View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Helsingin yliopiston digitaalinen arkisto

Viikki Doctoral Programme in Molecular Biosciences and
Integrative Life Science Doctoral Programme, Doctoral School in Health Sciences and
Institute of Biotechnology and
Division of General Microbiology, Department of Biosciences,
Faculty of Biological and Environmental Sciences
University of Helsinki
Helsinki

Lactobacillus crispatus and *Propionibacterium freudenreichii*: A Genomic and Transcriptomic View

Teija Ojala

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 lecture room 228, Koetilantie 5, Viikki, on January 22nd 2016, at 12 o'clock noon.

Helsinki 2016

Supervisors

Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Finland

Professor Liisa Holm
Institute of Biotechnology and
Department of Biosciences
University of Helsinki
Finland

Thesis Advisory Committee Members

Professor Benita Westerlund-Wickström
Department of Biosciences
University of Helsinki
Finland

Professor Kaarina Sivonen
Department of Food and Environmental
Sciences
University of Helsinki
Finland

Reviewers

Assistant Professor Mirko Rossi
Department of Food Hygiene and
Environmental Health
University of Helsinki
Finland

Professor Per Saris
Department of Food and Environmental
Sciences
University of Helsinki
Finland

Opponent

Docent Samuel Myllykangas
Department of Biosciences
University of Helsinki
Finland

Custos

Professor Sarah Butcher
Department of Biosciences
University of Helsinki
Finland

Published in *Dissertationes Scholae Doctoralis Ad Sanitatem Investigandam Universitatis Helsinkiensis*

ISBN 978-951-51-1843-1 (pbk.)
ISSN 2342-3161

ISBN 978-951-51-1844-8 (PDF)
ISSN 2342-317X
http://ethesis.helsinki.fi

Hansaprint Oy
Vantaa 2015

Abstract

Lactobacilli and propionibacteria are Gram-positive bacteria with major implications for human lives. The genera *Lactobacillus* and *Propionibacterium* both include industrially relevant members as well as inhabitants of a human host. However, the molecular mechanisms underlying their industrially desirable properties or microbe-host-interactions are relatively poorly understood. Attractive means to advance the understanding of these traits are offered by modern sequencing technologies, which have opened up numerous possibilities to investigate the bacterial physiology and ecology, to delineate phenotype-genotype relationships, and to identify the factors underlying bacterial traits as well as their significance in different conditions. In this thesis, an important member of the vaginal normal flora, *Lactobacillus crispatus*, and a central dairy culture, *Propionibacterium freudenreichii*, were studied using modern sequencing technologies and subsequent analyses to uncover how these particular bacteria can live in and interact with their environments.

Annotated whole-genome sequences of a representative member of each of the species *L. crispatus* and *P. freudenreichii* were produced and subjected to further functional genomics investigations. Specifically, the sequenced genome of *L. crispatus* ST1, a chicken isolate, was analyzed in conjunction with publically available genome sequences of nine vaginal *L. crispatus* isolates. These analyses were performed to determine the scale and scope of the genetic variation within the species and to identify species-wide mechanisms by which this prominent member of the human vaginal flora may promote urogenital health. The cheese starter, *P. freudenreichii* ssp. *shermanii* JS, in turn, was subjected to transcriptome and genome sequencing to gain a deeper understanding of the role of this bacterium in industrial cheese ripening. Given that an important part of the afore-mentioned analyses involved understanding the biological function of the predicted protein-coding genes, various methods for the functional annotation of bacterial proteins were also tested and evaluated.

The comparative genomics analysis of a total of ten *L. crispatus* isolates revealed the common genetic backbone of *L. crispatus*, providing the first insights into the collective molecular mechanisms of this species. The common core genome of the ten analyzed isolates comprised 1,224 ortholog groups, whereas the isolates harbored 2,705 ortholog groups in total. Extrapolations of these core and pan-genome data suggested the common features of these isolates to offer a rather good representation of the species-wide core proteins, whereas additional *L. crispatu*s genome sequences will be required to fully capture the genetic variation within the species. Notably, several of the detected *L. crispatus* core features were predicted to be of potential importance to vaginal health. Among these features was a previously characterized adhesin, which was in this thesis identified as a likely antagonist to the harmful vaginal bacterium *Gardnerella vaginalis*. Importantly, antibody fragments specific for this adhesin efficiently inhibited also *G. vaginalis* adhesion to human cell line, indicating the significance of *L. crispatus* core proteins in the protective function that this species is considered to have in the vagina.

In turn, the genomic and transcriptomic analysis of *P. freudenreichii* ssp. *shermanii* JS shed light on the flavor-forming abilities of this strain at the different stages of cheese

ripening. Specifically, the study revealed the genome of strain JS to be highly similar to those of other *P. freudenreichii* strains and revealed several enzymes and metabolic pathways involved in the formation of flavor compounds, such as propionate and acetate. Sequencing of the transcriptomes extracted from industrial cheese samples revealed nearly 15% of the 2,377 protein-coding genes of strain JS to be significantly differentially expressed between the warm and the subsequent cold ripening stages of cheese manufacture. A notable portion of the genes up-regulated in the cold were associated with mobile genetic elements, whereas several of the flavor-associated genes exhibited higher expression levels in the warm than the cold, corroborating the hypothesis that *P. freudenreichii* contributes more to the cheese flavor during the warm ripening than cold ripening period.

