful to professor Willem de Vos and Dr Soile Tynkkynen for their through guidance. My co-authors and colleagues at the Institute of Biotechnology are thanked for genome sequencing services and helpful discussions on topics of bioinformatics. Professors Mauno Vihinen and docent David Fewer are thanked for reviewing my thesis. Finally, I want to thank my family for supporting me during my long days and nights.

References

Abascal F, Valencia A: Automatic annotation of protein function based on family identification. *Proteins* 2003, 53(3):683-692.

Abriouel H, Benomar N, Pérez Pulido R, Cañamero MM, Gálvez A: Annotated Genome Sequence of Lactobacillus pentosusMP-10, Which Has Probiotic Potential, from Naturally Fermented Aloreña Green Table Olives. *J Bacteriol* 2011, 193(17):4559-4560.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod J, Mayer P, Kawashima E: Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* 2000, 28(20):e87-e87.

Ahrne S, Nobaek S, Jeppsson B, Adlerberth I, Wold A, Molin G: The normal *Lactobacillus* flora of healthy human rectal and oral mucosa. *J Appl Microbiol* 1998, 85(1):88-94.

Ai L, Chen C, Zhou F, Wang L, Zhang H, Chen W, Guo B: Complete Genome Sequence of the Probiotic Strain *Lactobacillus casei* BD-II. *J Bacteriol* 2011, 193(12):3160-3161.

Akhter S, Aziz RK, Edwards RA: PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* 2012, 40(16):e126.

Ali A, Soaeres SC, Barbosa E, Santos AR, Barh D, Bakhtiar SM, Hassan SS, Ussery DW, Silva A, Miyoshi A, Azevedo V: Microbial Comparative Genomics: An Overview of Tools and Insights Into The *Genus Corynebacterium*. *J Bacteriol Parasitol* 2013, 4(2):167.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C: Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol* 2012, 8(5):e1002514.

Altermann E, Russell WM, Azcarate-Peril MA, Barrangou R, Buck BL, McAuliffe O, Souther N, Dobson A, Duong T, Callanan M*, et al*: Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. *Proc Natl Acad Sci U S A* 2005, 102(11):3906-3912.

Altman T, Travers M, Kothari A, Caspi R, Karp P: A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* 2013, 14(1):112.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.

Anderson S: Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res* 1981, 9(13):3015-3027.

Andorf C, Dobbs D, Honavar V: Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics* 2007, 8(1):284.

Angelova, M, Slobodan K, Ljupco K: Computational methods for gene finding in prokaryotes. *ICT Innovations* 2010: 11-20.

Angiuoli S, Hotopp JD, Salzberg S, Tettelin H: Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* 2011, 12(1):272.

Antikainen J, Anton L, Sillanpää J, Korhonen TK: Domains in the S-layer protein CbsA of *Lactobacillus crispatus* involved in adherence to collagens, laminin and lipoteichoic acids and in self-assembly. *Mol Microbiol* 2002, 46(2):381-394.

Anukam KC, Osazuwa E, Osemene GI, Ehigiagbe F, Bruce AW, Reid G: Clinical study comparing probiotic *Lactobacillus* GR-1 and RC-14 with metronidazole vaginal gel to treat symptomatic bacterial vaginosis. *Microbes Infect* 2006, 8(12-13):2772-2776.

Arita M: The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci U S A* 2004, 101(6):1543-1547.

Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes H, Horn M, Rattei T: Sequence-based prediction of type III secreted proteins. *PLoS Pathog* 2009, 5(4):e1000376.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25(1):25-29.

Aureli P, Capurso L, Castellazzi AM, Clerici M, Giovannini M, Morelli L, Poli A, Pregliasco F, Salvini F, Zuccotti GV: Probiotics and health: an evidence-based review. *Pharmacol Res* 2011, 63(5):366-376.

Axelsson L, Rud I, Naterstad K, Blom H, Renckens B, Boekhorst J, Kleerebezem M, van Hijum S, Siezen RJ: Genome Sequence of the Naturally Plasmid-Free *Lactobacillus plantarum* Strain NC8 (CCUG 61730). *J Bacteriol* 2012, 194(9):2391-2392.

Azcarate-Peril MA, Altermann E, Goh YJ, Tallon R, Sanozky-Dawes RB, Pfeiler EA, O'Flaherty S, Buck BL, Dobson A, Duong T, *et al*: Analysis of the genome sequence of *Lactobacillus gasseri* ATCC 33323 reveals the molecular basis of an autochthonous intestinal organism. *Appl Environ Microbiol* 2008, 74(15):4610-4625.

Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M, *et al*: The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 2008, 9(1):75.

Badger JH, Olsen GJ: CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 1999, 16(4):512-524.

Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Puy GA, Axelsen K, Baratin D, Blatter M, Boeckmann B, *et al*: The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2008, 36(suppl 1):D190-D195.

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

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, *et al*: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012, 19(5):455-477.

Barabási A, Oltvai ZN: Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004, 5(2):101-113.

Bartels D, Kespohl S, Albaum S, Drüke T, Goesmann A, Herold J, Kaiser O, Pühler A, Pfeiffer F, Raddatz G, *et al*: BACCardI—a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison. *Bioinformatics* 2005, 21(7):853-859.

Båth K, Roos S, Wall T, Jonsson H: The cell surface of *Lactobacillus reuteri* ATCC 55730 highlighted by identification of 126 extracellular proteins from the genome sequence. *FEMS Microbiol Lett* 2005, 253(1):75-82.

Baumbach J, Tauch A, Rahmann S: Towards the integrated analysis, visualization and reconstruction of microbial gene regulatory networks. *Brief Bioinform* 2009, 10(1):75-83.

Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 2005, 6(1):167.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: GenBank. *Nucleic Acids Res* 2013, 41(Database issue):D36-42.

Bentley DR: Whole-genome re-sequencing. *Curr Opin Genet Dev* 2006, 16(6):545-552.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, *et al*: Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008, 456(7218):53-59.

Berger B, Pridmore RD, Barretto C, Delmas-Julien F, Schreiber K, Arigoni F, Brüssow H: Similarity and Differences in the *Lactobacillus acidophilus* Group Identified by Polyphasic Analysis and Comparative Genomics. *J Bacteriol* 2007, 189(4):1311-1321.

Berlec A, Zadravec P, Jevnikar Z, Štrukelj B: Identification of Candidate Carrier Proteins for Surface Display on *Lactococcus lactis* by Theoretical and Experimental Analyses of the Surface Proteome. *Appl Environ Microbiol* 2011, 77(4):1292-1300.

Bernardeau M, Guguen M, Vernoux JP: Beneficial lactobacilli in food and feed: long‐term use, biodiversity and proposals for specific and realistic safety assessments. *FEMS Microbiol Rev* 2006, 30(4):487-513.

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

Besemer J, Lomsadze A, Borodovsky M: GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001, 29(12):2607-2618.

Beuf KD, Schrijver JD, Thas O, Criekinge WV, Irizarry RA, Clement L: Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model. *BMC Bioinformatics* 2012, 13(1):303.

Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, Jia S, Deng Z, Rajakumar K, Ou H: ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res* 2012, 40(D1):D621-D626.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED*, et al*: Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res* 2004, 14(4):708-715.

Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P: CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007, 8(1):209.

Boekhorst J, Helmer Q, Kleerebezem M, Siezen RJ: Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology* 2006, 152(1):273-280.

Boekhorst J, Siezen RJ, Zwahlen M, Vilanova D, Pridmore RD, Mercenier A, Kleerebezem M, de Vos WM, Brüssow H, Desiere F: The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content. *Microbiology* 2004, 150(11):3601-3611.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011, 27(4):578-579.

Boisvert S, Laviolette F, Corbeil J: Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 2010, 17(11):1519-1533.

Bottacini F, Medini D, Pavesi A, Turroni F, Foroni E, Riley D, Giubellini V, Tettelin H, van Sinderen D, Ventura M: Comparative genomics of the genus *Bifidobacterium*. *Microbiology* 2010, 156(11):3243-3254.

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D: Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* 2004, 5(5):R35.

Broadbent JR, Neeno-Eckwall EC, Stahl B, Tandee K, Cai H, Morovic W, Horvath P, Heidenreich J, Perna NT, Barrangou R*, et al*: Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation. *BMC Genomics* 2012, 13:533-2164-13-533.

Brouwer RW, van den Hout MC, Grosveld FG, van Ijcken WF: NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics* 2012, 28(2):284-285.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: Universal trees based on large combined protein sequence data sets. *Nat Genet* 2001, 28(3):281-285.

Burstein D, Zusman T, Degtyar E, Viner R, Segal G, Pupko T: Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog* 2009, 5(7):e1000508.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008, 18(5):810-820.

Cai H, Thompson R, Budinich MF, Broadbent JR, Steele JL: Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution. *Genome Biol* Evol 2009, 1:239.

Callanan M, Kaleta P, O'Callaghan J, O'Sullivan O, Jordan K, McAuliffe O, Sangrador-Vegas A, Slattery L, Fitzgerald GF, Beresford T*, et al*: Genome Sequence of *Lactobacillus helveticus*, an Organism Distinguished by Selective Gene Loss and Insertion Sequence Element Expansion. *J Bacteriol* 2008, 190(2):727-735.

Canchaya C, Claesson MJ, Fitzgerald GF, van Sinderen D, O'Toole PW: Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. *Microbiology* 2006, 152(11):3185-3196.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* 2009, 37(suppl 1):D233-D238.

Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, Parkhill J: ACT: the Artemis comparison tool. *Bioinformatics* 2005, 21(16):3422-3423.

Chabot S, Yu HL, De Léséleuc L, Cloutier D, Van Calsteren MR, Lessard M, Roy D, Lacroix M, Oth D: Exopolysaccharides from *Lactobacillus rhamnosus* RW-9595M stimulate TNF, IL-6 and IL-12 in human and mouse cultured immunocompetent cells, and IFN-γ in mouse splenocytes. *Le Lait* 2001, 81(6):683-697.

Chaillou S, Champomier-Verges MC, Cornet M, Crutz-Le Coq AM, Dudez AM, Martin V, Beaufils S, Darbon-Rongere E, Bossy R, Loux V, *et al*: The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. *Nat Biotechnol* 2005, 23(12):1527-1533.

Chan RC, Reid G, Irvin RT, Bruce AW, Costerton JW: Competitive exclusion of uropathogens from human uroepithelial cells by *Lactobacillus* whole cells and cell wall fragments. *Infect Immun* 1985, 47(1):84-89.

Chen F, Mackey AJ, Vermunt JK, Roos DS: Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2007, 2(4):e383.

Chen C, Ai L, Zhou F, Wang L, Zhang H, Chen W, Guo B: Complete Genome Sequence of the Probiotic Bacterium *Lactobacillus casei* LC2W. *J Bacteriol* 2011, 193(13):3419-3420.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, Suhai S: Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004, 14(6):1147-1159.

Cho Y, Choi JK, Kim J, Lim Y, Ham J, Kang D, Chun J, Paik H, Kim G: Genome Sequence of *Lactobacillus salivarius* GJ-24, a Probiotic Strain Isolated from Healthy Adult Intestine. *J Bacteriol* 2011, 193(18):5021-5022.

Chou C, Chang W, Chiu C, Huang C, Huang H: FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res* 2009, 37(suppl 2):W129-W134.

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

Claesson MJ, Li Y, Leahy S, Canchaya C, van Pijkeren JP, Cerdeño-Tárraga AM, Parkhill J, Flynn S, O'Sullivan GC, Collins JK, *et al*: Multireplicon genome architecture of *Lactobacillus salivarius*. *Proc Natl Acad Sci U S A* 2006, 103(17):6718-6723.

Clark WT, Radivojac P: Analysis of protein function and its prediction from amino acid sequence. *Proteins* 2011, 79(7):2086-2096.

Claudel-Renard C, Chevalet C, Faraut T, Kahn D: Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 2003, 31(22):6633-6639.

Collado M, Delgado S, Maldonado A, Rodriguez J: Assessment of the bacterial diversity of breast milk of healthy women by quantitative real-time PCR. *Lett Appl Microbiol* 2009, 48(5):523-528.

Collins FS, Weissman SM: Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proc Natl Acad Sci U S A* 1984, 81(21):6812-6816.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005, 15(7):901-913.

Corr SC, Li Y, Riedel CU, O'Toole PW, Hill C, Gahan CG: Bacteriocin production as a mechanism for the antiinfective activity of *Lactobacillus salivarius* UCC118. *Proc Natl Acad Sci U S A* 2007, 104(18):7617-7621.

Cotter PD, Hill C, Ross RP: Bacteriocins: developing innate immunity for food. *Nat Rev Microbiol* 2005, 3(10):777-788.

Cox MP, Peterson DA, Biggs PJ: SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010, 11(1):485.

Cruveiller S, Le Saux J, Vallenet D, Lajus A, Bocs S, Medigue C: MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes. *Nucleic Acids Res* 2005, 33(suppl 2):W471-W479.

Daeschel MA: Antimicrobial factors from lactic acid bacteria for use as food preservatives. *Food Technol* 1989, 43:484–490

Dal Bello F, Hertel C: Oral cavity as natural reservoir for intestinal lactobacilli. *Syst Appl Microbiol* 2006, 29(1):69-76.

Damelin LH, Paximadis M, Mavri-Damelin D, Birkhead M, Lewis DA, Tiemessen CT: Identification of predominant culturable vaginal *Lactobacillus* species and associated bacteriophages from women with and without vaginal discharge syndrome in South Africa. *J Med Microbiol* 2011, 60(Pt 2):180-183.

Dandekar T, Snel B, Huynen M, Bork P: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 1998, 23(9):324-328.

Darling AE, Mau B, Perna NT: progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010, 5(6):e11147.

Darling AC, Mau B, Blattner FR, Perna NT: Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004, 14(7):1394-1403.

Darzentas N: Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* 2010, 26(20):2620-2621.

De Keersmaecker SC, Verhoeven TL, Desair J, Marchal K, Vanderleyden J, Nagy I: Strong antimicrobial activity of *Lactobacillus rhamnosus* GG against *Salmonella typhimurium* is due to accumulation of lactic acid. *FEMS Microbiol Lett* 2006, 259(1):89-96.

Delcher AL, Bratke KA, Powers EC, Salzberg SL: Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 2007, 23(6):673-679.

De Man, JC, Rogosa D, and Sharpe ME: A medium for the cultivation of lactobacilli. *J appl Bacteriol* 1960, 23(1):130-135.

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

Devos D, Valencia A: Intrinsic errors in genome annotation. *Trends Genet* 2001, 17(8):429-431.

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 2005, 15(2):330-340.

Donnarumma G, Molinaro A, Cimini D, De Castro C, Valli V, De Gregorio V, De Rosa M, Schiraldi C: *Lactobacillus crispatus* L1: high cell density cultivation and exopolysaccharide structure characterization to highlight potentially beneficial effects against vaginal pathogens. *BMC Microbiol* 2014, 14(1):137.

Douillard FP, Ribbera A, Järvinen HM, Kant R, Pietilä TE, Randazzo C, Paulin L, Laine PK, Caggia C, von Ossowski I, et al: Comparative Genomic and Functional Analysis of *Lactobacillus casei* and *Lactobacillus rhamnosus* Strains Marketed as Probiotics. *Appl Environ Microbiol.* 2013a, 79(6):1923-1933.

Douillard FP, Ribbera A, Kant R, Pietila TE, Jarvinen HM, Messing M, Randazzo CL, Paulin L, Laine P, Ritari J, *et al*: Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet* 2013b, 9(8):e1003683.

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

du Toit M, Engelbrecht L, Lerm E, Krieger-Weber S: *Lactobacillus*: the next generation of malolactic fermentation starter cultures—an overview. *Food Bioprocess Technol* 2011, 4(6):876-906.

Dubchak I, Poliakov A, Kislyuk A, Brudno M: Multiple whole-genome alignments without a reference organism. *Genome Res* 2009, 19(4):682-689.

Durot M, Bourguignon P, Schachter V: Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev* 2009, 33(1):164-190.

Earl D, Bradnam K, St. John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, *et al*: Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res* 2011, 21(12):2224-2241.

Eddy SR: Accelerated profile HMM searches. *PLoS Comput Biol* 2011, 7(10):e1002195.

Eddy SR: Computational genomics of noncoding RNA genes. *Cell* 2002, 109(2):137-140.

Edelman S, Leskela S, Ron E, Apajalahti J, Korhonen TK: In vitro adhesion of an avian pathogenic *Escherichia coli* O78 strain to surfaces of the chicken intestinal tract and to ileal mucus. *Vet Microbiol* 2003, 91(1):41-56.

Edelman S, Westerlund-Wikstrom B, Leskela S, Kettunen H, Rautonen N, Apajalahti J, Korhonen TK: In Vitro Adhesion Specificity of Indigenous Lactobacilli within the Avian Intestinal tract. *Appl Environ Microbiol* 2002, 68(10):5155-5159.

Edelman SM, Lehti TA, Kainulainen V, Antikainen J, Kylväjä R, Baumann M, Westerlund-Wikstrom B, Korhonen TK: Identification of a high-molecular-mass *Lactobacillus* epithelium adhesin (LEA) of *Lactobacillus crispatus* ST1 that binds to stratified squamous epithelium. *Microbiology* 2012, 158(Pt 7):1713-1722.

Edgar RC: PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* 2007, 8(1):18.

Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, 32(5):1792-1797.

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

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, *et al*: Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* 2009, 323(5910):133-138.

Eisen JA: Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res* 1998, 8(3):163-167.

Ely B, Scott LE: Correction of the *Caulobacter crescentus* NA1000 Genome Annotation. *PLoS ONE* 2014, 9(3):e91668.

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol* 2005, 1(5):e45.

Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE: Combo: a whole genome comparative browser. *Bioinformatics* 2006, 22(14):1782-1783.