Automated function prediction of the bacterial proteins greatly facilitated the genomics investigations of *L. crispatus* and *P. freudenreichii.* The different methods provided functional descriptions for ~77% and ~88% of the proteins predicted to be encoded in the newly sequenced genomes of strains ST1 and JS, respectively. Moreover, re-annotation of the *L. crispatus* proteomes included in the comparative genomics study notably increased the portion of *L. crispatus* proteins with functional descriptions. The different methods varied in their prediction capabilities and were therefore complementary. These results support the use of more than one function prediction method in a bacterial genome project. Moreover, the performance evaluation of different annotation strategies using more standardized test data support the use of annotation strategies that base their functional descriptions on more than one hit. Interestingly, such strategies were noted to benefit from the avoidance of extremely strict thresholds in the homology searches used for functional annotation. Strict thresholds were observed to unnecessarily restrict the pool of hits available for annotation transfer, hampering both the annotation quality and the fraction of functionally annotated proteins.

Taken together, the utilized sequencing approaches coupled with suitable downstream analyses proved effective in deciphering the physiology of lactobacilli and propionibacteria and offered novel insights into the urogenitally important properties of *L. crispatus* and the flavor-forming capabilities of *P. freudenreichii*.

Contents

Abstract

List of original publications

Abbreviations

1 Introduction 1

2 Review of the Literature 3

2.1 Sequencing-based Characterization of Bacteria 3

2.1.1 Sequencing Technologies 5

2.1.2 Resolving Bacterial Genome Sequences 9

2.1.3 Bacterial Genome Annotation 11

2.1.3.1 Structural Annotation 11

2.1.3.2 Functional Annotation 13

2.1.4 Comparative Bacterial Genomics 15

2.1.5 Resolving Bacterial Transcriptomes via Sequencing 17

2.2 Lactobacilli 20

2.2.1 *Lactobacillus* genomes 22

2.2.2 Comparative Genomics of Lactobacilli 23

2.2.3 Sequencing-based Investigations of Vaginal Lactobacilli 26

2.3 Propionibacteria 29

2.3.1 *Propionibacterium* Genomes 30

2.3.2 Comparative Genomics of Propionibacteria 31

2.3.3 Transcriptomics of Propionibacteria 33

3 Aims of the Study 35

4 Materials and Methods 36

5 Results and Discussion 38

5.1 Genomics of *L. crispatus* ST1 39

5.2 Comparative Genomics of *L. crispatus* 41

5.3 Genomics of *P. freudenreichii* ssp. *shermanii* JS 45

5.4 Activity of *P. freudenreichii* ssp. *shermanii* JS during Cheese Ripening 48

5.5 Annotation Transfer Evaluation 50

6 Conclusions 53

Acknowledgements 55

References 56

List of original publications

This thesis is based on the following publications:

I **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

II Kankainen M, **Ojala T**, Holm L. (**2012**) Blannotator – a practical method for homology based annotations. *BMC bioinformatics* 13:33

III **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* suggest novel mechanisms for the competitive exclusion of *Gardnerella vaginalis*. *BMC Genomics* 15:1070

IV **Ojala T,** Laine PK, Ahlroos T, Tanskanen J, Pitkänen S, Salusjärvi T, Kankainen M, Tynkkynen S, Paulin L, Auvinen P. Functional genomics provides insights into the role of *Propionibacterium freudenreichii* ssp. *shermanii* JS in cheese ripening. *Submitted manuscript.*

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

Abbreviations

bp	base pair
BLAST	basic local alignment search tool
BV	bacterial vaginosis
Cas	CRISPR-associated protein
CDD	conserved domain database
cDNA	complementary-DNA
CDS	coding DNA sequence
COG	cluster of orthologous groups
CRISPR	clustered regularly interspaced short palindromic repeat
cfu	colony forming unit
DE	description line
EC	enzyme commission
EMPP	Embden-Meyerhof-Parnas pathway
EPS	exopolysaccharide
GI	genomic island
GIT	gastrointestinal tract
GO	gene ontology
GRAS	generally regarded as safe
GUT	genitourinary tract
HGT	horizontal gene transfer
HMM	hidden Markov model
HMP	human microbiome project
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAB	lactic acid bacteria
LEA	*Lactobacillus* epithelium adhesin
mRNA	messenger-RNA
NGS	next generation sequencing
RNA-seq	RNA-sequencing
RPKM	reads per kilobase per million mapped reads
rRNA	ribosomal RNA
tRNA	transfer RNA
PCR	polymerase chain reaction
PPP	pentose phosphate pathway
PTS	phosphotransferase system
RAST	rapid annotation using subsystem technology
RBS	ribosomal binding site
SNP	single nucleotide polymorphism
ssp.	subspecies
TC	transport comission

1 Introduction

Although individually miniscule in size, the combined biomass of prokaryotic cells is estimated to rival that of the vegetation on Earth (Whitman et al. 1998). Divided into Bacteria and Archaea (Figure 1), prokaryotes perform many vital tasks in the biosphere and have a profound effect on our everyday lives. Prokaryotes play a fundamental part in geochemical processes such as nitrogen (Canfield et al. 2010) and carbon cycling (Shively et al. 2001), and have a major role, for example, in food production (Caplice & Fitzgerald 1999). In addition, prokaryotes are an inherent part of our own bodies, and the bacterial cells in and on the human body outnumbering the human cells by a factor of ten (Wilson 2008). These bacteria have central responsibilities relating to food digestion and absorption, vitamin synthesis, development of the human immune system, and resistance against pathogens (McFarland 2000, Maynard et al. 2012, Hooper et al. 2012).

Research on this important group of life forms has been revolutionized by the tremendous advancements in sequencing technologies and downstream analysis methods, which provide universal tools to study the characteristics and functions of bacteria and other organisms (Forde & O'Toole 2013, Medini et al. 2008, Hall 2007, Loman et al. 2012). The advancements have made the determination of the whole genome sequence of an organism feasible, enabling previously unimaginable insights into the characteristics and biology of the isolate or bacterial group

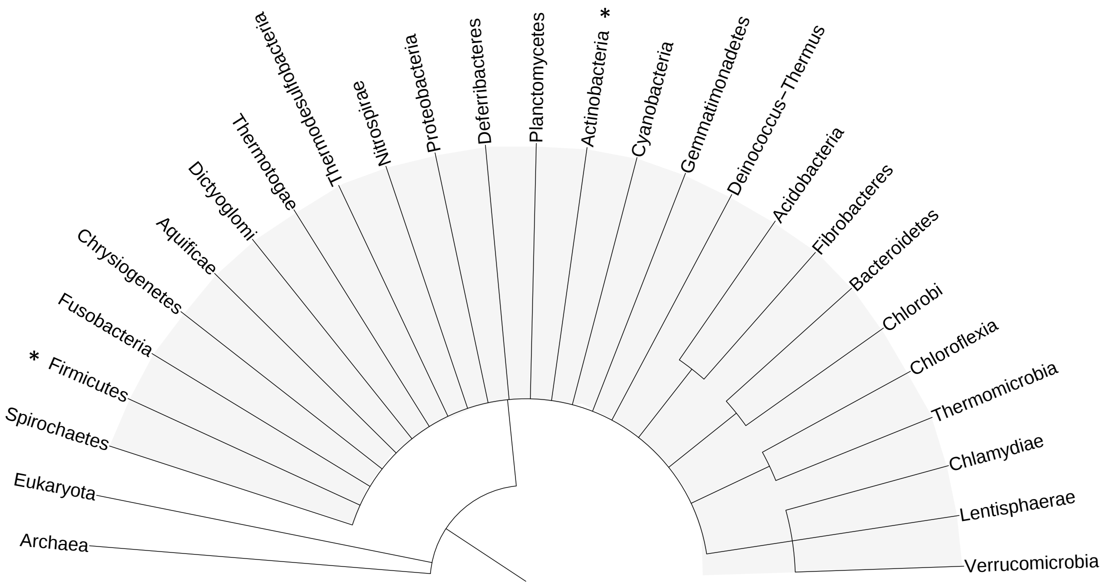


Figure 1 *Tree of life. In this tree, phyla within the domain Bacteria (grey background) are shown along the domains/superkingdoms Eukaryota and Archaea. The bacterial phyla were chosen according to the Taxonomic Outline of Bacteria and Archaea (Garrity et al. 2007b), and the tree is based on the NCBI taxonomy. The phylum* Chloroflexi *is represented by its clas*s Chloroflexia. *The tree was constructed using phyloT (database version 2015,1) and visualized with iTol (Letunic & Bork 2007). The asterisks mark the phyla most relevant to this thesis.*

of interest (Hall 2007, Medini et al. 2008, Forde & O'Toole 2013, Fleischmann et al. 1995). Whole-genome sequencing and subsequent functional genomics analyses have, for instance, shed light on the molecular mechanisms that underpin bacterial food fermentation processes (Smid & Hugenholtz 2010) or the harmful (Pallen & Wren 2007) or beneficial properties of bacteria (Ventura et al. 2009). Technological advancements have also enabled the detailed genomic characterization of uncultivable microorganisms (Simon & Daniel 2011) as well as the identification and quantification of the expression of genes in a given sample (van Vliet 2010). These analyses add a new layer of information regarding the properties and functions of bacteria. For example, transcriptome profiling has provided cues as to how bacteria respond and adjust to different environments (Yoder-Himes et al. 2009) and has revealed antibiotic resistance mechanisms that were unidentifiable via traditional or genome-based methods (Haaber et al. 2015).

In this thesis, modern sequencing technologies and subsequent functional genomics analyses were employed to provide novel genomic and functional information on two important bacterial species, *Lactobacillus crispatus* and *Propionibacterium freudenreichii* of the phyla *Firmicutes* and *Actinobacteria*, respectively (Figure 1).

2 Review of the Literature

2.1 Sequencing-based Characterization of Bacteria

Determining the genome sequence of an organism opens new avenues in deciphering the phenotypic characteristics, physiology, and ecology of the organism in question. This approach has been particularly popular in extending the genomic knowledge of bacteria (Figure 2A). After the first complete bacterial genome sequence was published in 1995 (Fleischmann et al. 1995), the number of bacterial genome sequences in public databases began to gradually increase until the massively parallel sequencing methods termed next generation sequencing (NGS) techniques (Mardis 2008, Shendure et al. 2005, Margulies et al. 2005) became available. This advance resulted in an explosion of bacterial genome information. In addition to providing invaluable information and testable explanations on the molecular biology of individual bacterial isolates (Kankainen et al. 2009, Crossman et al. 2008, Kube et al. 2013), genome sequencing and genomic comparisons between the members of the same species or more distant relatives have illuminated evolutionary events associated with, for example, the emergence of virulence (Reuter et al. 2014, Langridge et al. 2015) or niche specification (Broadbent et al. 2012, Smokvina et al. 2013). Furthermore, comparisons of isolates belonging to the same bacterial group has revealed (i) that the genetic repertoire of a given bacterial group is much larger than the gene content of any of the individual isolates and (ii) that horizontal gene transfer (HGT) from other bacteria has been one of the driving forces in the evolution of bacteria (Tettelin et al. 2005, Tettelin et al. 2008, Medini et al. 2005, Broadbent et al. 2012, Bentley 2009).