Erlich Y, Mitra PP: Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* 2008, 5(8):679-682.

Ewing B, Green P: Base-calling of automated sequencer traces usingPhred. II. error probabilities. *Genome Res* 1998, 8(3):186-194.

Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998, 8(3):175-185.

Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia E, Velasco R, Fontana P: Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics* 2012, 13:S14.

Felis GE, Dellaglio F: Taxonomy of Lactobacilli and Bifidobacteria. *Curr Issues Intest Microbiol* 2007, 8(2):44-61.

Felsenstein J: Phylip; Phylogeny Inference Package Version 3.2. *Cladistics* 1989, 5:164-166.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*: Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995, 269(5223):496-512.

Fleischner H: *Eulerian graphs and related topics:* Elsevier; 1990.

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

Forde BM, Neville BA, O'Donnell MM, Riboulet-Bisson E, Claesson MJ, Coghlan A, Ross RP, O'Toole PW: Genome sequences and comparative genomics of two *Lactobacillus ruminis* strains from the bovine and human intestinal tracts. *Microb Cell Fact* 2011, 10(Suppl 1):S13.

Fouts DE: Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 2006, 34(20):5839-5851.

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, *et al*: The phylogeny of prokaryotes. *Science* 1980, 209(4455):457-463.

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Res* 2003, 13(1):1-12.

Fredricks DN, Fiedler TL, Thomas KK, Oakley BB, Marrazzo JM: Targeted PCR for detection of vaginal bacteria associated with bacterial vaginosis. *J Clin Microbiol* 2007, 45(10):3270-3276.

Frese SA, Benson AK, Tannock GW, Loach DM, Kim J, Zhang M, Oh PL, Heng NC, Patil PB, Juge N, *et al*: The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet* 2011, 7(2):e1001314.

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

Frishman D, Mironov A, Mewes H, Gelfand M: Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res* 1998, 26(12):2941-2947.

Frith MC, Hamada M, Horton P: Parameters for accurate genome alignment. *BMC Bioinformatics* 2010, 11(1):80.

Frost LS, Leplae R, Summers AO, Toussaint A: Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005, 3(9):722-732.

Gao F, Zhang C: Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics* 2008, 9(1):79.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X: Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009, 25(12):i54-i62.

Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z: RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res* 2011, 39(14):5845-5852.

Garg A, Gupta D: VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics* 2008, 9(1):62.

Gascuel O: BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol* 1997, 14(7):685-695.

Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, *et al*: PATRIC: the Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infect Immun* 2011, 79(11):4286-4298.

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

Goudenège D, Avner S, Lucchetti-Miganeh C, Barloy-Hubler F: CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC microbiol* 2010, 10(1):88.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: Rfam: an RNA family database. *Nucleic Acids Res* 2003, 31(1):439-441.

Grissa I, Vergnaud G, Pourcel C: CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 2007, 35(suppl 2):W52-W57.

Guandalini S, Pensabene L, Zikri MA, Dias JA, Casali LG, Hoekstra H, Kolacek S, Massar K, Micetic-Turk D, Papadopoulou A, *et al*: *Lactobacillus* GG administered in oral rehydration solution to children with acute diarrhea: a multicenter European trial. *J Pediatr Gastroenterol Nutr* 2000, 30(1):54-60.

Guinane CM, Kent RM, Norberg S, Hill C, Fitzgerald GF, Stanton C, Ross RP: Host specific diversity in *Lactobacillus johnsonii* as evidenced by a major chromosomal inversion and phage resistance mechanisms. *PLoS One* 2011, 6(4):e18740.

Guindon S, Gascuel O: A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* 2003, 52(5):696-704.

Guo F, Ou H, Zhang C: ZCURVE: a new system for recognizing protein‐coding genes in bacterial and archaeal genomes. *Nucleic Acids Res* 2003, 31(6):1780-1789.

Haft DH, Selengut J, Mongodin EF, Nelson KE: A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* 2005, 1(6):e60.

Haft DH, Selengut JD, White O: The TIGRFAMs database of protein families. *Nucleic Acids Res* 2003, 31(1):371-373.

Ham J, Kim H, Seol K, Jang A, Jeong S, Oh M, Kim D, Kang D, Kim G, Cha C: Genome Sequence of *Lactobacillus salivarius* NIAS840, Isolated from Chicken Intestine. *J Bacteriol* 2011, 193(19):5551-5552.

Hammes WP, Vogel RF: The genus *Lactobacillus*. In *The Genera of Lactic Acid Bacteria. Volume 2*. Edited by Wood BJB, Holzapfel WH. Springer; 1995:19-54.

Haridas S, Breuill C, Bohlmann J, Hsiang T: A biologist's guide to de novo genome assembly using next-generation sequence data: A test with fungal genomes. *J Microbiol Methods* 2011, 86(3):368-375.

Hatakka K, Savilahti E, Ponka A, Meurman JH, Poussa T, Nase L, Saxelin M, Korpela R: Effect of long term consumption of probiotic milk on infections in children attending day care centres: double blind, randomised trial. *BMJ* 2001, 322(7298):1327.

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

Heath AP, Bennett GN, Kavraki LE: Finding metabolic pathways using atom tracking. *Bioinformatics* 2010, 26(12):1548-1555.

Heavens D, Tailford LE, Crossman L, Jeffers F, MacKenzie DA, Caccamo M, Juge N: Genome Sequence of the Vertebrate Gut Symbiont *Lactobacillus reuteri* ATCC 53608. *J Bacteriol* 2011, 193(15):4015-4016.

Hebert EM, Saavedra L, Taranto MP, Mozzi F, Magni C, Nader ME, Font de Valdez G, Sesma F, Vignolo G, Raya RR: Genome sequence of the bacteriocin-producing *Lactobacillus curvatus* strain CRL705. *J Bacteriol* 2012, 194(2):538-539.

Heller KJ: Probiotic bacteria in fermented foods: product characteristics and starter organisms. *Am J Clin Nutr* 2001, 73(2):374s-379s.

Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J: De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res* 2008, 18(5):802-809.

Hojsak I, Abdovic S, Szajewska H, Milosevic M, Krznaric Z, Kolacek S: *Lactobacillus* GG in the prevention of nosocomial gastrointestinal and respiratory tract infections. *Pediatrics* 2010, 125(5):e1171-7.

Horvath P, Coûté-Monvoisin A, Romero DA, Boyaval P, Fremaux C, Barrangou R: Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* 2009, 131(1):62-70.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, *et al*: InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 2012, 40(D1):D306-D312.

Hurmalainen V, Edelman S, Antikainen J, Baumann M, Lahteenmaki K, Korhonen TK: Extracellular proteins of *Lactobacillus crispatus* enhance activation of human plasminogen. *Microbiology* 2007, 153(Pt 4):1112-1122.

Hyatt D, Chen G, LoCascio P, Land M, Larimer F, Hauser L: Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010, 11(1):119.

Ihaka R, Gentleman R: R: A language for data analysis and graphics. *J Comp Graph Stat* 1996, 5(3):299-314.

Ilie L, Fazayeli F, Ilie S: HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics* 2011, 27(3):295-302.

Iliev ID, Tohno M, Kurosaki D, Shimosato T, He F, Hosoda M, Saito T, Kitazawa H: Immunostimulatory Oligodeoxynucleotide Containing TTTCGTTT Motif from *Lactobacillus rhamnosus* GG DNA Potentially Suppresses OVA-specific IgE Production in Mice. *Scand J Immunol* 2008, 67(4):370-376.

Isolauri E, Juntunen M, Rautanen T, Sillanaukee P, Koivula T: A human *Lactobacillus* strain (*Lactobacillus casei* sp strain GG) promotes recovery from acute diarrhea in children. *Pediatrics* 1991, 88(1):90-97.

Jack RW, Tagg JR, Ray B: Bacteriocins of gram-positive bacteria. *Microbiol Rev* 1995, 59(2):171-200.

Jacobsen CN, Nielsen VR, Hayford A, Møller P, Michaelsen K, Paerregaard A, Sandström B, Tvede M, Jakobsen M: Screening of probiotic activities of forty-seven strains of *Lactobacillus* spp. by in vitro techniques and evaluation of the colonization ability of five selected strains in humans. *Appl Environ Microbiol* 1999, 65(11):4949-4956.

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD: Extending assembly of short DNA sequences to handle error. *Bioinformatics* 2007, 23(21):2942-2944.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P: eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 2008, 36(suppl 1):D250-D254.

Ji Y, Mitra R, Quintana F, Jara A, Mueller P, Liu P, Lu Y, Liang S: BM-BC: a Bayesian method of base calling for Solexa sequence data. *BMC Bioinformatics* 2012, 13 Suppl 13:S6.

Jiménez E, Langa S, Martín V, Arroyo R, Martín R, Fernández L, Rodríguez JM: Complete Genome Sequence of *Lactobacillus fermentum* CECT 5716, a Probiotic Strain Isolated from Human Milk. *J Bacteriol* 2010a, 192(18):4800-4800.

Jiménez E, Martín R, Maldonado A, Martín V, Gómez de Segura A, Fernández L, Rodríguez JM: Complete Genome Sequence of *Lactobacillus salivarius* CECT 5713, a Probiotic Strain Isolated from Human Milk and Infant Feces. *J Bacteriol* 2010b, 192(19):5266-5267.

Jiménez-Díaz R, Rios-Sánchez RM, Desmazeaud M, Ruiz-Barba JL, Piard JC: Plantaricins S and T, Two New Bacteriocins Produced by *Lactobacillus plantarum* LPCO10 Isolated from a Green Olive Fermentation. *Appl Environ Microbiol* 1993, 59(5):1416-1424.

Joss MJ, Koenig JE, Labbate M, Polz MF, Gillings MR, Stokes HW, Doolittle WF, Boucher Y: ACID: annotation of cassette and integron data. *BMC Bioinformatics* 2009, 10(1):118.

Juan D, Pazos F, Valencia A: High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A* 2008, 105(3):934-939.

Jünemann S, Prior K, Albersmeier A, Albaum S, Kalinowski J, Goesmann A, Stoye J, Harmsen D: GABenchToB: A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers. *PLoS ONE* 2014, 9(9):e107014.

Joerger MC, Klaenhammer TR: Cloning, expression, and nucleotide sequence of the *Lactobacillus helveticus* 481 gene encoding the bacteriocin helveticin J. *J Bacteriol* 1990, 172(11):6339-6347.

Kajander K, Hatakka K, Poussa T, Farkkila M, Korpela R: A probiotic mixture alleviates symptoms in irritable bowel syndrome patients: a controlled 6-month intervention. *Aliment Pharmacol Ther* 2005, 22(5):387-394.

Kajander K, Myllyluoma E, Rajilic-Stojanovic M, Kyronpalo S, Rasmussen M, Jarvenpää S, Zoetendal EG, de Vos WM, Vapaatalo H, Korpela R: Clinical trial: multispecies probiotic supplementation alleviates the symptoms of irritable bowel syndrome and stabilizes intestinal microbiota. *Aliment Pharmacol Ther* 2008, 27(1):48-57.

Kalliomäki M, Salminen S, Arvilommi H, Kero P, Koskinen P, Isolauri E: Probiotics in primary prevention of atopic disease: a randomised placebo-controlled trial. *Lancet* 2001, 357(9262):1076-1079.

Kalliomäki M, Salminen S, Poussa T, Arvilommi H, Isolauri E: Probiotics and prevention of atopic disease: 4-year follow-up of a randomised placebo-controlled trial. *Lancet* 2003, 361(9372):1869-1871.

Kalliomäki M, Salminen S, Poussa T, Isolauri E: Probiotics during the first 7 years of life: a cumulative risk reduction of eczema in a randomized, placebo-controlled trial. *J Allergy Clin Immunol* 2007, 119(4):1019-1021.

Kandler O, Weiss N: Genus Lactobacillus Beijerinck 1901, 212^{AL}. In *Bergey's manual of systematic bacteriology*. *Volume 2*. Edited by Williams ST, Sharpe ME, Holt JG, Williams and Wilkins: 1986:1209-1234.

Kang HJ, Paterson NG, Gaspar AH, Ton-That H, Baker EN: The *Corynebacterium diphtheriae* shaft pilin SpaA is built of tandem Ig-like modules with stabilizing isopeptide and disulfide bonds. *Proc Natl Acad Sci U S A* 2009, 106(40):16967-16971.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, 32(suppl 1):D277-D280.

Kant R, Blom J, Palva A, Siezen RJ, de Vos WM: Comparative genomics of *Lactobacillus*. *Microb Biotechnol* 2011a, 4(3):323-332.

Kant R, Paulin L, Alatalo E, de Vos WM, Palva A: Genome sequence of *Lactobacillus amylovorus* GRL1118, isolated from pig ileum. *J Bacteriol* 2011b, 193(12):3147-3148.

Kant R, Paulin L, Alatalo E, de Vos WM, Palva A: Genome sequence of *Lactobacillus amylovorus* GRL1112. *J Bacteriol* 2011c, 193(3):789-790.

Kant R, Rintahaka J, Yu X, Sigvart-Mattila P, Paulin L, Mecklin JP, Saarela M, Palva A, von Ossowski I: A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*. *PLoS One* 2014, 9(7):e102762.

Kao W, Chan AH, Song YS: ECHO: a reference-free short-read error correction algorithm. *Genome Res* 2011, 21(7):1181-1192.

Kao W, Stevens K, Song YS: BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res* 2009, 19(10):1884-1895.

Karp PD, Paley S, Romero P: The pathway tools software. *Bioinformatics* 2002, 18(suppl 1):S225-S232.

Kergourlay G, Messaoudi S, Dousset X, Prévost H: Genome Sequence of *Lactobacillus salivarius* SMXD51, a Potential Probiotic Strain Isolated from Chicken Cecum, Showing Anti-Campylobacter Activity. *J Bacteriol* 2012, 194(11):3008-3009.

Kim D, Choi S, Kim D, Kim RN, Nam S, Kang A, Kim A, Park H: Genome Sequence of *Lactobacillus versmoldensis* KCTC 3814. *J Bacteriol* 2011a, 193(19):5589-5590.

Kim D, Choi S, Kang A, Nam S, Kim D, Kim RN, Kim A, Park H: Draft Genome Sequence of *Lactobacillus zeae* KCTC 3804. *J Bacteriol* 2011b, 193(18):5023-5023.

Kim D, Choi S, Kang A, Nam S, Kim D, Kim RN, Kim A, Park H: Draft Genome Sequence of *Lactobacillus malefermentans* KCTC 3548. *J Bacteriol* 2011c, 193(19):5537-5537.

Kim D, Choi S, Kang A, Nam S, Kim D, Kim RN, Kim A, Park H: Draft Genome Sequence of *Lactobacillus mali* KCTC 3596. *J Bacteriol* 2011d, 193(18):5037-5037.

Kim JW, Rajagopal SN: Antimicrobial activities of *Lactobacillus crispatus* ATCC 33820 and *Lactobacillus gasseri* ATCC 33323. *J Microbiol* 2001, 39:146-148.

Kingsford CL, Ayanbule K, Salzberg SL: Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007, 8(2):R22.

Kircher M, Stenzel U, Kelso J: Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* 2009, 10(8):R83.

Kislyuk AO, Katz LS, Agrawal S, Hagen MS, Conley AB, Jayaraman P, Nelakuditi V, Humphrey JC, Sammons SA, Govil D*, et al*: A computational genomics pipeline for prokaryotic sequencing projects. *Bioinformatics* 2010, 26(15):1819-1826.

Klassen J, Currie C: Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. *BMC Genomics* 2012, 13(1):14.

Klassen JL, Currie CR: ORFcor: Identifying and Accommodating ORF Prediction Inconsistencies for Phylogenetic Analysis. *PLoS One* 2013, 8(3):e58387.

Kleerebezem M, Boekhorst J, van Kranenburg R, Molenaar D, Kuipers OP, Leer R, Tarchini R, Peters SA, Sandbrink HM, Fiers MW, *et al*: Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A* 2003, 100(4):1990-1995.

Kleerebezem M, Hols P, Bernard E, Rolain T, Zhou M, Siezen RJ, Bron PA: The extracellular biology of the lactobacilli. *FEMS Microbiol Rev* 2010, 34(2):199-230.

Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S, *et al*: The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res* 2009, 37(suppl 1):D216-D223.

Konstantinidis KT, Tiedje JM: Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 2005, 187(18):6258-6264.

Konstantinov SR, Smidt H, de Vos WM, Bruijns SCM, Singh SK, Valence F, Molle D, Lortal S, Altermann E, Klaenhammer TR, et al: S layer protein A of *Lactobacillus acidophilus* NCFM regulates immature dendritic cell and T cell functions. *Proc Natl Acad Sci U S A* 2008, 105(49):19474-19479.

Koonin EV, Galperin, MY: Sequence-Evolution-Function: Computational Approaches in Comparative Genomics, Kluwer Academic: 2003.

Koponen J, Laakso K, Koskenniemi K, Kankainen M, Savijoki K, Nyman TA, de Vos WM, Tynkkynen S, Kalkkinen N, Varmanen P: Effect of acid stress on protein expression and phosphorylation in *Lactobacillus rhamnosus* GG. *J Proteomics* 2012, 75(4):1357-1374.

Korbel JO, Snel B, Huynen MA, Bork P: SHOT: a web server for the construction of genome phylogenies. *Trends Genet* 2002, 18(3):158-162.

Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P: Systematic Association of Genes to Phenotypes by Genome and Literature Mining. *PLoS Biol* 2005, 3(5):e134.

Koskenniemi K, Laakso K, Koponen J, Kankainen M, Greco D, Auvinen P, Savijoki K, Nyman TA, Surakka A, Salusjarvi T, *et al*: Proteomics and transcriptomics characterization of bile stress response in probiotic *Lactobacillus rhamnosus* GG. *Mol Cell Proteomics* 2011, 10(2):M110.002741.