Moreover, modern sequencing technologies have been applied in sequencing the genomic DNA extracted from natural bacterial communities. These metagenomics studies offer a more in-depth view of the bacterial communities than is obtainable via sequencing of biomarker genes amplified from the sample (Forde & O'Toole 2013). Metagenomics studies have revealed, for instance, that the bacterial communities residing in and on the human body have a vast functional potential to process and biosynthesize a wide variety of substances (Qin et al. 2010, Oh et al. 2014) and act as important sources of antibiotic resistance genes (Forslund et al. 2013, Oh et al. 2014).

Lastly, deciphering the active parts of genomes can reveal important insights into the characteristics and properties of bacteria as well as into the biological functions in which they participate. A promising approach in this functional characterization is to examine the expression of genes. Interrogation of the whole bacterial transcriptomes has its roots in the hybridization-based studies (Selinger et al. 2000, Cho et al. 2013); however, since their emergence, modern sequencing technologies have become increasingly popular in the generation of bacterial genome-wide expression data

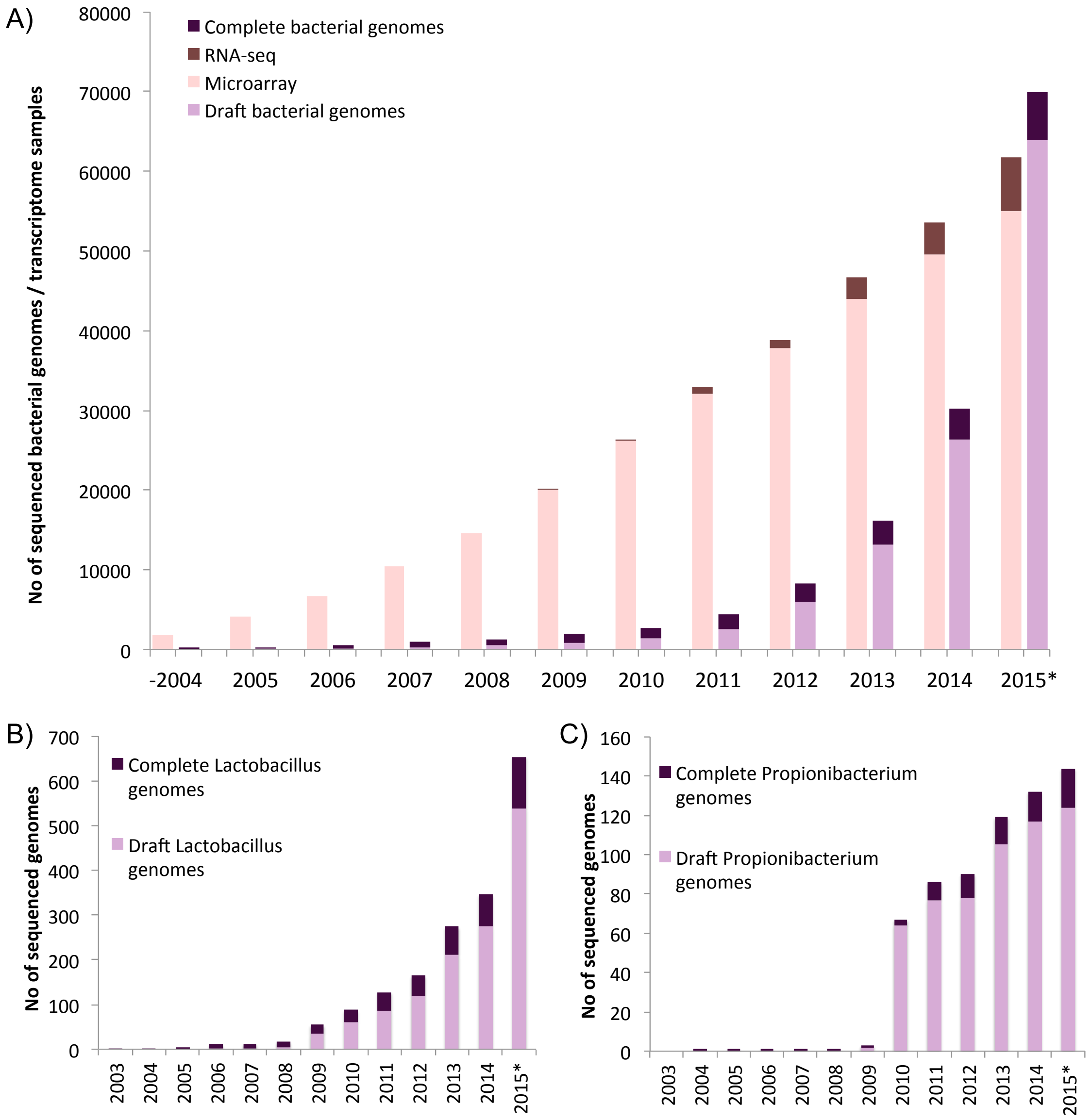


Figure 2 *Accumulation of bacterial genome and transcriptome data in NCBI. The cumulative number bacterial genome entries and bacterial transcriptome samples (A) and the cumulative number of* Lactobacillus *(B)* and Propionibacterium *genome entries (C) are given in the charts. The number of complete (complete genomes and chromosomes) and draft genome (contigs and scaffolds) assemblies, as well as the number of microarray and RNA-seq samples, are represented by different colors (given in the legends). The genome entry data were retrieved from GenBank (Benson et al. 2013) and the transcriptome sample data from Gene Expression Omnibus (GEO) (Barrett et al. 2013) in September 2015. *Numbers for 2015 were extrapolated from the existing data under the assumption that the average number of data entries per month will constant throughout the year 2015.*

(van Vliet 2010) (Figure 2A). For example, sequencing of the bacterial transcriptomes has unveiled the infection-related adaptation strategies of facultative (Mandlik et al. 2011) or opportunistic (Jorth et al. 2013) pathogens. In addition, the sequencing of random RNA fragments extracted from community samples (*i.e.,* metatranscriptomics) has revealed community-wide bacterial gene expression responses to a variety of conditions, such as bacterial vaginosis (BV) (Macklaim et al. 2013), periodontitis (Jorth et al. 2014), antibiotics and other xenobiotics (Maurice et al. 2013), and different diets (David et al. 2014). In addition, transcriptome sequencing is useful in aiding the accurate detection of genes (Sorek & Cossart 2010) and operons (Sharma et al. 2010).

With the increasing amounts of genome and transcriptome data (Figure 2), the cornerstone of modern bacterial genomics can be considered to be the extraction of biological information encoded in the DNA or RNA sequence. This is a multi-level process that involves various types of bioinformatics analyses. For example, the analysis of a genome sequence involves the identification of sequence features (structural annotation) and connecting them with relevant biological information and processes (functional annotation) (Stein 2001, Reed et al. 2006). In turn, the main tasks of transcriptome data analysis include quantifying the expression levels of genetic elements and identifying features that are differentially expressed between different conditions (Oshlack et al. 2010). In the following subsections, the aspects of bacterial genomics and transcriptomics most relevant to this thesis are discussed in more detail.

2.1.1 Sequencing Technologies

For nearly 30 years, DNA sequencing methods based on the Sanger procedure (Sanger *et al.* 1977) were vastly more common than other sequencing techniques as the method of choice in sequencing projects (Schuster 2008, Hutchison 2007). This dominance began to abate just a decade ago, when the more efficient NGS technologies emerged and enabled faster and more affordable DNA sequencing than was previously possible (Schuster 2008, Metzker 2005, Margulies et al. 2005, Shendure et al. 2005). Although the power of the first NGS method to come to market was demonstrated by the *de novo* generation of a *Mycoplasma genitalium* genome (Margulies et al. 2005), NGS methods were initially considered to be better suited for re-sequencing of bacterial genomes and widely applied in *de novo* sequencing projects only after NGS methods were capable of producing longer read lengths (MacLean et al. 2009, van Dijk et al. 2014). Despite having been replaced by other methods in wide-scale sequencing projects, Sanger-based technologies still hold a place in smaller scale sequencing, such as sequencing polymerase chain reaction (PCR) products and aiding gap closing in whole genome sequencing projects.

DNA sequencing with the Sanger procedure is based on the use of chain-terminating dideoxy-nucleotides in the synthesis of complementary strands to the template (Sanger et al. 1977). With appropriate mixture of the deoxy- and dideoxy-nucleotides, the latter of which lack the 3’-OH group required for the incorporation of the next nucleotide, the synthesized DNA strands differ in length from each other by one nucleotide. After denaturation, the strands are separated by gel or capillary electrophoresis and the last base incorporated is identified

(Hutchison 2007). Originally, this identification was based on performing chain elongation in four separate reactions, each containing just one type of chain-terminating nucleotide, and the use of a radioisotope-labeled primers (Sanger et al. 1977). However, the development of terminators labeled with fluorescent dyes and their laser-based identification systems have enabled pooled reactions and have sped the process notably (Hutchison 2007).

Compared even to the automated Sanger-based sequencing, the NGS platforms are advantageous in their ability to deliver massive amounts of sequence data rapidly and at thousands of times lower cost (Liu et al. 2012). The high throughput of the sequence data is the result of the massive parallelization of the sequencing, a common feature to all of the NGS technologies. However, speed and high throughput often comes at the cost of accuracy and length of the generated reads (Liu et al. 2012, Hutchison 2007), although the read length of some NGS instruments far exceeds those of the Sanger-based technologies (van Dijk et al. 2014, Liu et al. 2012). The read length, sample preparation, sequencing chemistry, and other specific features vary between the different NGS platforms, affecting their suitability for different applications (Table 1).

The workflows of different NGS technologies resemble each other. The NGS workflows usually begin with the conversion of the source DNA or RNA into a sequencing library of size-selected molecules with platform-specific adaptors in both ends, after which the library is either amplified and subjected to sequencing or sequenced directly (Loman et al. 2012, van Dijk et al. 2014) (Figure 3). The 454 (Margulies et al. 2005), Ion Torrent (Rothberg et al. 2011), and, until recently, SOLiD (Shendure et al. 2005) technologies have relied on emulsion-PCR for library amplification. In this amplification method, the library fragments are amplified on water-in-oil microreactors, each containing a single bead-like particle having one DNA fragment bound on its surface. After the PCR amplification, the particles covered with the clonal copies of the template fragment are deposited onto a Picotiterplate (454), Semiconductor Sequencing Chip (Ion torrent), or glass slide (SOLiD) for sequencing. However, the SOLiD system recently replaced the emulsion-PCR with a “Wildfire” approach. In this approach, the library fragments are captured on the sequencing slide and amplified *in situ* with isothermal template walking, resulting in dense sequencing colonies. This amplification approach resembles that of Illumina technology (Bentley 2006, Bentley et al. 2008), in which the library fragments are denatured and immobilized on a solid surface covered with amplification primers and amplified *in situ* via bridge-PCR under isothermal conditions.