Koskinen P, Toronen P, Nokso-Koivisto J, Holm L: PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* 2015.

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2004, 32(suppl 1):D438-D442.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001, 305(3):567-580.

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

Krutmann J: Pre-and probiotics for human skin. *J Dermatol Sci* 2009, 54(1):1-5.

Kunin V, Ouzounis CA: Clustering the annotation space of proteins. *BMC Bioinformatics* 2005, 6(1):24.

Kurtz S, Schleiermacher C: REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* 1999, 15(5):426-427.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: Versatile and open software for comparing large genomes. *Genome Biol* 2004, 5(2):R12.

Laakso K, Koskenniemi K, Koponen J, Kankainen M, Surakka A, Salusjarvi T, Auvinen P, Savijoki K, Nyman TA, Kalkkinen N, *et al*: Growth phase-associated changes in the proteome and transcriptome of *Lactobacillus rhamnosus* GG in industrial-type whey medium. *Microb Biotechnol* 2011, 4(6):746-766.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H, Rognes T, Ussery DW: RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007, 35(9):3100-3108.

Langille MG, Brinkman FS: IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009, 25(5):664-665.

Langille MG, Hsiao WW, Brinkman FS: Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 2010, 8(5):373-382.

Langille M, Hsiao W, Brinkman F: Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 2008, 9(1):329.

Larsen TS, Krogh A: EasyGene–a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 2003, 4(1):21.

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

Lebeer S, Vanderleyden J, De Keersmaecker SC: Genes and molecules of lactobacilli supporting probiotic action. *Microbiol Mol Biol Rev* 2008, 72(4):728-764.

Lebeer S, Claes IJ, Verhoeven TL, Vanderleyden J, De Keersmaecker SC: Exopolysaccharides of *Lactobacillus rhamnosus* GG form a protective shield against innate immune factors in the intestine. *Microb Biotechnol* 2011, 4(3):368-374.

Lebeer S, Claes I, Tytgat HL, Verhoeven TL, Marien E, von Ossowski I, Reunanen J, Palva A, Vos WM, Keersmaecker SC, *et al*: Functional analysis of *Lactobacillus rhamnosus* GG pili in relation to adhesion and immunomodulatory interactions with intestinal epithelial cells. *Appl Environ Microbiol* 2012, 78(1):185-193.

Lebeer S, Verhoeven TL, Francius G, Schoofs G, Lambrichts I, Dufrene Y, Vanderleyden J, De Keersmaecker SC: 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* 2009, 75(11):3554-3563.

Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thevenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF: Orthology detection combining clustering and synteny for very large datasets. *PLoS One* 2014, 9(8):e105015.

Lee Y, Salminen S: The coming of age of probiotics. *Trends Food Sci Technol* 1995, 6(7):241-245.

Lee JH, Chae JP, Lee JY, Lim J, Kim G, Ham J, Chun J, Kang D: Genome Sequence of *Lactobacillus johnsonii* PF01, Isolated from Piglet Feces. *J Bacteriol* 2011a, 193(18):5030-5031.

Lee JH, Valeriano VD, Shin Y, Chae JP, Kim G, Ham J, Chun J, Kang D: Genome Sequence of *Lactobacillus mucosae* LM1, Isolated from Piglet Feces. *J Bacteriol* 2012, 194(17):4766-4766.

Lee S, Cho Y, Lee AH, Chun J, Ha N, Ko G: Genome Sequence of *Lactobacillus ruminis* SPM0211, Isolated from a Fecal Sample from a Healthy Korean. *J Bacteriol* 2011b, 193(18):5034-5034.

Leplae R, Hebrant A, Wodak SJ, Toussaint A: ACLAME: a CLAssification of Mobile genetic Elements. *Nucleic Acids Res* 2004, 32(suppl 1):D45-D49.

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

Li L, Stoeckert CJ, Roos DS: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003, 13(9):2178-2189.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, *et al*: De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010, 20(2):265-272.

Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, *et al*: HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* 2009, 37(suppl 1):D471-D478.

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

Lippi M, Passerini A, Punta M, Rost B, Frasconi P: MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* 2008, 24(18):2094-2095.

Liu B, Pop M: ARDB—antibiotic resistance genes database. *Nucleic Acids Res* 2009, 37(suppl 1):D443-D447.

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

Liu M, Siezen RJ, Nauta A: In Silico Prediction of Horizontal Gene Transfer Events in *Lactobacillus bulgaricus* and *Streptococcus thermophilus* Reveals Protocooperation in Yogurt Manufacturing. *Appl Environ Microbiol* 2009, 75(12):4120-4129.

Liu M, Nauta A, Francke C, Siezen RJ: Comparative Genomics of Enzymes in Flavor-Forming Pathways from Amino Acids in Lactic Acid Bacteria. *Appl Environ Microbiol* 2008, 74(15):4590-4600.

Liu S, Leathers TD, Copeland A, Chertkov O, Goodwin L, Mills DA: Complete Genome Sequence of *Lactobacillus buchneri* NRRL B-30929, a Novel Strain from a Commercial Ethanol Plant. *J Bacteriol* 2011, 193(15):4019-4020.

Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ: High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* 2012, 10(9):599-606.

Lowe TM, Eddy SR: tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, 25(5):0955-964.

Lukjancenko O, Ussery DW, Wassenaar TM: Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera. *Microb Ecol* 2012, 63(3):651-673.

Ma B, Forney LJ, Ravel J: Vaginal microbiome: rethinking health and disease. *Annu Rev Microbiol* 2012, 66:371-389.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* 2001, 29(22):4724-4735.

Macklaim JM, Gloor GB, Anukam KC, Cribby S, Reid G: At the crossroads of vaginal health and disease, the genome sequence of *Lactobacillus iners* AB-1. *Proc Natl Acad Sci U S A* 2011, 108 Suppl 1:4688-4695.

Madigan MT, Martinko JM, Parker J, Brock TD: *Brock biology of microorganisms*. 13th edition. Benjamin Cummings. 2010.

Madupu R, Richter A, Dodson RJ, Brinkac L, Harkins D, Durkin S, Shrivastava S, Sutton G, Haft D: CharProtDB: a database of experimentally characterized protein annotations. *Nucleic Acids Res* 2012, 40(D1):D237-D241.

Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL: GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 2013, 29(14):1718-1725.

Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, *et al*: Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A* 2006, 103(42):15611-15616.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, *et al*: CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2011, 39(suppl 1):D225-D229.

Marcotte EM, Pellegrini M, Ng H, Rice DW, Yeates TO, Eisenberg D: Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999, 285(5428):751-753.

Mardis ER: Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008, 9:387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005, 437(7057):376-380.

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE: Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. *Appl Environ Microbiol* 2011, 77(22):8071-8079.

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al*: IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res* 2012, 40(D1):D115-D122.

Marraffini LA, Sontheimer EJ: CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* 2010, 11(3):181-190.

Martin DH: The microbiota of the vagina and its influence on women's health and disease. *Am J Med Sci* 2012, 343(1):2.

Martin DM, Berriman M, Barton GJ: GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* 2004, 5(1):178.

Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011, 17(1):pp. 10-12.

Martín R, Olivares M, Marín ML, Fernández L, Xaus J, Rodríguez JM: Probiotic potential of 3 lactobacilli strains isolated from breast milk. *J Hum Lact* 2005, 21(1):8-17.

Massingham T, Goldman N: All Your Base: a fast and accurate probabilistic approach to base calling. *Genome Biol* 2012, 13(2):R13.

Matsuda K, Tsuji H, Asahara T, Matsumoto K, Takada T, Nomoto K: Establishment of an analytical system for the human fecal microbiota, based on reverse transcription-quantitative PCR targeting of multicopy rRNA molecules. *Appl Environ Microbiol* 2009, 75(7):1961-1969.

Mavromatis K, Ivanova NN, Chen IA, Szeto E, Markowitz VM, Kyrpides NC: The DOE-JGI Standard operating procedure for the annotations of microbial genomes. *Stand Genomic Sci* 2009, 1(1):63.

Maxam AM, Gilbert W: A new method for sequencing DNA. *Proc Natl Acad Sci U S A* 1977, 74:560-564.

McHardy AC, Goesmann A, Pühler A, Meyer F: Development of joint application strategies for two microbial gene finders. *Bioinformatics* 2004, 20(10):1622-1631.

Mc Grath S, Fitzgerald GF, van Sinderen D: Bacteriophages in dairy products: pros and cons. *Biotechnol J* 2007, 2(4):450-455.

McNulty NP, Yatsunenko T, Hsiao A, Faith JJ, Muegge BD, Goodman AL, Henrissat B, Oozeer R, Cools-Portier S, Gobert G, *et al*: The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci Transl Med* 2011, 3(106):106ra106.

Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R: antiSMASH: rapid identification, annotation and

analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 2011, 39(suppl 2):W339-W346.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R: The microbial pan-genome. *Curr Opin Genet Dev* 2005, 15(6):589-594.

Meijerink M, van Hemert S, Taverne N, Wels M, de Vos P, Bron PA, Savelkoul HF, van Bilsen J, Kleerebezem M, Wells JM: Identification of genetic loci in *Lactobacillus plantarum* that modulate the immune response of dendritic cells using comparative genome hybridization. *PLoS One* 2010, 5(5):e10632.

Miettinen M, Pietilä TE, Kekkonen RA, Kankainen M, Latvala S, Pirhonen J, Osterlund P, Korpela R, Julkunen I: Nonpathogenic *Lactobacillus rhamnosus* activates the inflammasome and antiviral responses in human macrophages. *Gut Microbes* 2012, 3(6):510-522.

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

Mills S, Stanton C, Hill C, Ross RP: New developments and applications of bacteriocins and peptides in foods. *Annu Rev Food Sci Technol* 2011, 2:299-329.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV: Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 2003, 3:2.

Mohammed Y, Lee B, Kang Z, Du G: Capability of *Lactobacillus reuteri* to Produce an Active Form of Vitamin B12 under Optimized Fermentation Conditions. *Journal of Academia and Industrial Research (JAIR)* 2014, 2(11):617.

Molin G, Jeppsson B, Johansson M, Ahrne S, Nobaek S, Ståhl M, Bengmark S: Numerical taxonomy of *Lactobacillus* spp. associated with healthy and diseased mucosa of the human intestines. *J Appl Microbiol* 1993, 74(3):314-323.

Morgat A, Coissac E, Coudert E, Axelsen KB, Keller G, Bairoch A, Bridge A, Bougueleret L, Xenarios I, Viari A: UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* 2012, 40(D1):D761-D769.

Morita H, Toh H, Fukuda S, Horikawa H, Oshima K, Suzuki T, Murakami M, Hisamatsu S, Kato Y, Takizawa T, *et al*: Comparative Genome Analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* Reveal a Genomic Island for Reuterin and Cobalamin Production. *DNA Res* 2008, 15(3):151-161.

Morita H, Toh H, Oshima K, Murakami M, Taylor TD, Igimi S, Hattori M: Complete Genome Sequence of the Probiotic *Lactobacillus rhamnosus* ATCC 53103. *J Bacteriol* 2009, 191(24):7630-7631.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 2007, 35(suppl 2):W182-W185.

Moura A, Soares M, Pereira C, Leitão N, Henriques I, Correia A: INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* 2009, 25(8):1096-1098.

Munoz-Merida A, Viguera E, Claros MG, Trelles O, Perez-Pulido AJ: Sma3s: a three-step modular annotator for large sequence datasets. *DNA Res* 2014, 21(4):341-353.

Mayra-Makinen A, Bigret M: Industrial use and production of lactic acid bacteria. In *Lactic Acid Bacteria – Microbiology and Functional Aspects.* Edited by Salminen S, Wright AV. Marcel Dekker Inc: 1998: 73–102.

Nagarajan N, Read TD, Pop M: Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* 2008, 24(10):1229-1235.

Nam S, Choi S, Kang A, Kim D, Kim D, Kim RN, Kim A, Park H: Genome Sequence of *Lactobacillus coryniformis* subsp. *coryniformis* KCTC 3167. *J Bacteriol* 2011a, 193(4):1014-1015.

Nam S, Choi S, Kang A, Kim D, Kim RN, Kim A, Kim D, Park H: Genome Sequence of *Lactobacillus farciminis* KCTC 3681. *J Bacteriol* 2011b, 193(7):1790-1791.

Nam S, Choi S, Kang A, Kim D, Kim RN, Kim A, Kim D, Park H: Genome Sequence of *Lactobacillus animalis* KCTC 3501. *J Bacteriol* 2011c, 193(5):1280-1281.

Nam S, Choi S, Kang A, Kim D, Kim RN, Kim D, Kim A, Park H: Genome Sequence of *Lactobacillus suebicus* KCTC 3549. *J Bacteriol* 2011d, 193(19):5532-5533.

Nam S, Choi S, Kang A, Lee KS, Kim D, Kim RN, Kim D, Park H: Genome Sequence of *Lactobacillus fructivorans* KCTC 3543. *J Bacteriol* 2012, 194(8):2111-2112.

Nawrocki EP, Kolbe DL, Eddy SR: Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009, 25(10):1335-1337.

Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, *et al*: A catalog of reference genomes from the human microbiome. *Science* 2010, 328(5981):994-999.

Nes IF, Johnsborg O: Exploration of antimicrobial potential in LAB by genomics. *Curr Opin Biotechnol* 2004, 15(2):100-104.

Nijkamp J, Winterbach W, van den Broek M, Daran J, Reinders M, de Ridder D: Integrating genome assemblies with MAIA. *Bioinformatics* 2010, 26(18):i433-i439.

Nilsen T, Nes IF, Holo H: Enterolysin A, a cell wall-degrading bacteriocin from *Enterococcus faecalis* LMG 2333. *Appl Environ Microbiol* 2003, 69(5):2975-2984.

Nusbaum C, Ohsumi TK, Gomez J, Aquadro J, Victor TC, Warren RM, Hung DT, Birren BW, Lander ES, Jaffe DB: Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. *Nat Methods* 2008, 6(1):67-69.

Nyquist OL, McLeod A, Brede DA, Snipen L, Aakra Å, Nes IF: Comparative genomics of *Lactobacillus sakei* with emphasis on strains from meat. *Mol Genet Genomics* 2011, 285(4):297-311.

Oberhardt MA, Palsson BØ, Papin JA: Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009, 5(1).

Oelschlaeger TA: Mechanisms of probiotic actions–a review. Int *J Med Microbiol* 2010, 300(1):57-62.

Ofran Y, Mysore V, Rost B: Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007, 23(13):i347-i353.

Ofran Y, Punta M, Schneider R, Rost B: Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov Today* 2005, 10(21):1475-1482.

Oh S, Roh H, Ko H, Kim S, Kim KH, Lee SE, Chang IS, Kim S, Choi I: Complete Genome Sequencing of *Lactobacillus acidophilus* 30SC, Isolated from Swine Intestine. *J Bacteriol* 2011, 193(11):2882-2883.

Oksaharju A, Kankainen M, Kekkonen RA, Lindstedt KA, Kovanen PT, Korpela R, Miettinen M: Probiotic *Lactobacillus rhamnosus* downregulates FCER1 and HRH4 expression in human mast cells. *World J Gastroenterol* 2011, 17(6):750-759.

O'Sullivan O, O'Callaghan J, Sangrador-Vegas A, McAuliffe O, Slattery L, Kaleta P, Callanan M, Fitzgerald G, Ross RP, Beresford T: Comparative genomics of lactic acid bacteria reveals a niche-specific gene set. *BMC microbiol* 2009, 9(1):50.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H, Cohoon M, de Crécy-Lagard V, Diaz N, Disz T, Edwards R, *et al*: The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. *Nucleic Acids Res* 2005, 33(17):5691-5702.

Overbeek R, Larsen N, Walunas T, D'Souza M, Pusch G, Selkov E, Liolios K, Joukov V, Kaznadzey D, Anderson I, *et al*: The ERGO™ genome analysis and discovery system. *Nucleic Acids Res* 2003, 31(1):164-171.

Patel RK, Jain M: NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012, 7(2):e30619.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E: Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 2008, 18(11):1814-1828.

Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC: GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 2010, 7(6):455-457.

Péant B, LaPointe G, Gilbert C, Atlan D, Ward P, Roy D: Comparative analysis of the exopolysaccharide biosynthesis gene clusters from four strains of *Lactobacillus rhamnosus*. *Microbiology* 2005, 151(6):1839-1851.

Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 1988, 85(8):2444-2448.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci U S A* 1999, 96(8):4285-4288.

Peng Y, Leung HCM, Yiu SM, Chin FYL: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012, 28(11):1420-1428.

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

Pevzner PA, Tang H, Waterman MS: An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 2001, 98(17):9748-9753.

Pfeiler EA, Klaenhammer TR: The genomics of lactic acid bacteria. *Trends Microbiol* 2007, 15(12):546-553.

Pichon C, Felden B: Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics* 2008, 24(24):2807-2813.

Pitcher DG, Saunders NA, Owen RJ: Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Lett Appl Microbiol* 1989, 8(4):151-156.

Pittet V, Ewen E, Bushell BR, Ziola B: Genome Sequence of *Lactobacillus rhamnosus* ATCC 8530. *J Bacteriol* 2012, 194(3):726-726.

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

Pop M, Kosack DS, Salzberg SL: Hierarchical scaffolding with Bambus. *Genome Res* 2004a, 14(1):149-159.

Pop M, Phillippy A, Delcher AL, Salzberg SL: Comparative genome assembly. *Brief Bioinform* 2004b, 5(3):237-248.

Pot B, Ludwig W, Kersters K, Schleifer K: Taxonomy of lactic acid bacteria. In *Bacteriocins of lactic acid bacteria.* Edited by De Vuyst L, Vandamme EJ. Chapman and Hall; 1994:13-90.