The different NGS technologies use different methods to sequence, in parallel, the millions of template clusters that are generated in the amplification step. The 454 technology (Margulies et al. 2005) utilizes pyrosequencing (Ronaghi et al. 1998, Ronaghi et al. 1996), in which the incorporation of a base in the elongating strand generates an inorganic pyrophosphate that is enzymatically converted to ATP and used in a light-producing enzymatic reaction. Each nucleotide species is flushed over the Picotiter plate one at a time in sequential order, and the light emission is monitored in real time after the nucleotide flush (Margulies et al. 2005). The Ion Torrent technology also monitors the incorporation of new bases to the synthesized strand, but differs from the 454 technology by detecting minor changes in pH caused by the release of a hydrogen ion (proton) during the base incorporation (Rothberg et al. 2011). Illumina technology (Bentley et al. 2008), on the other hand, is based on the sequencing-by-synthesis with reversible terminator nucleotides. Each nucleotide species has specific fluorescent label and a termination moiety, enabling the parallel analysis of the incorporation of all of the nucleotide species. The SOLiD technology (Shendure et al. 2005) is based on the ligation of fluorescently labeled oligonucleotides (octamers) to the template. First, a sequencing primer is hybridized to the adapter, and the octamers are then allowed to hybridize to the template. However, only those annealing perfectly are ligated to the primer. After imagining, the ligated octamer is cleaved between position 5 and 6, removing the label, and the cycle is repeated. To capture all of the bases, sequencing needs to be repeated with primers of different lengths.

The PacBio sequencing technology (Eid et al. 2009) differs from the above-mentioned NGS technologies as it does not require the sequencing library amplification step but is capable of single molecule sequencing (Figure 3, Table 1). In this system, single DNA-polymerases are immobilized within zero-mode waveguide detectors, and the incorporation of fluorescent-labeled nucleotides in the growing chain is detected in real time with continuous imagining. In addition to the identification of the incorporated nucleotide, this method allows the identification of the methylation modification status of the template base (Flusberg et al. 2010). The PacBio technology also ensures that the same insert can be read multiple times, which can improve sequencing accuracy. Consequently, the sequencing of large inserts increases the read length but lowers the accuracy because the large insert cannot be read as many times as the shorter ones.

Recently, a portable single-molecule sequencer, MinION, was launched by Oxford Nanopore Technologies (Table 1). This system is based on measuring ionic current passing through single nanometer-scale biological pores (nanopores), through which the template DNA is translocated one strand and one base at a time. The translocation of the DNA strand disrupts the flow of the ion current through the nanopore, and each base can be identified from the characteristic drop of the current (Ip et al. 2015, Mikheyev & Tin 2014). Similar to the PacBio sequencing, the library preparation in Nanopore technology does not require an amplification step. The PacBio and Nanopore technologies are occasionally referred to as "Third Generation Sequencing technologies" as they both are single-molecule sequencing technologies that capture the signal in real time (Schadt et al. 2010).

Table 1. *Summary of selected NGS platforms. This table is compiled from the data presented in Reuter et al. 2015, Liu et al. 2012, Loman et al. 2012, and van Dijk et al. 2014.*

	454, GS FLX+	SOLiD, 5500 W series	Illumina, HiSeq 2500	Ion PGM System, Ion 318 Chip v2	PacBio, RS II	Nanopore, MinION
Release of the first version	2005	2006	2006	2010	2010	2014
Amplification	Emulsion PCR	Isothermal template walking	Bridge PCR	Emulsion PCR	-	-
Sequencing chemistry	Pyrosequencing	Sequencing by ligation	Reversible dye termination	Ion semiconductor sequencing	Single molecule real time sequencing	Nanopore-based sequencing
Max read length (nt) for fragment-based run	1000	75	2 x 250	400	60000	60000
Advantages	Long read length, relatively fast run time	High throughput, low sequencing cost, accuracy	Highest throughput, best cost-effectiveness	Fast run times	Very long read lengths, fast run time	Very long read lengths, low machine cost
Disadvantages	Relatively low throughput, high reagent cost, errors in homopolymer repeats, planned discontinuation in 2016	Very short read length, long run time	Long run time, short read length, substitution errors	Errors in homopolymer repeats	High cost, high overall error rate, low throughput	High error rate, insertion/deletion errors, high run failure rate, low throughput
Most suitable microbiological applications	*de novo* genome sequencing, rapid draft *de novo* genome sequencing, 16S-based phylogenetic community analysis	Transcriptome sequencing, SNP detection	Draft *de novo* genome sequencing, genome re-sequencing, transcriptome sequencing, metagenomics	Rapid draft *de novo* genome sequencing	*de novo* genome sequencing, rapid draft *de novo* genome sequencing, transcriptomics	*de novo* genome sequencing, genome scaffolding

2.1.2 Resolving Bacterial Genome Sequences

Although some bacteria are known to harbor more than one chromosome, the bacterial genome typically consists of a single chromosome, which can be accompanied by smaller DNA molecules termed plasmids. Bacterial DNA molecules are most often circular, but both linear chromosomes and plasmids are known to exist (Casjens 1998). In bacterial whole-genome sequencing, the order of the nucleotides in all of the DNA molecules of the bacterium under interest is determined. These genome-sequencing projects generally start with the isolation of the genomic DNA, followed by the generation of sequencing libraries, amplification, size selection, and sequencing of the DNA fragments. Lastly, the obtained sequencing reads are used to decipher the genome sequence in a genome assembly process (Figure 3) (Loman et al. 2012).