Prajapati JB, Khedkar CD, Chitra J, Suja S, Mishra V, Sreeja V, Patel RK, Ahir VB, Bhatt VD, Sajnani MR, *et al*: Whole-Genome Shotgun Sequencing of *Lactobacillus rhamnosus* MTCC 5462, a Strain with Probiotic Potential. *J Bacteriol* 2012, 194(5):1264-1265.

Prajapati JB, Khedkar CD, Chitra J, Suja S, Mishra V, Sreeja V, Patel RK, Ahir VB, Bhatt VD, Sajnani MR, *et al*: Whole-Genome Shotgun Sequencing of an Indian-Origin *Lactobacillus helveticus* Strain, MTCC 5463, with Probiotic Potential. *J Bacteriol* 2011, 193(16):4282-4283.

Price AL, Jones NC, Pevzner PA: De novo identification of repeat families in large genomes. *Bioinformatics* 2005, 21(suppl 1):i351-i358.

Pridmore RD, Berger B, Desiere F, Vilanova D, Barretto C, Pittet A, Zwahlen M, Rouvet M, Altermann E, Barrangou R, *et al*: The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc Natl Acad Sci U S A* 2004, 101(8):2512-2517.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 1987, 238(4825):336-341.

Punta M, Ofran Y: The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol* 2008, 4(10):e1000160.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: The Pfam protein families database. *Nucleic Acids Res* 2012, 40(D1):D290-D301.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012, 13(1):341.

Quinlan AR, Stewart DA, Strömberg MP, Marth GT: Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 2008, 5(2):179-181.

Raes J, Harrington ED, Singh AH, Bork P: Protein function space: viewing the limits or limited by our view? *Curr Opin Struct Biol* 2007, 17(3):362-369.

Raftis EJ, Salvetti E, Torriani S, Felis GE, O'Toole PW: Genomic Diversity of *Lactobacillus salivarius*. *Appl Environ Microbiol* 2011, 77(3):954-965.

Rahman O, Cummings SP, Harrington DJ, Sutcliffe IC: Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria. *World J Microbiol Biotech* 2008, 24(11):2377-2382.

Ravcheev DA, Best AA, Sernova NV, Kazanov MD, Novichkov PS, Rodionov DA: Genomic reconstruction of transcriptional regulatory networks in lactic acid bacteria. *BMC Genomics* 2013, 14(1):94.

Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO*, et al*: Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci U S A* 2011, 108 Suppl 1:4680-4687.

Rawlings ND, Tolle DP, Barrett AJ: MEROPS: the peptidase database. *Nucleic Acids Res* 2004, 32(suppl 1):D160-D164.

Regalia M, Rosenblad MA, Samuelsson T: Prediction of signal recognition particle RNA genes. *Nucleic Acids Res* 2002, 30(15):3368-3377.

Remm M, Storm CE, Sonnhammer EL: Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 2001, 314(5):1041-1052.

Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X: CSS-Palm 2.0: an updated software for palmitoylation sites prediction. *Protein Eng Des Sel* 2008, 21(11):639-644.

Rentzsch R, Orengo CA: Protein function prediction–the power of multiplicity. *Trends Biotechnol* 2009, 27(4):210-219.

Reunanen J, von Ossowski I, Hendrickx AP, Palva A, de Vos WM: Characterization of the SpaCBA pilus fibers in the probiotic *Lactobacillus rhamnosus* GG. *Appl Environ Microbiol* 2012, 78(7):2337-2344.

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

Rho M, Wu YW, Tang H, Doak TG, Ye Y: Diverse CRISPRs evolving in human microbiomes. *PLoS Genet* 2012, 8(6):e1002441.

Richardson EJ, Watson M: The automatic annotation of bacterial genomes. *Brief Bioinform* 2013, 14(1):1-12.

Richter DC, Schuster SC, Huson DH: OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics* 2007, 23(13):1573-1579.

Riley MA, Wertz JE: Bacteriocins: evolution, ecology, and application. *Annu Rev Microbiol* 2002, 56:117-137.

Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT: Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 2009, 25(16):2071-2073.

Rivas E, Eddy SR: Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001, 2(1):8.

Rivera-Espinoza Y, Gallardo-Navarro Y: Non-dairy probiotic products. *Food Microbiol* 2010, 27(1):1-11.

Roberts RJ, Vincze T, Posfai J, Macelis D: REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* 2010, 38(suppl 1):D234-D236.

Rokas A, Williams BL, King N, Carroll SB: Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 2003, 425(6960):798-804.

Rost B: Enzyme function less conserved than anticipated. *J Mol Biol* 2002, 318(2):595-608.

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

Roth AC, Gonnet GH, Dessimoz C: Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 2008, 9(1):518.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, *et al*: An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011, 475(7356):348-352.

Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F: Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 2008, 9(1):431.

Ryan KA, Jayaraman T, Daly P, Canchaya C, Curran S, Fang F, Quigley EM, O'Toole PW: Isolation of lactobacilli with probiotic properties from the human stomach. *Lett Appl Microbiol* 2008, 47(4):269-274.

Sachdeva G, Kumar K, Jain P, Ramachandran S: SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* 2005, 21(4):483-491.

Saier MH, Tran CV, Barabote RD: TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 2006, 34(suppl 1):D181-D186.

Salichos L, Rokas A: Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* 2011, 6(4):e18755.

Salmela L: Correction of sequencing errors in a mixed set of reads. *Bioinformatics* 2010, 26(10):1284-1290.

Salvatore S, Salvatore S, Cattoni E, Siesto G, Serati M, Sorice P, Torella M: Urinary tract infections in women. *Eur J Obstet Gynecol Reprod Biol* 2011, 156(2):131-136.

Salvetti E, Torriani S, Felis G: The Genus *Lactobacillus*: A Taxonomic Update. *Probiotics Antimicrob Proteins* 2012, 4(4):217-226.

Samad A, Huff EF, Cai W, Schwartz DC: Optical mapping: a novel, single-molecule approach to genomic analysis. *Genome Research* 1995, 5(1):1-4.

Sanchez B, Schmitter JM, Urdaci MC: Identification of novel proteins secreted by *Lactobacillus rhamnosus* GG grown in de Mann-Rogosa-Sharpe broth. *Lett Appl Microbiol* 2009, 48(5):618-622.

Sanderson MJ, Purvis A, Henze C: Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol Evol* 1998, 13(3):105-109.

Sanger F, Coulson AR: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol* 1975, 94:441-448.

Salvetti E, Fondi M, Fani R, Torriani S, Felis GE: Evolution of lactic acid bacteria in the order *Lactobacillales* as depicted by analysis of glycolysis and pentose phosphate pathways. *Syst Appl Microbiol* 2013, 36(5):291-305.

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

Santos F, Vera JL, Lamosa P, de Valdez GF, de Vos WM, Santos H, Sesma F, Hugenholtz J: Pseudovitamin is the corrinoid produced by *Lactobacillus reuteri* CRL1098 under anaerobic conditions. *FEBS Lett* 2007, 581(25):4865-4870.

Saulnier DM, Santos F, Roos S, Mistretta TA, Spinler JK, Molenaar D, Teusink B, Versalovic J: Exploring metabolic pathway reconstruction and genome-wide expression profiling in *Lactobacillus reuteri* to define functional probiotic features. *PLoS One* 2011, 6(4):e18783.

Saxelin M, Myllyluoma E, Korpela R: Developing LGG®Extra, a Probiotic Multispecies Combination. In *Probiotics and health claims*. Edited by Kneifel W, Salminen S. Wiley-Blackwell; 2011:249-262

Saxelin M, Lassig A, Karjalainen H, Tynkkynen S, Surakka A, Vapaatalo H, Jarvenpää S, Korpela R, Mutanen M, Hatakka K: Persistence of probiotic strains in the gastrointestinal tract when administered as capsules, yoghurt, or cheese. *Int J Food Microbiol* 2010, 144(2):293-300.

Saxelin M, Tynkkynen S, Mattila-Sandholm T, de Vos WM: Probiotic and other functional microbes: from markets to mechanisms. *Curr Opin Biotechnol* 2005, 16(2):204-211.

Saxena RK, Anand P, Saran S, Isar J: Microbial production of 1,3-propanediol: Recent developments and emerging opportunities. *Biotechnol Adv* 2009, 27(6):895-913.

Scaria J, Chandramouli U, Verma SK: Antibiotic Resistance Genes Online (ARGO): A Database on vancomycin and βlactam resistance genes. *Bioinformation* 2005, 1(1):5.

Schmieder R, Edwards R: Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011, 27(6):863-864.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC: Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 2009, 5(12):e1000605.

Scott JR, Zähner D: Pili with strong attachments: Gram-positive bacteria do it differently. *Mol Microbiol* 2006, 62(2):320-330.

Seepersaud R, Hanniffy SB, Mayne P, Sizer P, Le Page R, Wells JM: Characterization of a Novel Leucine-Rich Repeat Protein Antigen from Group B *Streptococci* That Elicits Protective Immunity. *Infect Immun* 2005, 73(3):1671-1683.

Segers M, Lebeer S: Towards a better understanding of *Lactobacillus rhamnosus* GG - host interactions. *Microb Cell Fact* 2014, 13:S7.

Sevin EW, Barloy-Hubler F: RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol* 2007, 8(8):R155.

Shao Y, Harrison EM, Bi D, Tai C, He X, Ou H, Rajakumar K, Deng Z: TADB: a web-based resource for Type 2 toxin–antitoxin loci in bacteria and archaea. *Nucleic Acids Res* 2011, 39(suppl 1):D606-D611.

Sheikh MA, Erlich Y: Base-Calling for Bioinformaticians. In *Bioinformatics for High Throughput Sequencing.* Edited by Rodriguez-Ezpeleta N, Hackenberg M, Aransay AM. Springer; 2012:67-83.

Shendure J, Ji H: Next-generation DNA sequencing. *Nat Biotechnol* 2008, 26(10):1135-1145.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005, 309(5741):1728-1732.

Shipitsyna E, Roos A, Datcu R, Hallen A, Fredlund H, Jensen JS, Engstrand L, Unemo M: Composition of the vaginal microbiota in women of reproductive age-sensitive and specific molecular diagnosis of bacterial vaginosis is possible? *PLoS One* 2013, 8(4):e60670.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, *et al*: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011, 7:539.

Siezen RJ, van Hylckama Vlieg JET: Genomic diversity and versatility of *Lactobacillus plantarum*, a natural metabolic engineer. *Microb Cell Fact* 2011, 10.

Siezen RJ, Wilson G: Probiotics genomics. *Microb Biotechnol* 2010, 3(1):1-9.

Siezen RJ, Francke C, Renckens B, Boekhorst J, Wels M, Kleerebezem M, van Hijum SA: Complete resequencing and reannotation of the *Lactobacillus plantarum* WCFS1 genome. *J Bacteriol* 2012, 194(1):195-196.

Siezen RJ, Tzeneva VA, Castioni A, Wels M, Phan HT, Rademaker JL, Starrenburg MJ, Kleerebezem M, Molenaar D, van Hylckama Vlieg JE: Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches. *Environ Microbiol* 2010, 12(3):758-773.

Siguier P, Pérochon J, Lestrade L, Mahillon J, Chandler M: ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 2006, 34(suppl 1):D32-D36.

Silva M, Jacobus NV, Deneke C, Gorbach SL: Antimicrobial substance from a human *Lactobacillus* strain. *Antimicrobial Agents and Chemotherapy* 1987, 31(8):1231-1233.

Simpson JT, Durbin R: Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 2012, 22(3):549-556.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol İ: ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009, 19(6):1117-1123.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE: Fluorescence detection in automated DNA sequence analysis. *Nature* 1986, 321(6071):674-679.

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

Snel B, Bork P, Huynen MA: Genome phylogeny based on gene content. *Nat Genet* 1999, 21(1):108-110.

Sommer D, Delcher A, Salzberg S, Pop M: Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* 2007, 8(1):64.

Sonnhammer EL, Koonin EV: Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 2002, 18(12):619-620.

Srinivasan S, Hoffman NG, Morgan MT, Matsen FA, Fiedler TL, Hall RW, Ross FJ, McCoy CO, Bumgarner R, Marrazzo JM, et al: Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One* 2012, 7(6):e37818.

Srividhya K, Alaguraj V, Poornima G, Kumar D, Singh G, Raghavenderan L, Katta AM, Mehta P, Krishnaswamy S: Identification of prophages in bacterial genomes by dinucleotide relative abundance difference. *PLoS One* 2007, 2(11):e1193.

Staden R, Beal K, Bonfield J: The Staden Package, 1998. In *Bioinformatics methods and protocols. Volume 132.* Edited by Misener S, Krawetz S. Springer; 1999:115-130.

Steele J, Broadbent J, Kok J: Perspectives on the contribution of lactic acid bacteria to cheese flavor development. *Curr Opin Biotechnol* 2012, 24(2):135-141.

Storm CE, Sonnhammer EL: Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002, 18(1):92-99.

Sun Z, Chen X, Wang J, Zhao W, Shao Y, Guo Z, Zhang X, Zhou Z, Sun T, Wang L, *et al*: Complete Genome Sequence of *Lactobacillus delbrueckii* subsp. *bulgaricus* Strain ND02. *J Bacteriol* 2011, 193(13):3426-3427.

Suomalainen TH, Mäyrä-Mäkinen AM: Propionic acid bacteria as protective cultures in fermented milks and breads. *Le Lait* 1999, 79(1):165-174.

Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD: A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc* 2012, 7(7):1260-1284.

Szajewska H, Mrukowicz JZ: Probiotics in the treatment and prevention of acute infectious diarrhea in infants and children: a systematic review of published randomized, double-blind, placebo-controlled trials. *J Pediatr Gastroenterol Nutr* 2001, 33 Suppl 2:S17-25.

Tahara T, Kanatani K: Isolation and partial characterization of crispacin A, a cell-associated bacteriocin produced by *Lactobacillus crispatus* JCM 2009. *FEMS Microbiol Lett* 1997, 147:287-290.

Tanenbaum DM, Goll J, Murphy S, Kumar P, Zafar N, Thiagarajan M, Madupu R, Davidsen T, Kagan L, Kravitz S*, et al*: The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand Genomic Sci* 2010, 2(2):229-237.

Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, 278(5338):631-637.

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr Biol* 1996, 6(3):279-291.

Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A*, et al*: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 2003, 4(1):41.

Tech M, Merkl R: YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol* 2003, 3(4):441-451.

Telford JL, Barocchi MA, Margarit I, Rappuoli R, Grandi G: Pili in gram-positive pathogens. *Nat Rev Microbiol* 2006, 4(7):509-519.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, *et al*: Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* 2005, 102(39):13950-13955.

Tillier ERM, Collins RA: Genome rearrangement by replication-directed translocation. *Nat Genet* 2000, 26(2):195-197.

Tompkins TA, Barreau G, de Carvalho VG: Draft genome sequence of probiotic strain *Lactobacillus rhamnosus* R0011. *J Bacteriol* 2012, 194(4):902-902.

Ton-That H, Schneewind O: Assembly of pili on the surface of *Corynebacterium diphtheriae. Mol Microbiol* 2003, 50(4):1429-1438.

Torto-Alalibo T, Collmer C, Gwinn-Giglio M: The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. *BMC microbiol* 2009, 9(Suppl 1):S1.

Touzain F, Petit M, Schbath S, El Karoui M: DNA motifs that sculpt the bacterial chromosome. *Nat Rev Microbiol* 2010, 9(1):15-26.

Trachana K, Larsson TA, Powell S, Chen W, Doerks T, Muller J, Bork P: Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 2011, 33(10):769-780.

Tu Q, Ding D: Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol Lett* 2003, 221(2):269-275.

Tuomola EM, Ouwehand AC, Salminen SJ: Chemical, physical and enzymatic pre-treatments of probiotic lactobacilli alter their adhesion to human intestinal mucus glycoproteins. *Int J Food Microbiol* 2000, 60(1):75-81.

Uehara S, Monden K, Nomoto K, Seno Y, Kariyama R, Kumon H: A pilot study evaluating the safety and effectiveness of *Lactobacillus* vaginal suppositories in patients with recurrent urinary tract infection. *Int J Antimicrob Agents* 2006, 28 Suppl 1:S30-4.

Ulrich LE, Zhulin IB: The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res* 2010, 38(suppl 1):D401-D407.

Valencia A: Automatic annotation of protein function. *Curr Opin Struct Biol* 2005, 15(3):267-274.

van de Guchte M, Penaud S, Grimaldi C, Barbe V, Bryson K, Nicolas P, Robert C, Oztas S, Mangenot S, Couloux A, *et al*: The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution. *Proc Natl Acad Sci U S A* 2006, 103(24):9274-9279.

van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP: BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res* 2013, 41(Web Server issue):W448-53.

van Hemert S, Meijerink M, Molenaar D, Bron P, de Vos P, Kleerebezem M, Wells J, Marco M: Identification of *Lactobacillus plantarum* genes modulating the cytokine

response of human peripheral blood mononuclear cells. *BMC microbiol* 2010, 10(1):293.

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

Varnam A: *Lactobacillus*: occurrence and significance in non-dairy foods. *Microbiol Today* 2002, 29:13-17.

Vaughan EE, Heilig HG, Ben‐Amor K, Vos WM: Diversity, vitality and activities of intestinal lactic acid bacteria and bifidobacteria assessed by molecular approaches. *FEMS Microbiol Rev* 2005, 29(3):477-490.

Vélez MP, Petrova MI, Lebeer S, Verhoeven TL, Claes I, Lambrichts I, Tynkkynen S, Vanderleyden J, De Keersmaecker SC: Characterization of MabA, a modulator of *Lactobacillus rhamnosus* GG adhesion and biofilm formation. *FEMS Immunol Med Microbiol* 2010, 59(3):386-398.

Vélez MP, Verhoeven TLA, Draing C, Von Aulock S, Pfitzenmaier M, Geyer A, Lambrichts I, Grangette C, Pot B, Vanderleyden J, et al: Functional Analysis of d-Alanylation of Lipoteichoic Acid in the Probiotic Strain *Lactobacillus rhamnosus* GG. *Appl Environ Microbiol* 2007, 73(11):3595-3604.

Ventura M, O'Flaherty S, Claesson MJ, Turroni F, Klaenhammer TR, van Sinderen D, O'Toole PW: Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nat Rev Microbiol* 2008, 7(1):61-71.

Ventura M, Canchaya C, Bernini V, Altermann E, Barrangou R, McGrath S, Claesson MJ, Li Y, Leahy S, Walker CD, et al: Comparative Genomics and Transcriptional Analysis of Prophages Identified in the Genomes of *Lactobacillus gasseri*, *Lactobacillus salivarius*, and *Lactobacillus casei*. *Appl Environ Microbiol* 2006, 72(5):3130-3146.

Vernikos GS, Parkhill J: Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 2006, 22(18):2196-2203.

Vinayagam A, del Val C, Schubert F, Eils R, Glatting K, Suhai S, König R: GOPET: a tool for automated predictions of Gene Ontology terms. *BMC Bioinformatics* 2006, 7(1):161.

Voelkerding KV, Dames SA, Durtschi JD: Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009, 55(4):641-658.

Vogel RF, Pavlovic M, Ehrmann MA, Wiezer A, Liesegang H, Offschanka S, Voget S, Angelov A, Bocker G, Liebl W: Genomic analysis reveals *Lactobacillus sanfranciscensis* as stable element in traditional sourdoughs. *Microb Cell Fact* 2011, 10 Suppl 1:S6.

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005, 33(suppl 1):D433-D437.

von Ossowski I, Reunanen J, Satokari R, Vesterlund S, Kankainen M, Huhtinen H, Tynkkynen S, Salminen S, de Vos WM, Palva A: Mucosal Adhesion Properties of the Probiotic *Lactobacillus rhamnosus* GG SpaCBA and SpaFED Pilin Subunits. *Appl Environ Microbiol* 2010, 76(7):2049-2057.

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

Vuyst L, Degeest B: Heteropolysaccharides from lactic acid bacteria. *FEMS Microbiol Rev* 1999, 23(2):153-177.

Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R: Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006, 7(1):142.

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

Wagner I, Volkmer M, Sharan M, Villaveces JM, Oswald F, Surendranath V, Habermann BH: morFeus: a web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. *BMC Bioinformatics* 2014, 15:263.

Wall D, Fraser H, Hirsh A: Detecting putative orthologs. *Bioinformatics* 2003, 19(13):1710-1711.

Walter J: Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research. *Appl Environ Microbiol* 2008, 74(16):4985-4996.

Walter J: The microecology of lactobacilli in the gastrointestinal tract. In *Probiotics and Prebiotics: Scientific Aspects.* Edited by Tannock GW. Horizon Scientific Press; 2005:51-82.

Wang Y, Wang J, Ahmed Z, Bai X, Wang J: Complete Genome Sequence of *Lactobacillus kefiranofaciens* ZW3. *J Bacteriol* 2011a, 193(16):4280-4281.

Wang Y, Chen C, Ai L, Zhou F, Zhou Z, Wang L, Zhang H, Chen W, Guo B: Complete Genome Sequence of the Probiotic *Lactobacillus plantarum* ST-III. *J Bacteriol* 2011b, 193(1):313-314.

Warren AS, Archuleta J, Feng W, Setubal JC: Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* 2010, 11(1):131.

Warren RL, Sutton GG, Jones SJ, Holt RA: Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 2007, 23(4):500-501.

Washietl S, Hofacker IL, Stadler PF: Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* 2005, 102(7):2454-2459.

Wass MN, Sternberg MJ: ConFunc—functional annotation in the twilight zone. *Bioinformatics* 2008, 24(6):798-806.

Wassenaar TM, Binnewies TT, Hallin PF, Ussery DW: Tools for Comparison of Bacterial Genomes. In *Handbook of Hydrocarbon and Lipid Microbiology.* Edited by Timmis K. Springer; 2010:4313-4327.

Webb EC: *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes:* San Diego: Academic Press; 1992.

Wegmann U, Overweg K, Horn N, Goesmann A, Narbad A, Gasson MJ, Shearman C: Complete Genome Sequence of *Lactobacillus johnsonii* FI9785, a Competitive Exclusion Agent against Pathogens in Poultry. *J Bacteriol* 2009, 191(22):7142-7143.

Williams TA, Sarah EH: An introduction to phylogenetics and the tree of life. *Meth Microbiol* 2014, 41:2-6.

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA: DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* 2008, 36(suppl 1):D88-D92.

Woese CR: Bacterial evolution. *Microbiol Rev* 1987, 51(2):221-271

Wu H, Irizarry RA, Bravo HC: Intensity normalization improves color calling in SOLiD sequencing. *Nat Methods* 2010, 7(5):336-337.

Wu R, Kaiser AD: Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol* 1968, 35:523-537.

Wu R, Taylor E: Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol* 1971, 57:491-511.

Yan F, Cao H, Cover TL, Whitehead R, Washington MK, Polk DB: Soluble proteins produced by probiotic bacteria regulate intestinal epithelial cell survival and growth. *Gastroenterology* 2007, 132(2):562-575.

Yang X, Chockalingam SP, Aluru S: A survey of error-correction methods for next-generation sequencing. *Brief Bioinform* 2013, 14(1):56-66.

Yang X, Dorman KS, Aluru S: Reptile: representative tiling for short read error correction. *Bioinformatics* 2010, 26(20):2526-2533.

Yang Z, Rannala B: Molecular phylogenetics: principles and practice. *Nat Rev Genet* 2012, 13(5):303-314.

Yao G, Ye L, Gao H, Minx P, Warren WC, Weinstock GM: Graph accordance of next-generation sequence assemblies. *Bioinformatics* 2012, 28(1):13-16.

Yao Z, Weinberg Z, Ruzzo WL: CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006, 22(4):445-452.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, *et al*: PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010, 26(13):1608-1615.

Yusuf D, Marz M, Stadler P, Hofacker I: Bcheck: a wrapper tool for detecting RNase P RNA genes. *BMC Genomics* 2010, 11(1):432.

Zdobnov EM, Apweiler R: InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 2001, 17(9):847-848.

Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008, 18(5):821-829.

Zhang T, Luo Y, Liu K, Pan L, Zhang B, Yu J, Hu S: BIGpre: a quality assessment package for next-generation sequencing data. Genomics Proteomics *Bioinformatics* 2011, 9(6):238-244.

Zhang W, Yu D, Sun Z, Wu R, Chen X, Chen W, Meng H, Hu S, Zhang H: Complete Genome Sequence of *Lactobacillus casei* Zhang, a New Probiotic Strain Isolated from Traditional Homemade Koumiss in Inner Mongolia, China. *J Bacteriol* 2010, 192(19):5268-5269.

Zhang Z, Liu C, Zhu Y, Zhong Y, Zhu Y, Zheng H, Zhao G, Wang S, Guo X: Complete Genome Sequence of *Lactobacillus plantarum* JDM1. *J Bacteriol* 2009, 191(15):5020-5021.

Zhao W, Chen Y, Sun Z, Wang J, Zhou Z, Sun T, Wang L, Chen W, Zhang H: Complete Genome Sequence of *Lactobacillus helveticus* H10. *J Bacteriol* 2011, 193(10):2666-2667.

Zheng HJ, Wang BF, Zhang XL, Han H, Lu G, Jin L, Pu SY, Hu QP, Zhu GF, Wang SY, *et al*: The complete genome sequence of *Lactobacillus delbrueckii* subsp. bulgaricus 2038. *Trends Cell Mol Biol* 2008, 3:15-30.

Zhou C, Smith J, Lam M, Zemla A, Dyer MD, Slezak T: MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res* 2007, 35(suppl 1):D391-D394.

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

Zhou M, Boekhorst J, Francke C, Siezen RJ: LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics* 2008, 9(1):173.

Zhou M, Theunissen D, Wels M, Siezen R: LAB-Secretome: a genome-scale comparative analysis of the predicted extracellular and surface-associated proteins of Lactic Acid Bacteria. *BMC Genomics* 2010, 11(1):651.

Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: PHAST: a fast phage search tool. *Nucleic Acids Res* 2011, 39(suppl 2):W347-W352.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA: The MaSuRCA genome assembler. *Bioinformatics* 2013, 29(21):2669-2677.

Zmasek CM, Eddy SR: RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002, 3(1):14.

Appendixes

Appendix Table 1. Software and databases that can be used as sources of bacterial genome projects.

Method	Description	Ref
Base calling in Sanger technology		
Phred	Base caller for Sanger sequence traces published in 1998. Introduced the logarithmically related base-calling error probabilities	Ewing & Green, 1998
KB Basecaller	Base caller developed by ABI	-
Base calling in NGS technologies		
Rolexa	Parametric base caller for Illumina sequence data. IUPAC codes used to describe base calling quality	Rougemont *et al.*, 2008
Alta-Cyclic	Mixed Parametric base caller for Illumina data that uses SVMs	Erlich & Mitra, 2008
BayesCall	Parametric base caller for Illumina sequence data	Kao *et al.*, 2009
All Your Base	Method that uses a completely empirical model for generalising both cross-talk and phasing	Massingham & Goldman, 2012
BM-BC	Bayesian method of base calling for Illumina sequence data. Uses a hierarchical model that accounts for three sources of noise in the data	Ji *et al.*, 2012
Ibis	Fully empirical SVM-based base caller for Illumina data	Kircher *et al.*, 2009
Pyrobayes	Base caller for 454 sequence data. Adapts an empirical prior on the homopolymer length	Quinlan *et al.*, 2008
HPCall	454 base calling method that make use of a weighted Hurdle Poisson model	Beuf *et al.*, 2012
Rsolid	R package for normalizing intensity data from SOLiD platform. Should be applied before calling colors	Wu *et al.*, 2010
Quality analysis and read manipulation		
FastQC	Quality control tool for high throughput sequence data	-
PrinSeq	Bioinformatic tool for quality control and data preprocessing of genomic datasets.	Schmieder & Edwards, 2011
BIGpre	Quality assessment package for NGS data	Zhang *et al.*, 2011
NGS QC Toolkit	Standalone and open source application for quality check and filtering of high-quality data	Patel & Jain, 2012

NARWHAL	Software pipeline to automate the primary analysis of Illumina data. Includes de-multiplexin component	Brouwer *et al.*, 2012
Cutadapt	Finds and removes adapter sequences from high-throughput sequencing data	Martin, 2011
fastx	Toolkit for NGS data preprocessing	-
Staden package	Package including a number of tools for DNA sequence assembly, editing and sequence analysis	Staden *et al.*, 1999
Read correction		
Reptile	*k*-spectrum based error correction algorithm. Detects and corrects substitution sequencing errors	Yang *et al.*, 2010
HiTEC	Error correction algorithm that uses suffix trees to identify and correct substitution errors	Ilie *et al.*, 2011
ECHO	*k*-spectrum based error correction algorithm. Does not require user-specified parameters	Kao *et al.*, 2011
Hybrid-SHREC	Read correction tool designated to account substitutions, insertions and deletions	Salmela, 2010
Greedy assemblers		
SSAKE	*De novo* short read DNA assembler. Searches progressively for perfect 3' read matches	Warren *et al.*, 2007
VCAKE	Greedy short read data assembler with robust error correction	Jeck *et al.*, 2007
Overlap-based genome assembly		
Newbler	Whole-genome assembler designed for 454 sequence data	Margulies *et al.*, 2005
EDENA	Exact *de novo* assembler dedicated to process very short reads	Hernandez *et al.*, 2008
SGA	Set of memory efficient assembly algorithms based on the FM-index	Simpson & Durbin, 2012
MIRA	Whole-genome sequence assembler for Sanger, 454 and Illumina	Chevreux *et al.*, 2004
De Bruijn graph based genome assembers		
SPAdes	De Bruijn graph assembly at multiple *k*-mer sizes. Ideal for genomes of varying coverage	Bankevich *et al.*, 2012
MaSuRCA	Algorithm that transforms paired-end reads into longer super-reads. Capable of handling reads from different sequencing platforms	Zimin *et al.*, 2013
RAY	Assembler for simultaneous assembly based on reads from a combination of sequencing	Boisvert *et al.*, 2010

	platforms	
ALL-PATHS	Assembler for large genomes. Requires at least two types of read libraries to work	Butler *et al.*, 2008
SOAPdenovo	Short-read assembly method designed for Illuman reads	Li *et al.*, 2010
Velvet	Sequence assembler dedicated to short read sequences. Can make use of multiple sequence libraries	Zerbino & Birney, 2008
ABySS	*De novo* sequence assembler that supports parallel computing. Capable of assembling larger genomes and one of the best assemblers when used with default parameters	Simpson *et al.*, 2009
Reference-based genome assemblers		
VAAL	Variant ascertainment algorithm. Assemble reads using reference genome for assistance	Nusbaum *et al.*, 2008
Amos-Cmp	Comparative assembler based on MUMmer	Pop *et al.*, 2004b
Scaffolders		
Bambus	General purpose scaffolder that can use mate-pair data or reference genome alignments	Pop *et al.*, 2004a
SSPACE	Stand-alone program for scaffolding pre-assembled contigs using paired-read data	Boetzer *et al.*, 2011
SOMA	Scaffolding using optical restriction mapping	Nagarajan *et al.*, 2008
OSLay	Scaffolder that order and orient contigs based on matching sequences in a target assembly and a reference assembly	Richter *et al.*, 2007
BACCardI	Uses gene pairs from a related reference genome. Reference genome genes are mapped to the analysed genome at the protein level	Bartels *et al.*, 2005
PAGIT	Post-assembly genome-improvement toolkit that has modules for both mate pair and reference genome based scaffolding	Swain *et al.*, 2012
Assembly integrators		
Minimus2	Basic genome assembler for merging one or two sequence sets	Sommer *et al.*, 2007
MAIA	Graph-based algorithm for integration of several *de novo* and comparative assemblies	Nijkamp, *et al.*, 2010
GAA	Graph accordance assembly program	Yao *et al.*, 2012
***Ab initio* CDS predictors**		
Glimmer	Interpolated Markov model based CDS predictor. Uses long orfs to train CDS models	Delcher *et al.*, 2007
GeneMark	Popular gene caller suite. Includes GeneMarkS that uses a self-training procedure to derive model parameters. RBS information used to locate the correct start site	Besemer *et al.*, 2001

EasyGene	HMM-based gene finder. Extensions of similarities in Swiss-Prot are used to learn parameters	Larsen & Krogh, 2003
Prodigal	Prokaryotic gene caller that uses dynamic programming to find CDSs. Suitable also for organisms with high GC genomes	Hyatt *et al.*, 2010
ZCURVE	Gene caller based on the Z-curve representation of the DNA sequences	Guo *et al.*, 2003
Evidence-base CDS predictors		
ORPHEUS	Gene caller that guide gene prediction based on database similarity search	Frishman *et al.*, 1998
CRITICA	Gene calling system that uses comparative analysis to derive statistical characteristics of CDSs. Initial coding regions are detected based on non-synonymous mutations	Badger & Olsen, 1999
CDS model integrators		
Reganor	Combines evidence from Glimmer and CRITICA to produce a consensus gene model	McHardy *et al.*, 2004
YACOP	Produces a consensus gene model by integrating gene models predictions of CRITICA, Glimmer and ZCURVE	Tech & Merkl, 2003
CDS model refinement tools		
GenePRIMP	Gene prediction improvement pipeline. Handle various types of gene calling anomalies	Pati *et al.*, 2010
Mugsy-Annotator	Identifies anomalies in annotated gene structures based on comparative genomics. Useful for standardising annotations across closely related organisms	Angiuoli *et al.*, 2011
ORFCor	Corrects annotation inconsistencies based on consensus start and stop positions derived from sets of closely related orthologs	Klassen & Currie, 2013
MICheck	Microbial genome checker is a tool for finding missed or inaccurate gene annotations and frameshifts	Cruveiller *et al.*, 2005
ncRNA predictors		
RNAmotif	RNA secondary structure definition and search algorithm	Macke *et al.*, 2001
RNAmmer	HMM-based gene predictor to annotate 5s, 16s and 23s ribosomal RNA in full genome sequences	Lagesen *et al.*, 2007
QRNA	Gene finding system that uses comparative genome sequence analysis to guide RNA prediction	Rivas & Eddy, 2001

RNAz	Predicts structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments	Washietl *et al.*, 2005
Aragorn	Employs heuristic algorithms to predict tRNA and tmRNAs based on secondary structure on homology with known RNA consensus sequences and ability to form a base-paired cloverleaf	Laslett & Canback, 2004
tRNAscan-SE	Gene caller that uses covariance models to detect tRNAs	Lowe & Eddy, 1997
Infernal	General RNA structure and sequence similarity search tool that uses covariance models for searching DNA sequence databases	Nawrocki *et al.*, 2009
SRP-scan	Gene finding system that uses covariance models to annotate SRP RNA genes in genomic DNA sequences	Regalia *et al.*, 2002
Bcheck	Detects RNase P RNA genes using pattern matching and covariance models	Yusuf *et al.*, 2010
CMFinder	RNA motif finding algorithm that uses both expectation maximization and covariance models	Yao *et al.*, 2006
Intrinsic terminators		
TransTermHP	Annotates low-energy hairpins followed by a stretch of thymines in bacterial genomes	Kingsford *et al.*, 2007
RNIE	Bioinformatic software that uses covariance models to find intrinsic terminators	Gardner *et al.*, 2011
CRISPR arrays		
CRT tool	CRISPR recognition tool. Searches for series of k-mers separated by a similar distance by sliding windows	Bland *et al.*, 2007
PILER-CR	Extension to the PILER family of repeat analysis algorithms for fast and accurate identification of CRISPR repeats	Edgar, 2007
CRISPRFinder	Web tool offering tools to detect CRISPRs and for their comprative analysis	Grissa *et al.*, 2007
Repeats		
REPuter	Software for repeat analysis on a genomic scale	Kurtz & Schleiermacher, 1999
RepeatScout	Discovers repetitive substrings in DNA	Price *et al.*, 2005
Insertion sequences		
ISfinder database	Database of insertion sequences isolated from eubacteria and archae	Siguier *et al.*, 2006
IScan	Package to identify known insertion sequences in bacterial genomes	Wagner *et al.*, 2007