The most typical approach in the determination of the bacterial genome sequence is the whole-genome shotgun (WGS) sequencing approach (Staden 1979). This approach involves the sequencing of randomly fragmented DNA molecules. Many of the DNA reads obtained in WGS sequencing overlap, forming the basis for the assembly of the reads into longer contiguous sequences. Traditionally, fragments were inserted into an appropriate vector and replicated in a bacterial culture to gain sufficient amplification of the fragment for sequencing (Staden 1979, Shendure & Ji 2008). Subsequently, PCR and massively parallel modern sequencing methods, some of which can even directly use the insert as a template (Table 1), have made the *in vivo* amplification step redundant (Loman et al. 2012, Shendure & Ji 2008). The templates are then sequenced either from one or both ends to generate single-end reads or paired reads, respectively (Loman et al. 2012, Pop 2009).

Before they can be used for biological analysis, the signals generated by the incorporation of bases must be translated into a human-readable format. This process is termed base calling and includes the transformation of the intensity signals into base calls, the assignment of a quality score to each base, and adjustment of the data for platform-specific anomalies (Sheikh & Erlich 2012). The read data are usually also quality assessed and pre-processed, which includes the trimming of adapter sequences and poor quality regions as well as the filtering of the unqualified reads (Sheikh & Erlich 2012).

After the pre-processing steps, the overlapping sequence reads can be identified and joined into contiguous consensus sequences termed contigs based on the assumption that overlapping reads originate from the same location (Nagarajan & Pop 2013). The contigs can be further organized into genomic scaffolds containing both contigs and gaps with known contig orientation, order, and distance (Nagarajan & Pop 2013, Pop 2009). The scaffolding process is often aided by the generation of paired reads or mate-pair reads, both referring to reads obtained by sequencing both ends of a DNA fragment of known size (Figure 3) (Pop 2009, Loman et al. 2012). In addition, other strategies exist for improving the contiguity, such as the use of optical maps, *i.e.*, ordered restriction maps created directly from genomic DNA molecules, comparative genomics, and directed sequencing of the

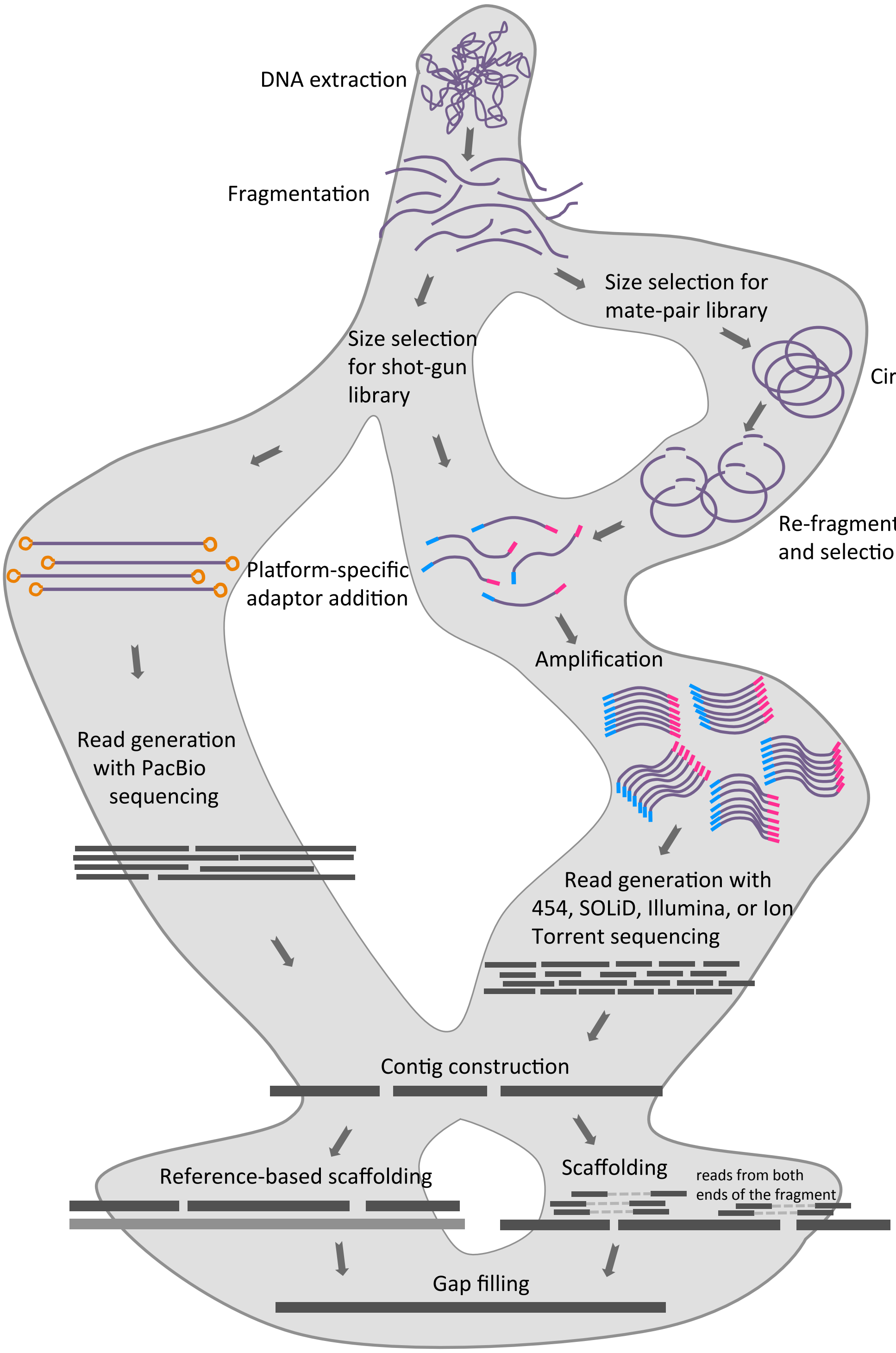

DNA extraction
Fragmentation
Size selection for mate-pair library
Size selection for shot-gun library
Circularization
Re-fragmentation and selection
Platform-specific adaptor addition
Amplification
Read generation with PacBio sequencing
Read generation with 454, SOLiD, Illumina, or Ion Torrent sequencing
Contig construction
Reference-based scaffolding
Scaffolding
reads from both ends of the fragment
Gap filling