ISsaga	Web application pipeline for insertion sequence annotation	Varani *et al.*, 2011
Prophages		
ACLAME	Database dedicated to the collection and classification of mobile genetic elements. Provides access to a list of functional ontologies	Leplae *et al.*, 2004
PhiPsy	Phage detection algorithm that calls prophage regions based on seven distinctive characteristics of prophages including similarity of phage proteins as well as other statistics	Akhter *et al.*, 2012
Phage_Finder	Heuristic computer program to identify prophage regions in bacterial genomes. Uses comparative genomics appraoches to locate prophage regions	Fouts, 2006
ProphageDB	Database of prophage elements and phage remnants	Srividhya *et al.*, 2007
PHAST	Homology-based tool to annotate prophage sequences	Zhou *et al.*, 2011
Prophinder	Detects genome regions with a significantly high density of known phage-like proteins from ACLAME	Lima-Mendez *et al.*, 2008
Genomic islands		
SIGI-HMM	Discriminative genomic island caller that uses HMMs to identify genome regions with abnormal sequence composition statistics	Waack *et al.*, 2006
IslandViewer	Web-server for predicting genomic islands. Provides access to three different tools	Langille & Brinkman, 2009
PAI-IDA	Combines dinucleotide frequency, G+C content and codon usage to predict the location of genomic islands in genomes	Tu & Ding, 2003
Alien_Hunter	Discriminative genomic island prediction method that uses interpolated variable order motifs to identify genome islands	Vernikos & Parkhill, 2006
IslandPick	One of the few comparative genomics-based genomic island identification tools	Langille *et al.*, 2008
Plasmids, integrative and conjugative elements, gene cassettes, integrons		
ICEberg	Web-based resource for integrative and conjugative elements found in bacteria	Bi *et al.*, 2012
INTEGRALL	Database and search engine for integrons, integrases and gene cassettes.	Moura *et al.*, 2009
ACID	Community resource for annotation of cassette and integron data	Joss *et al.*, 2009
cBar	Distinguishes plasmid-derived from chromosome-derived sequence fragments based on their sequence composition	Zhou *et al.*, 2010

Origin of replication		
Ori-Finder	Finds origins of replication. Combines evidence from base composition asymmetry, distribution of DnaA boxes and the occurrence of genes frequently close to oriCs to produce predictions	Gao & Zhang, 2008
Sequence database search		
BLAST and PSI-BLAST	Set of programs to find similarity between a query protein or DNA sequence and a sequence database	Altschul *et al.*, 1997
FASTA	Suite of sequence analysis tools for searching Protein or DNA sequence databases. More accurate than BLAST, but slower to run	Pearson & Lipman, 1988
HMMER3	Sequence comparison toolikit implemeting profile hidden Markov models. Accurate and fast to run	Eddy, 2011
Biological databases		
GenBank	One of the largest genetic sequence databases. Maintained by NCBI	Benson *et al.*, 2013
UniProt	UniProt is a catalog of information on proteins. Includes sequence and annotation data	Bairoch *et al.*, 2008
PATRIC	Information system that integrates bacterial genome information with rich data and analysis tools	Gillespie *et al.*, 2011
CharProtDB	Resource of expertly curated, experimentally characterised proteins described in published literature	Madupu *et al.*, 2012
Rfam	Collection of RNA families	Griffiths-Jones *et al.*, 2003
COG	Database of clusters of orthologous groups of proteins. Orthologue groups build by merging triangles with a common side. Original versions classified sequences into 23 functional categories	Tatusov *et al.*, 1997, Tatusov *et al.*, 2003
SEED	Subsystem-based approach and database to high-throughput genome annotation	Overbeek *et al.*, 2005
Protein signature databases		
PFAM	Collection of protein domain families represented by multiple sequence alignments and HMMs	Punta *et al.*, 2012
TigrFAM	Collection of protein families featuring curated multiple sequence alignments, HMMs and associated information. Most signature models represent full-length proteins	Haft *et al.*, 2003

HAMAP	Collection of manually curated family profiles for protein classification and associated manually created annotation rules. Most signature profiles represent full-length proteins	Lima *et al.*, 2009
Interpro	Integrative protein signature database of protein families, domains and functional sites	Hunter *et al.*, 2012
CDD	Conserved domains and protein classification system	Marchler-Bauer *et al.*, 2011
General function		
FunCut	Method for generating functional annotations based on multiple homologs	Abascal & Valencia, 2003
CLAN	Clusters proteins according to both annotation and sequence similarity. Provides to highlight protein function assignment inconsistencies among similar sequences	Kunin & Ouzounis, 2005
Gotcha	Method for predicting gene product function by annotation with GO terms. Uses multiple homologs	Martin *et al.*, 2004
GOPET	GO term prediction and evaluation tool that uses SVMs	Vinayagam *et al.*, 2006
ARGOT2	Annotates query sequences based on similarity-score weighted GO terms and by taking into account the semantic similarity relations of GO terms described by the Gene Ontology	Falda *et al.*, 2012
BLANNOTATOR	Protein function prediction system that groups sequences identified by BLAST into subsets according to their GO annotation before performing annotation transfer from the best subset to the query	Study II
PFP	Sequence-based predictor of GO functional terms. Links query proteins also with GO terms that are highly associated to those terms associated to sequence hits	Hawkins *et al.*, 2006
PANNZER	High-throughput functional annotation of uncharacterized proteins based on weighted k-nearest neighbour method with statistical testing	Koskinen *et al.*, 2015
Sma3s	Accurate and flexible protein function prediction tool specifically designed for the annotation of large collections of sequences. Has a component for defining optimal sequence similarity thresholds for the given sequences	Muñoz-Mérida *et al.*, 2014
ConFunc	Protein function prediction system that uses conserved residues to generate sequence profiles to infer function	Wass & Sternberg, 2008
SIFTER	Statistical inference of function through evolutionary relationships	Engelhardt *et al.*, 2005

InterProScan	Tool that combines different protein signature recognition methods into one resource and allows to query different databases at one go	Zdobnov & Apweiler, 2001
Advanced function classification		
BAGEL3	Web-based bacteriocin genome mining tool	van Heel *et al.*, 2013
RASTA-Bacteria	Bioinformatic tool for identifying toxin-antitoxin loci in bacteria	Sevin & Barloy-Hubler, 2007
TADB	Resource for type 2 toxin-antitoxin loci in prokaryotes	Shao *et al.*, 2011
ARGO	Database of βlactam and vancomycin resistance genes	Scaria *et al.*, 2005
MvirDB	Database of protein toxins, virulence factors and antibiotic resistance genes	Zhou *et al.*, 2007
ARDB	Antibiotic resistance gene database	Liu & Pop, 2009
DBD	Transcription factor database	Wilson *et al.*, 2008
antiSMASH	Software pipeline for secondary metabolite biosynthesis gene cluster identification	Medema *et al.*, 2011
Mist2	Microbial signal transduction database	Ulrich & Zhulin, 2010
LOCP	Tool for locating pilus operons in gram-positive bacteria	Study I
Metabolism related genes		
PRIAM	Method for automated enzyme detection. Relies on sets of position-specific score matrices tailored for each enzyme class	Claudel-Renard *et al.*, 2003
KAAS	Annotation server that provides functional annotation of CDSs by BLAST comparisons against the known KEGG genes. Constructs also KEGG pathway mappings	Moriya *et al.*, 2007
TCDB	Database and classification system for membrane transport proteins	Saier *et al.*, 2006
CAZy	Database of enzymes that degrade, modify, or create glycosidic bonds	Cantarel *et al.*, 2009
MEROPS	Database of peptidase and proteins that inhibit them. Proteins classified into families and homology clans	Rawlings *et al.*, 2004
REBASE	Restriction enzyme database	Roberts *et al.*, 2010
Context-based protein function prediction		
ContextMirror	Method to find co-evoluting genes. Builds upon family tree similarities	Juan *et al.*, 2008
String	Database of known and predicted protein-protein interactions. Makes use of homology information and information on genomic context, high-throughput experiments, co-expression and text-mining	von Mering *et al.*, 2005

Prolinks	Database of predicted protein-protein interactions. Interactions dervied from phylogenetic profile, fusion gene, gene neighbor and gene cluster predictions	Bowers *et al.*, 2004
***Ab initio* protein function prediction**		
CSS-Palm	Predicts palmitoylation sites	Ren *et al.*, 2008
DISIS	*Ab initio* DNA-binding residue prediction tool that uses multiple classifying algorithms	Ofran *et al.*, 2007
MetalDetector	Annotates metal binding sites in proteins from sequence information alone	Lippi *et al.*, 2008
VirulentPred	Virulent protein prediction algorithm that uses SVM	Garg & Gupta, 2008
SPAAN	Neural network prediction of adhesins and adhesin-like proteins	Sachdeva *et al.*, 2005
Subcellular location		
PSORTb v3.0	Subcellular localization prediction tool that uses both SVMs and Bayesian networks	Yu *et al.*, 2010
tatP	Ppredicts the presence and location of Twin-arginine signal peptide cleavage sites in bacterial proteins	Bendtsen *et al.*, 2005
LipoP	Lipoprotein signal peptide predictor that uses HMMs	Rahman *et al.*, 2008
SignalP	Signal sequence prediction utility that uses a combination of several artificial neural networks to product signal peptide cleavage sites	Petersen *et al.*, 2011
TMHMM	HMM-based protein topology predictor. Identifies transmembrane helices in proteins	Krogh *et al.*, 2001
LocateP	Genome-scale subcellular-location prediction suite. Distinguishes between seven cellular locations for gram-positive proteins	Zhou *et al.*, 2008
EffectiveT3	Sequence-based prediction of type III secreted proteins	Arnold *et al.*, 2009
CoBaltDB	Complete prokaryote protein subcellular localization database and associated resources. The database integrates the results of 43 localization predictors for over 700 complete prokaryotes proteomes	Goudenège *et al.*, 2010
Orthology prediction		
RBH	Reciprocal best hit approach. Find two-way best hits among genes in two organisms	Tatusov *et al.*, 1996
RSD	Reciprocal smallest distance algorithm. Uses global sequence alignment and maximum likelihood estimation of evolutionary distances to predict orthologs	Wall *et al.*, 2003
RIO	Mehtod for automated phylogenomics using explicit phylogenetic inference	Zmasek & Eddy, 2002
OrthoStrapper	Program for predicting orthologs between two species. Calculates orthology support	Storm & Sonnhammer, 2002

	values	
InParanoid	Finds orthologous genes and paralogous genes that arose after some speciation event between two species	Remm *et al.*, 2001
PoFF	Orthology grouping by combining clustering, sequence similarity, and conservation of gene order	Lechner *et al.*, 2014
morFeus	Program to call remotely conserved orthologs. Uses relaxed sequence similarity searches, iterative reciprocal BLAST searches and network score calculation to find orthologous relationships between sequences	Wagner et al., 2014
EggNOG	The orthology prediction method used by STRING database. Similar to that of COG	Jensen *et al.*, 2008
OrthoMCL	Scalable method for constructing orthologue groups across multiple species. Uses Markov cluster algorithm to group orthologs and in-paralogs	Li *et al.*, 2003
Protein cluster	NCBI's collection of related protein sequences. Orthologue groups created by finding maximum cliques	Klimke *et al.*, 2009
OMA	Database of orthologs among publicly available, complete genomes. Evolutionary distances and maximum-weight clique algorithm used to create orthologue groups	Roth *et al.*, 2008
Multiple sequence aligners		
Muscle	Multiple sequence comparison by log-expectation. Good average accuracy and speed	Edgar, 2004
ProbCons	Probabilistic consistency-based multiple alignment program for amino acid sequences	Do *et al.*, 2005
Clustal Omega	General purpose multiple sequence alignment program for protein and DNA/RNA sequences	Sievers *et al.*, 2011
Whole-genome aligners		
MUMmer	Ultra-fast alignment program for DNA and protein sequences. Generates alignments between two sequences. Does not support duplications	Darling *et al.*, 2004; Darling *et al.*, 2010
TBA	Threaded blockset aligner. Supports multiple sequence alignments. Does not require a reference genome, but aligment blocks need to be projected against a reference genome	Blanchette *et al.*, 2004
Pecan	Practical global multiple sequence alignment program. Can report duplications, does not	Paten *et al.*, 2008

	require a reference genome, but needs a guide tree	
Mauve aligner	Constructs multiple genome alignments in the presence of rearrangements and insertions. No support for duplications. Mauve contig mover is a popular scaffolder	Rissman *et al.*, 2009
Gepard	Rapid and sensitive dotplot tool. Dotplots are useful in detecting inversions, duplications and rerarrangements between two sequences	Krumsiek *et al.*, 2007
BLASTatlas	Maps of genome homology of a list of sequences against a reference genome. Good in highlighting conserved and unconserved genome areas in the reference genome	Wassenaar *et al.*, 2010
Prediction of constrained elements		
GERP	Identifies constrained elements in multiple sequence alignments	Cooper *et al.*, 2005
SiPhy	Software package for detecing bases under selection from a multiple alignment data	Garber *et al.*, 2009
Phylogenetic trees		
PhyML	Method to estimate maximum-likelihood phylogenies	Guindon *et al.*, 2003
BioNJ	Neighbor-joining algorithm to estimate phylogenies	Gascuel, 1997
Phylip	Package of programs for inferring phylogenies and evolutionary relationships between sequences	Felsenstein, 1989
Metabolic reconstruction		
Pathway tools	Utility for capturing and integrating metabolic information. Uses MetaCyc pathways	Karp *et al.*, 2002
FMM	Web-resource for reconstructing metabolic pathways form one metabolite to the other one by combining KEGG maps	Chou *et al.*, 2009
MetaCyc	Database of metabolic pathways and enzymes. Pathways in MetaCyc are more compact and contain on average fewer reactions than those of KEGG	Krieger *et al.*, 2004
UniPathway	Manually curated resource of enzyme-catalyzed and spontaneous chemical reactions and pathways	Morgat *et al.*, 2012
KEGG	Database of metabolic pathways and enzymes	Kanehisa *et al.*, 2004
Annotation pipelines		
IMG	System to support the annotation and analysis microbial genome datasets. Genomes are associated with rich annotation data. Various visualization tools	Markowitz *et al.*, 2012

RAST	Server for high quality genome annotation of prokaryotes. Includes a component that can be used to produce a draft metabolic mode	Aziz *et al.*, 2008
DOE-JGI MAP	Microbial annotation pipeline	Mavromatis *et al.*, 2009
CG pipeline	Resource for assembling sequence data and running feature prediction and annotation tools on the assembly	Kislyuk *et al.*, 2010
ERGO	Commercial microbial genome analysis and discovery system	Overbeek *et al.*, 2003
PGAAP	NCBI prokaryotic genome annotation pipeline	-
Broad Institute	Broad prokaryotic genome annotation pipeline	-
BCM	Baylor Prokaryotic Annotation Pipeline	-
JCVI	The automated prokaryotic annotation pipeline of TIGR/JCVI institute	Tanenbaum *et al.*, 2010
Visualization		
Genome atlas	Circular plots of chromosomes or plasmids on which general properties of the DNA molecule are plotted as colors	Wassenaar *et al.*, 2010
ACT	Genome browser for feature viewing and annotation. Can display pairwise comparisons between two or more DNA sequences	Carver *et al.*, 2005
Combo	Argo comparative genome viewer	Engels *et al.*, 2006
Circoletto	Tool for the generation of circularly composited renditions of genomic data and sequence similarity	Darzentas, 2010

Appendix Table 2. Characteristics of the sequenced *Lactobacillus* and their genomes.

Organism	Clade	Source	Release Date	Length (Mb)	Status	CDSs	Ortholog groups	RNA genes	16S	Original reference
L. farciminis KCTC 3681	*L. ali*	Sausage	17.12.2010	2.5	Draft	2440	349	57	1	Nam *et al.*, 2011b
L. versmoldensis KCTC 3814	*L. ali*	Fermented salami	21.06.2011	2.4	Draft	2354	290	55	2	Kim *et al.*, 2011a
L. brevis ATCC 27305	*L. bre*	Wine	12.02.2009	3.1	Draft	3041	254	61	1	Nelson *et al.*, 2010
L. brevis ATCC 367	*L. bre*	Silage	13.10.2006	2.3	Complete	2218	235	81	5	Makarova *et al.*, 2006
L. buchneri ATCC 11577	*L. buc*	Human oral cavity	12.02.2009	2.9	Draft	3002	228	60	1	Nelson *et al.*, 2010
L. buchneri NRRL B-30929	*L. buc*	Ethanol production plant	18.04.2011	2.6	Complete	2392	134	78	5	Liu *et al.*, 2011
L. hilgardii ATCC 8290	*L. buc*	Wine	19.02.2009	2.6	Draft	2791	190	60	1	Nelson *et al.*, 2010
L. kisonensis F0435	*L. buc*	Human oral cavity	10.01.2012	3.0	Draft	3325	574	46	1	Nelson *et al.*, 2010
L. parafarraginis F0439	*L. buc*	Human oral cavity	16.12.2011	2.9	Draft	3183	506	44	0	Nelson *et al.*, 2010
L. casei ATCC 334	*L. cas*	Swiss cheese	13.10.2006	2.9	Complete	2771	72	75	5	Makarova *et al.*, 2006
L. casei BD-II	*L. cas*	Koumiss	01.04.2011	3.1	Complete	3204	42	74	5	Ai *et al.*, 2011
L. casei BL23	*L. cas*	Laboratory strain	19.06.2008	3.1	Complete	3044	83	0	0	Cai *et al.*, 2009
L. casei LC2W	*L. cas*	Mongolian dairy product	01.04.2011	3.1	Complete	3164	55	73	5	Chen *et al.*, 2011
L. casei Zhang	*L. cas*	Koumiss	12.07.2010	2.9	Complete	2848	64	74	5	Zhang *et al.*, 2010
L. paracasei 8700:2	*L. cas*	Human GIT	26.08.2008	3.0	Draft	3021	140	49	0	Nelson *et al.*, 2010
L. paracasei ATCC 25302	*L. cas*	Dairy product	20.02.2009	2.9	Draft	3042	189	58	1	Nelson *et al.*, 2010

L. rhamnosus ATCC 21052	*L. cas*	Human faeces	18.11.2011	2.9	Draft	3014	165	46	1	Nelson *et al.*, 2010
L. rhamnosus ATCC 8530	*L. cas*	-	03.11.2011	3.0	Complete	2887	65	75	5	Pittet *et al.*, 2012
L. rhamnosus GG	*L. cas*	Human faeces	02.09.2009	3.0	Complete	2944	39	72	5	Study III
L. rhamnosus GG	*L. cas*	Human faeces	25.09.2009	3.0	Complete	2834	20	71	5	Morita *et al.*, 2009
L. rhamnosus HN001	*L. cas*	Cheese	11.09.2008	2.9	Draft	2758	69	53	1	-
L. rhamnosus LC705	*L. cas*	Starter culture	02.09.2009	3.0	Complete	2992	44	76	5	Study III
L. rhamnosus LMS2-1	*L. cas*	Human GIT	16.03.2009	3.1	Draft	3155	102	51	1	Nelson *et al.*, 2010
L. rhamnosus MTCC 5462	*L. cas*	Human faecal	07.04.2011	2.5	Draft	3255	942	53	3	Prajapati *et al.*, 2012
L. rhamnosus R0011	*L. cas*	Dairy starter culture	18.11.2011	2.9	Draft	2719	15	63	3	Tompkins *et al.*, 2012
L. zeae KCTC 3804	*L. cas*	Corn steep liquor	21.06.2011	3.1	Draft	2958	281	57	3	Kim *et al.*, 2011b
L. acidophilus 30SC	*L. del*	Swine GIT	07.03.2011	2.1	Complete	2059	72	75	4	Oh *et al.*, 2011
L. acidophilus ATCC 4796	*L. del*	-	10.03.2009	2.0	Draft	2020	79	61	1	Nelson *et al.*, 2010
L. acidophilus NCFM	*L. del*	Human faeces	27.01.2005	2.0	Complete	1862	29	74	4	Altermann *et al.*, 2005
L. amylolyticus DSM 11664	*L. del*	Acidified beer wort	23.04.2010	1.5	Draft	1684	182	58	1	Nelson *et al.*, 2010
L. amylovorus GRL1112	*L. del*	Porcine feces	19.11.2010	2.1	Complete	2121	121	72	4	Kant *et al.*, 2011c
L. amylovorus GRL1118	*L. del*	Porcine GIT	29.03.2011	2.0	Complete	1920	57	74	4	Kant *et al.*, 2011b
L. crispatus 125-2-CHN	*L. del*	Human vagina	24.08.2009	2.3	Draft	2082	24	57	0	Nelson *et al.*, 2010
L. crispatus 214-1	*L. del*	Human vagina	11.03.2010	2.1	Draft	2163	50	55	1	Nelson *et al.*, 2010
L. crispatus CTV-05	*L. del*	Human vagina	29.10.2010	2.4	Draft	2248	115	51	1	Nelson *et al.*, 2010
L. crispatus JV-V01	*L. del*	Human vagina	27.04.2009	2.1	Draft	2209	63	66	1	Nelson *et al.*, 2010
L. crispatus MV-1A-US	*L. del*	Human vagina	24.08.2009	2.3	Draft	2151	22	62	0	Nelson *et al.*, 2010
L. crispatus MV-3A-US	*L. del*	Human vagina	01.10.2009	2.4	Draft	2330	46	57	0	Nelson *et al.*, 2010
L. crispatus SJ-3C-US	*L. del*	Human vagina	08.07.2010	2.1	Draft	2174	146	71	3	Nelson *et al.*, 2010
L. crispatus ST1	*L. del*	Chicken crop	21.04.2010	2.0	Draft	2024	72	76	4	Study IV
L. delbrueckii 2038	*L. del*	Dairy product	03.03.2011	1.9	Complete	1792	41	115	9	Zheng *et al.*, 2008
L. delbrueckii ATCC 11842	*L. del*	Bulgarian yogurt	26.05.2006	1.9	Complete	1562	67	122	9	van de Guchte *et al.*, 2006

L. delbrueckii ATCC BAA-365	*L. del*	Starter culture	13.10.2006	1.9	Complete	1721	87	126	9	Makarova *et al.*, 2006
L. delbrueckii CNCM I-1519	*L. del*	Yogurt starter culture	23.09.2011	1.8	Draft	1893	62	64	1	McNulty *et al.*, 2011
L. delbrueckii CNCM I-1632	*L. del*	Yogurt starter culture	07.09.2011	1.8	Draft	1850	54	69	1	McNulty *et al.*, 2011
L. delbrueckii DSM 20072	*L. del*	Cheese	10.03.2011	1.9	Draft	2006	155	74	1	Nelson *et al.*, 2010
L. delbrueckii ND02	*L. del*	Fermented yak milk	19.11.2010	2.1	Complete	2018	130	121	9	Sun *et al.*, 2011
L. delbrueckii PB2003/044-T3-4	*L. del*	Human vagina	13.07.2010	2.0	Draft	1909	77	69	2	Nelson *et al.*, 2010
L. gasseri 202-4	*L. del*	Human vagina	03.06.2009	1.8	Draft	1773	21	45	1	Nelson *et al.*, 2010
L. gasseri 224-1	*L. del*	Human vagina	30.12.2009	2.0	Draft	2252	160	97	5	Nelson *et al.*, 2010
L. gasseri ATCC 33323	*L. del*	Human	13.10.2006	1.9	Complete	1755	24	97	6	Makarova *et al.*, 2006
L. gasseri JV-V03	*L. del*	Human vagina	19.02.2009	2.0	Draft	1977	82	55	1	Nelson *et al.*, 2010
L. gasseri MV-22	*L. del*	Human vagina	14.10.2008	1.9	Draft	1945	67	39	1	Nelson *et al.*, 2010
L. gasseri SJ-9E-US	*L. del*	Human vagina	08.07.2010	1.8	Draft	1688	3	66	2	Nelson *et al.*, 2010
L. gasseri SV-16A-US	*L. del*	Human vagina	08.07.2010	2.0	Draft	1955	28	74	1	Nelson *et al.*, 2010
L. helveticus DPC 4571	*L. del*	Cheese	15.11.2007	2.1	Complete	1610	13	73	4	Callanan *et al.*, 2008
L. helveticus DSM 20075	*L. del*	Cheese	27.04.2009	1.8	Draft	2078	144	48	1	Nelson *et al.*, 2010
L. helveticus H10	*L. del*	Traditional fermented milk	16.02.2011	2.2	Complete	1978	65	74	4	Zhao *et al.*, 2011
L. helveticus MTCC 5463	*L. del*	Human vagina	07.04.2011	2.0	Draft	2239	421	68	2	Prajapati *et al.*, 2011
L. iners AB-1	*L. del*	Human vagina	02.02.2010	1.3	Draft	1209	13	49	5	Macklaim *et al.*, 2011
L. iners ATCC 55195	*L. del*	Human	27.12.2010	1.2	Draft	1144	11	49	1	Nelson *et al.*, 2010
L. iners DSM 13335	*L. del*	Human urine	27.04.2009	1.3	Draft	1214	9	47	1	Nelson *et al.*, 2010
L. iners LactinV 01V1-a	*L. del*	Human vagina	05.10.2010	1.3	Draft	1527	108	50	1	Nelson *et al.*, 2010
L. iners LactinV 03V1-b	*L. del*	Human vagina	05.10.2010	1.3	Draft	1459	84	51	1	Nelson *et al.*, 2010
L. iners LactinV 09V1-c	*L. del*	Human vagina	05.10.2010	1.3	Draft	1361	26	51	1	Nelson *et al.*, 2010

L. iners LactinV 11V1-d	*L. del*	Human vagina	05.10.2010	1.3	Draft	1338	27	51	1	Nelson *et al.*, 2010
L. iners LEAF 2052A-d	*L. del*	Human vagina	04.11.2010	1.3	Draft	1256	24	52	1	Nelson *et al.*, 2010
L. iners LEAF 2053A-b	*L. del*	Human vagina	04.11.2010	1.4	Draft	1277	45	52	1	Nelson *et al.*, 2010
L. iners LEAF 2062A-h1	*L. del*	Human vagina	04.11.2010	1.3	Draft	1265	13	49	1	Nelson *et al.*, 2010
L. iners LEAF 3008A-a	*L. del*	Human vagina	04.11.2010	1.3	Draft	1210	5	51	1	Nelson *et al.*, 2010
L. iners SPIN 1401G	*L. del*	Human vagina	15.04.2011	1.3	Draft	1238	27	29	0	Nelson *et al.*, 2010
L. iners SPIN 2503V10-D	*L. del*	Human vagina	05.10.2010	1.3	Draft	1273	21	49	1	Nelson *et al.*, 2010
L. iners UPII 143-D	*L. del*	Human vagina	07.03.2011	1.3	Draft	1186	17	53	1	Nelson *et al.*, 2010
L. iners UPII 60-B	*L. del*	Human vagina	07.03.2011	1.3	Draft	1276	23	51	1	Nelson *et al.*, 2010
L. jensenii 115-3-CHN	*L. del*	Human vagina	01.10.2009	1.6	Draft	1470	3	50	0	Nelson *et al.*, 2010
L. jensenii 1153	*L. del*	Human vagina	14.10.2008	1.7	Draft	1347	43	60	2	Nelson *et al.*, 2010
L. jensenii 269-3	*L. del*	Human vagina	03.06.2009	1.7	Draft	1575	21	46	1	Nelson *et al.*, 2010
L. jensenii 27-2-CHN	*L. del*	Human vagina	24.08.2009	1.7	Draft	1476	11	50	0	Nelson *et al.*, 2010
L. jensenii JV-V16	*L. del*	Human vagina	19.02.2009	1.6	Draft	1450	37	63	2	Nelson *et al.*, 2010
L. jensenii SJ-7A-US	*L. del*	Human vagina	01.10.2009	1.7	Draft	1630	62	49	0	Nelson *et al.*, 2010
L. johnsonii ATCC 33200	*L. del*	Human blood	19.02.2009	1.8	Draft	1838	96	55	1	Nelson *et al.*, 2010
L. johnsonii DPC 6026	*L. del*	Porcine GIT	20.04.2011	2.0	Complete	1772	17	68	4	Guinane *et al.*, 2011
L. johnsonii FI9785	*L. del*	Poultry GIT	04.11.2009	1.8	Complete	1735	57	67	4	Wegmann *et al.*, 2009
L. johnsonii NCC 533	*L. del*	Human GIT	02.02.2004	2.0	Complete	1821	27	97	6	Pridmore *et al.*, 2004
L. johnsonii PF01	*L. del*	Piglet feces	28.06.2011	1.9	Draft	1846	113	43	3	Lee *et al.*, 2011a
L. kefiranofaciens ZW3	*L. del*	Kefir grain	25.05.2011	2.4	Complete	2162	249	0	0	Wang *et al.*, 2011a
L. ultunensis DSM 16047	*L. del*	Human GIT	19.02.2009	2.2	Draft	2210	201	58	1	Nelson *et al.*, 2010
L. fructivorans KCTC 3543	*L. fru*	Spoiled sake	05.01.2011	1.4	Draft	1358	163	39	1	Nam *et al.*, 2012
L. sanfranciscensis TMW 1.1304	*L. fru*	Sourdough	06.09.2011	1.4	Complete	1284	127	82	7	Vogel *et al.*, 2011
L. malefermentans KCTC 3548	*L. mal*	Beer	22.06.2011	2.0	Draft	1968	196	65	2	Kim *et al.*, 2011c
L. pentosus MP-10	*L. pla*	Fermented green	23.05.2011	3.8	Draft	2755	34	64	1	Abriouel *et al.*, 2011

		olives								
L. plantarum ATCC 14917	*L. pla*	Pickled cabbage	20.02.2009	3.2	Draft	3154	153	62	1	Nelson *et al.*, 2010
L. plantarum JDM1	*L. pla*	-	17.07.2009	3.2	Complete	2948	63	78	5	Zhang *et al.*, 2009
L. plantarum NC8	*L. pla*	Grass silage	16.02.2012	3.2	Draft	2868	27	75	5	Axelsson *et al.* 2012
L. plantarum ST-III	*L. pla*	Kimchi	01.10.2010	3.3	Complete	3038	84	79	5	Wang *et al.*, 2011b
L. plantarum WCFS1	*L. pla*	Human oral cavity	05.02.2003	3.3	Complete	3108	83	85	5	Kleerebezem *et al.*, 2003
L. antri DSM 16041	*L. reu*	Human GIT	27.04.2009	2.2	Draft	2224	183	59	1	Nelson *et al.*, 2010
L. coleohominis 101-4-CHN	*L. reu*	Human vagina	24.08.2009	1.7	Draft	1652	131	57	0	Nelson *et al.*, 2010
L. coryniformis KCTC 3167	*L. reu*	Silage	17.11.2010	2.7	Draft	2678	152	35	1	Nam *et al.*, 2011a
L. coryniformis KCTC 3535	*L. reu*	Kimchi	17.12.2010	2.8	Draft	2818	216	54	1	-
L. fermentum 28-3-CHN	*L. reu*	Human vagina	01.10.2009	2.0	Draft	1880	40	53	0	Nelson *et al.*, 2010
L. fermentum ATCC 14931	*L. reu*	Fermented beets	12.02.2009	1.8	Draft	1866	93	60	1	Nelson *et al.*, 2010
L. fermentum CECT 5716	*L. reu*	Human Milk	29.06.2010	2.1	Complete	1051	15	74	5	Jiménez *et al.*, 2010a
L. fermentum IFO 3956	*L. reu*	Fermented plant material	15.04.2008	2.1	Complete	1843	34	69	5	Morita *et al.*, 2008
L. gastricus PS3	*L. reu*	Human milk	16.02.2012	1.9	Draft	1269	11	43	1	-
L. mucosae LM1	*L. reu*	-	21.02.2012	2.2	Draft	2039	320	58	1	Lee *et al.*, 2012
L. oris F0423	*L. reu*	Human oral cavity	25.07.2011	2.2	Draft	2050	55	84	5	Nelson *et al.*, 2010
L. oris PB013-T2-3	*L. reu*	Human vagina	04.11.2010	2.1	Draft	2038	69	62	1	Nelson *et al.*, 2010
L. reuteri 100-23	*L. reu*	Rat GIT	25.07.2008	2.3	Draft	2181	157	88	6	Frese *et al.*, 2011
L. reuteri ATCC 53608	*L. reu*	Swine GIT	24.03.2011	2.0	Draft	1931	69	76	3	Heavens *et al.*, 2011
L. reuteri CF48-3A	*L. reu*	Human faeces	11.03.2009	2.0	Draft	2164	63	55	2	Nelson *et al.*, 2010
L. reuteri DSM 20016	*L. reu*	Human faeces	01.06.2007	2.0	Complete	1900	4	86	6	Frese *et al.*, 2011
L. reuteri JCM 1112	*L. reu*	Human faeces	15.04.2008	2.0	Complete	1820	3	81	6	Morita *et al.*, 2008
L. reuteri lpuph	*L. reu*	Mouse	13.07.2010	2.1	Draft	2008	95	82	4	Frese *et al.*, 2011
L. reuteri mlc3	*L. reu*	Mouse	13.07.2010	2.0	Draft	1962	71	77	9	Frese *et al.*, 2011

L. reuteri MM2-3	*L. reu*	Human Milk	20.04.2009	1.9	Draft	2045	54	56	1	Nelson *et al.*, 2010
L. reuteri MM4-1A	*L. reu*	Human Milk	19.02.2009	2.1	Draft	2095	18	111	6	Nelson *et al.*, 2010
L. reuteri SD2112	*L. reu*	Human Milk	20.06.2011	2.3	Complete	2300	43	88	6	Nelson *et al.*, 2010
L. vaginalis ATCC 49540	*L. reu*	Human vagina	19.02.2009	1.8	Draft	1870	196	62	3	Nelson *et al.*, 2010
L. curvatus CRL 705	*L. sak*	Fermented sausage	26.10.2011	1.8	Draft	1862	145	56	1	Hebert *et al.*, 2012
L. sakei 23K	*L. sak*	French sausage	02.11.2005	1.9	Complete	1885	173	84	7	Chaillou *et al.*, 2005
L. acidipiscis KCTC 13900	*L. sal*	Cheese	21.06.2011	2.3	Draft	2262	386	55	3	-
L. animalis KCTC 3501	*L. sal*	Kimchi	09.12.2010	1.9	Draft	1823	223	29	1	Nam *et al.*, 2011c
L. mali KCTC 3596	*L. sal*	Apple juice	21.06.2011	2.7	Draft	2642	440	58	1	Kim *et al.*, 2011d
L. ruminis ATCC 25644	*L. sal*	Human faeces	19.02.2009	2.1	Draft	2251	292	59	1	Forde *et al.*, 2011
L. ruminis SPM0211	*L. sal*	Human faeces	09.06.2011	2.2	Draft	2326	369	62	1	Lee *et al.*, 2011b
L. salivarius ACS-116-V-Col5a	*L. sal*	Human vagina	19.07.2010	2.0	Draft	2121	108	58	1	Nelson *et al.*, 2010
L. salivarius ATCC 11741	*L. sal*	Human oral cavity	19.02.2009	2.0	Draft	1976	134	66	2	Nelson *et al.*, 2010
L. salivarius CECT 5713	*L. sal*	Human Milk	07.07.2010	2.1	Complete	1552	17	120	7	Jiménez *et al.*, 2010b
L. salivarius GJ-24	*L. sal*	Human faeces	09.06.2011	2.0	Draft	1876	84	83	4	Cho *et al.*, 2011
L. salivarius NIAS840	*L. sal*	Chicken faeces	31.05.2011	2.0	Draft	1869	136	103	7	Ham *et al.*, 2011
L. salivarius SMXD51	*L. sal*	Chicken GIT	14.03.2012	2.0	Draft	1771	57	102	8	Kergourlay *et al.*, 2012
L. salivarius UCC118	*L. sal*	Human GIT	30.03.2006	2.1	Complete	2014	69	99	7	Claesson *et al.*, 2006
L. suebicus KCTC 3549	*L. vac*	Apple mash	21.06.2011	2.7	Draft	2534	368	58	1	Nam *et al.*, 2011d
L. sp. 7_1_47FAA	-	Human GIT	03.10.2011	1.3	Draft	1181	20	76	1	Nelson *et al.*, 2010

Appendix Table 3. Summary of sequencing status of *Lactobacillus* genome projects.

Organism	Sequencing platform	Assembler	Gene calling	Functional Annotation	ACC
L. farciminis KCTC 3681	454	Newbler	RAST/Glimmer/RNAmmer/ tRNAscan-SE	RAST/BLAST/COG	AEOT01000000
L. versmoldensis KCTC 3814	454	Newbler	RAST/Glimmer/BLAST/ RNAmmer	BLAST	BACR01000000
L. brevis ATCC 27305	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGG01000000
L. brevis ATCC 367	Sanger	Jazz	GeneMarkS/tRNAscan-SE	COG/PSI-BLAST	CP000416
L. buchneri ATCC 11577	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGH01000000
L. buchneri NRRL B-30929	454/Illumina/Sanger	Newbler	IMG	IMG	CP002652
L. hilgardii ATCC 8290	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGP01000000
L. kisonensis F0435	Illumina	Velvet	GeneMark/Glimmer/BLAST/ tRNAscan-SE/RNAmmer/Rfam	BLAST/BER/KEGG	AGRJ01000000
L. parafarraginis F0439	Illumina	Velvet	GeneMark/Glimmer/BLAST/ tRNAscan-SE/RNAmmer/Rfam	BLAST/BER/KEGG	AGEY01000000
L. casei ATCC 334	Sanger	Jazz	GeneMarkS/tRNAscan-SE	COG/PSI-BLAST	CP000423
L. casei BD-II	454/Illumina/Sanger	Newbler	-	-	CP002618
L. casei BL23	454	Phrap	AGMIAL	AGMIAL	FM177140
L. casei LC2W	454/Illumina/Sanger	-	-	-	CP002616
L. casei Zhang	-	Phrap	-	-	CP001084
L. paracasei 8700:2	454	-	Broad Institute	JCVI	NZ_DS990485
L. paracasei ATCC 25302	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGY01000000
L. rhamnosus ATCC 21052	Illumina	Velvet	GeneMark/Glimmer/BLAST/ tRNAscan-SE/RNAmmer/Rfam	BLAST/BER/KEGG	AFZY00000000
L. rhamnosus ATCC 8530	454/Sanger	Newbler	IGS Annotation Engine	IGS Annotation Engine	CP003094

L. rhamnosus GG	454/Sanger	Newbler	Glimmer/tRNA-scan-SE/ RNAmmer/ERGO	BLAST/InterPro/TransAAP/ MEROPS/CaZy/TCDB/KAAS	FM179322
L. rhamnosus GG	Sanger	Phrap	Glimmer/BLAST/EMBOSS	-	AP011548
L. rhamnosus HN001	-	-	PGAAP	PGAAP	ABWJ01000000
L. rhamnosus LC705	454/Sanger	Newbler	Glimmer/tRNA-scan-SE/ RNAmmer	BLAST/InterPro/TransAAP/ MEROPS/CaZy/TCDB/KAAS	FM179323
L. rhamnosus LMS2-1	454/Illumina/Sanger	Newbler	BCM	JCVI	ACIZ01000000
L. rhamnosus MTCC 5462	454	Newbler	PGAAP	PGAAP	AEYM01000000
L. rhamnosus R0011	454	Newbler	PGAAP	PGAAP	AGKC01000000
L. zeae KCTC 3804	454	Newbler	RAST/Glimmer/tRNAscan-SE/RNAmmerm/BLAST	BLAST	BACQ01000000
L. acidophilus 30SC	454/Sanger	Newbler	RAST/PGAAP	RAST/PGAAP	CP002559
L. acidophilus ATCC 4796	454/Illumina/Sanger	Newbler	BCM	JCVI	ACHN01000000
L. acidophilus NCFM	Sanger	Phrap	Glimmer/tRNAscan-SE	GAMOLA/COG	CP000033
L. amylolyticus DSM 11664	454	Newbler	BCM	JCVI	ADNY01000000
L. amylovorus GRL 1112	454/Sanger	Phrap/Newbler	GeneMark/Glimmer/BLAST/ PGAAP	PGAAP	CP002338
L. amylovorus GRL1118	454/Sanger	Gap4	GeneMark/Glimmer/BLAST/ PGAAP	PGAAP	CP002609
L. crispatus 125-2-CHN	454	Newbler	GeneMark/Glimmer/Metagene / BLAST	BLAST/PFAM	NZ_GG698760
L. crispatus 214-1	454	-	JCVI	JCVI	ADGR01000000
L. crispatus CTV-05	454/Sanger	-	Broad Institute	JCVI	NZ_GL531736
L. crispatus JV-V01	454/Sanger	Newbler	BCM	JCVI	ACKR01000000
L. crispatus MV-1A-US	454	Newbler	GeneMark/Glimmer/Metagene	BLAST/PFAM	NZ_GG698827
L. crispatus MV-3A-US	454	Newbler	Broad Institute	JCVI	NZ_GG704606
L. crispatus SJ-3C-US	454	Newbler	GeneMark/Glimmer/Metagene / BLAST	BLAST/PFAM	ADDT01000000

L. crispatus ST1	454/Sanger	Phrap/Newbler	Glimmer/tRNA-scan-SE/RNAmmer	Blannotator/RAST/InterPro/MEROPS/TCDB/KAAS	FN692037
L. delbrueckii 2038	Sanger	Phrap	Glimmer	BLAST/RPS-BLAST/ScanProsite	CP000156
L. delbrueckii ATCC 11842	Sanger	-	AGMIAL/ERGO	Agmial/ERGO	CR954253
L. delbrueckii ATCC BAA-365	Sanger	Jazz	GeneMarkS/tRNAscan-SE	COG/PSI-BLAST	CP000412
L. delbrueckii CNCM I-1519	454	Newbler	-	BLASTP	AGHW00000000
L. delbrueckii CNCM I-1632	454	Newbler	-	BLASTP	AGFO00000000
L. delbrueckii DSM 20072	454	Newbler	BCM	JCVI	AEXU01000000
L. delbrueckii ND02	454/Illumina/Sanger	Newbler	-	-	CP002341
L. delbrueckii PB2003/044-T3-4	454	Newbler	JCVI	JCVI	AEAT01000000
L. gasseri 202-4	454	-	JCVI	JCVI	ACOZ01000000
L. gasseri 224-1	454	-	JCVI	JCVI	ADFT01000000
L. gasseri ATCC 33323	Sanger	Jazz	GeneMarkS/tRNAscan-SE	COG PSI-BLAST	CP000413
L. gasseri JV-V03	454	Newbler	BCM	JCVI	ACGO02000000
L. gasseri MV-22	454	-	Broad Institute	JCVI	NZ_DS995854
L. gasseri SJ-9E-US	454	Newbler	GeneMark/Glimmer/Metagene	BLAST/PFAM	ADDU01000000
L. gasseri SV-16A-US	454	Newbler	GeneMark/Glimmer/Metagene	BLAST/PFAM	ADDY01000000
L. helveticus DPC 4571	Sanger	Staden	Gamola/ERGO/Glimmer/tRNAscan-SE	ERGO/BLAST	CP000517
L. helveticus DSM 20075	454	Newbler	BCM	JCVI	ACLM01000000
L. helveticus H10	454/Illumina/Sanger	-	-	-	CP002429
L. helveticus MTCC 5463	454	Newbler	PGAAP	PGAAP	AEYL01000000
L. iners AB-1	454/Illumina/Sanger	Minimus	RAST/GeneMark/Glimmer/tRNAscan-SE	RAST/BLAST/PFAM/KAAS/SLEP/TransAAP	NZ_ADHG01000000
L. iners ATCC 55195	454	Newbler	BCM	JCVI	AEPX01000000
L. iners DSM 13335	454	Newbler	BCM	JCVI	ACLN01000000
L. iners LactinV 01V1-a	454	Newbler	JCVI	JCVI	AEHQ01000000

L. iners LactinV 03V1-b	454	Newbler	JCVI	JCVI	AEHP01000000
L. iners LactinV 09V1-c	454	Newbler	JCVI	JCVI	AEHO01000000
L. iners LactinV 11V1-d	454	Newbler	JCVI	JCVI	AEHN01000000
L. iners LEAF 2052A-d	454	Newbler	JCVI	JCVI	AEKI01000000
L. iners LEAF 2053A-b	454	Newbler	JCVI	JCVI	AEKH01000000
L. iners LEAF 2062A-h1	454	Newbler	JCVI	JCVI	AEKJ01000000
L. iners LEAF 3008A-a	454	Newbler	JCVI	JCVI	AEKK01000000
L. iners SPIN 1401G	454	Newbler	JCVI	JCVI	AEXP01000000
L. iners SPIN 2503V10-D	454	Newbler	JCVI	JCVI	AEHR01000000
L. iners UPII 143-D	454	Newbler	JCVI	JCVI	AEXJ01000000
L. iners UPII 60-B	454	Newbler	JCVI	JCVI	AEXK01000000
L. jensenii 115-3-CHN	454	Newbler	Broad Institute	JCVI	NZ_GG704741
L. jensenii 1153	454	HybridAssembler	Broad Institute	JCVI	ABWG02000000
L. jensenii 269-3	454	-	JCVI	JCVI	ACOY01000000
L. jensenii 27-2-CHN	454	-	GeneMark/Glimmer/Metagene / BLAST	BLATP/PFAM	NZ_GG698814
L. jensenii JV-V16	454	Newbler	BCM	JCVI	ACGQ02000000
L. jensenii SJ-7A-US	454	-	Broad Institute	JCVI	NZ_GG704682
L. johnsonii ATCC 33200	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGR01000000
L. johnsonii DPC 6026	454/Sanger	Phrap/Newbler	Gamola/Glimmer/RAST	BLAST/PROSITE/RBS finder	CP002464
L. johnsonii FI9785	454/Sanger	Phrap/Staden	RBSfinder/REGANOR5/CRITICA/ GenDB/Glimmer/tRNAscan-SE	BLAST/PFAM/InterPro	FN298497
L. johnsonii NCC 533	Sanger	Phrap	FrameD/BLAST/tRNAscan-SE	BLAST/PFAM/COG/TCDB	AE017198
L. johnsonii pf01	454	Newbler/ CodonCode	RAST/Glimmer	RAST	AFQJ01000000

L. kefiranofaciens ZW3	454/Illumina/Sanger	Phrap/Newbler	Glimmer/GeneMark	-	CP002764
L. ultunensis DSM 16047	454/Sanger	Newbler	BCM	JCVI	ACGU01000000
L. fructivorans KCTC 3543	454	Newbler	RAST/Glimmer/RNAmmer/ tRNAscan-SE	RAST/BLAST/COG	AEQY01000000
L. sanfranciscensis TMW 1.1304	454/Sanger	-	PEDANT/GenMark/Glimmer/ CRISPRFinder	PEDANT/BLAST	CP002461
L. malefermentans KCTC 3548	454	Newbler	RAST/Glimmer/RNAmmer/ tRNAscan-SE	BLAST/COG	BACN01000000
L. pentosus MP-10	454	Newbler	-	BLAST/InterPro	FR871817
L. plantarum ATCC 14917	454	Newbler	BCM	JCVI	ACGZ02000000
L. plantarum JDM1	454/SOLiD/Sanger	Phrap/Newbler	-	-	CP001617
L. plantarum NC8	454/Illumina	Newbler	IGS	BLAST/InterPro	AGRI00000000
L. plantarum ST-III	454/Illumina/Sanger	Newbler	-	-	CP002222
L. plantarum WCFS1	Sanger	-	-	-	AL935263
L. antri DSM 16041	454/Illumina/Sanger	Newbler	BCM	JCVI	ACLL01000000
L. coleohominis 101-4-CHN	454	Newbler	GeneMark/Glimmer/Metagene	BLAST/PFAM	NZ_GG698802
L. coryniformis KCTC 3167	454	Newbler	RAST/Glimmer/RNAmmer/ tRNAscan-SE	RAST/COG/BLAST	AELK01000000
L. coryniformis KCTC 3535	454	Newbler	PGAAP	PGAAP	AEOS01000000
L. fermentum 28-3-CHN	454	Newbler	Broad Institute	JCVI	NZ_GG704699
L. fermentum ATCC 14931	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGI01000000
L. fermentum CECT 5716	454	Newbler	-	-	CP002033
L. fermentum IFO 3956	Sanger	Phrap	Glimmer/BLAST/EMBOSS/ tRNAscan-SE	BLAST/PFAM/COG	AP008937
L. gastricus PS3	454	Newbler	-	-	AICN00000000

L. mucosae LM1	454/Illumina	CLCbio/Newbler/CodonCode	RAST	RAST/BLAST	AICN00000000
L. oris F0423	454	CeleraAssembler	JCVI	JCVI	AFTL01000000
L. oris PB013-T2-3	454	Newbler	JCVI	JCVI	AEKL01000000
L. reuteri 100-23	454/Sanger	-	IMG	IMG	AAPZ02000000
L. reuteri ATCC 53608	454	Newbler	Glimmer3/GeneMark	BLAST/InterPro	CACS02000000
L. reuteri CF48-3A	454/Illumina/Sanger	Newbler	BCM	JCVI	ACHG01000000
L. reuteri DSM 20016	454/Sanger	-	IMG	IMG	CP000705
L. reuteri JCM 1112	Sanger	Phrap	Glimmer/BLAST/Emboss/tRNAscan-SE	BLAST/PFAM/COG	AP007281
L. reuteri lpuph	454	Newbler	-	-	AEAX01000000
L. reuteri mlc3	454	Newbler	-	-	AEAW01000000
L. reuteri MM2-3	454/Sanger	Newbler	BCM	JCVI	ACLB01000000
L. reuteri MM4-1A	454	Newbler	BCM	JCVI	ACGX02000000
L. reuteri SD2112	454	Newbler	BCM	JCVI	CP002844
L. vaginalis ATCC 49540	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGV01000000
L. curvatus CRL 705	454	Newbler	ISGA/RAST/RAST/Glimmer/tRNAscan-SE/RNAmmer	ISGA/RAST	AGBU01000000
L. sakei 23K	Sanger	Phrap	AGMIAL	AGMIAL/InterPro/PFAM	CR936503
L. acidipiscis KCTC 13900	454	Newbler	PGAAP	PGAAP	BACS01000000
L. animalis KCTC 3501	454	Newbler	Glimmer/RNAmmer/RAST/tRNAscan-SE	BLAST	AEOF01000000
L. mali KCTC 3596	454	Newbler	RAST/Glimmer/RNAmmer/tRNAscan-SE	BLAST/COG	BACP01000000
L. ruminis ATCC 25644	454	Newbler	BCM	JCVI	ACGS02000000
L. ruminis SPM0211	454/Illumina/Sanger	CLCbio/Ne	-	-	AFOJ01000000

		wbler			
L. salivarius ACS-116-V-Col5a	454	Newbler	JCVI	JCVI	AEBA01000000
L. salivarius ATCC 11741	454/Illumina/Sanger	Newbler	BCM	JCVI	ACGT01000000
L. salivarius CECT 5713	454	Newbler	-	-	CP002034
L. salivarius GJ-24	454/Illumina	CLCbio/Newbler	RAST	RAST/KEGG/COG	AFOI01000000
L. salivarius NIAS840	454/Illumina	CLCbio/Newbler	RAST/BLAST	RAST	AFMN01000000
L. salivarius SMXD51	454	MIRA	RAST/Glimmer	RAST	AICL00000000
L. salivarius UCC118	Sanger	Phrap	ERGO/YACOP/glimmer/Zcurve /Orpheus/Critica/tRNAscan-SE	ERGO/COG/TCDB	CP000233
L. suebicus KCTC 3549	454	Newbler	RAST/Glimmer/tRNAscan-SE/RNAmmerm BLAST	BLAST	BACO01000000
L. sp. 7_1_47FAA	454	Newbler	Prodigal	JCVI	ACWR01000000