Figure 3 *Simplified representation of the workflow used in a bacterial genome-sequencing project. The genomic DNA is first extracted and fragmented, after which sequencing libraries are generated from fragments of suitable lengths. The DNA fragments can be used for whole-genome shotgun and/or mate-pair sequencing. Larger DNA fragments are often used for the PacBio shotgun libraries and mate-pair libraries. In the latter, the ends of a fragment are joined, and the joined segment with its surrounding DNA is used for sequencing. Fragments appropriate for the sequencing library are ligated with adaptors. The fragments can then be sequenced directly (PacBio) or after amplification (the other indicated NGS platforms). Overlapping*

shotgun reads are joined into contigs, which can be organized into scaffolds by alignment to a reference genome or by deducing the order, orientation, and distance of the contigs from each other from mate-pair reads or paired reads that map to two different contigs. A combination or the different approaches and sequencing technologies are often used. Filling the gaps in the scaffolds results in complete whole genome sequence. The grey arrows indicate the direction of the workflows.

gap regions (Shendure & Lieberman Aiden 2012, França et al. 2002, Bartels et al. 2005, Samad et al. 1995). As the generation of a complete genome sequence without any gaps can be costly and labor intensive, genome assemblies not capturing the whole genome but consisting of scaffolds or independent contigs, referred to as draft genomes, have become increasingly popular in past years (Figure 2) (Chain et al. 2009). This practice, however, may fall out of favor given the ability of some NGS instruments (Table 1) to produce very long reads.

In addition to the above-described *de novo* bacterial genome sequencing and assembly, a comparative approach can be used to order reads into contigs (Bentley 2006, Pop 2009, Pop et al. 2004). Because a reference genome that is highly similar to the genome under interest is required in this approach, genomic discrepancies originating from, for example, HGT events and genomic re-arrangements, can result in erroneous assemblies and impede reference-based genome inference (Pop 2009, Pop et al. 2004).

2.1.3 Bacterial Genome Annotation

Once the bacterial genome sequence has been resolved, it is subjected to an annotation process, in which functional information is inferred from the genomic sequence. This genome annotation process includes both the identification of sequence features and the subsequent association of biological information with the identified features. Genome annotation is usually performed computationally given that the genome-wide identification and characterization of the features and their products would be too costly, time-consuming, and laborious to be conducted experimentally (Friedberg 2006, Lee et al. 2007).

2.1.3.1 Structural Annotation

A bacterial genome contains various sequence features that can be identified in the structural annotation phase. A key aspect of this annotation phase is gene prediction, or more specifically, the identification of genetic regions that code for proteins. In addition, genes transcribed into non-protein-coding or functional RNA-molecules, various regulatory features, and regions of foreign origin can be identified. Genome-wide structural annotation is usually performed with various computational approaches, which is often followed by manual curation (Stein 2001).

Bacterial genomes are full of protein-coding genes; the majority of bacteria have 85-90% of their genomes composed of coding sequences (CDSs) although this portion can vary a great deal (Kuo et al. 2009). The initial identification of CDSs typically relies on algorithms that distinguish CDS from non-CDS regions using statistical coding-region models (Delcher et al. 2007, Hyatt et al. 2010, Larsen & Krogh 2003). By applying complex statistical rules, the coding potential of a genomic region is quantified, and CDSs are thereby distinguished from other regions. The statistical models employed by these algorithms can be built based on the characteristics of known genes (Larsen & Krogh 2003) or genomic segments from the given genome that are assumed to represent genuine CDSs (Delcher et al. 2007, Hyatt et al. 2010). In some tools (Hyatt et al. 2010, Delcher et al. 2007), the presence of sequence motifs associated with the start of the CDSs, such as ribosomal binding sites (RBSs), is employed in the identification of the CDS and its correct start site. Different methods have been reported to detect over 98% of experimentally verified bacterial CDSs and to accurately locate the start site in over 90% of the cases (Hyatt et al. 2010). However, short CDSs and CDSs with an abnormal sequence composition are considered particularly difficult to predict, while gene overlaps and the scarcity of stop codons in genomes with a high GC% may also be problematic (Hyatt et al. 2010, Warren et al. 2010, Poptsova & Gogarten 2010).

Complementary to the use of statistical gene evaluation models, CDSs can be identified through sequence similarity searches against a database of known protein and gene sequences using the basic local alignment search tool (BLAST) (Altschul et al. 1997) or other sequence similarity search algorithms. This approach has been shown to identify many CDSs that are missed in the initial CDS-finding step (Warren et al. 2010). In addition, missed genes can be detected and CDS predictions validated using either RNA expression (Sorek & Cossart 2010) or proteomics data (Ansong et al. 2008), but this approach is more expensive and laborious than using purely computational strategies. Gene expression data can also be employed in the identification of operon structures, *i.e.*, genes that are co-transcribed in one polycistronic mRNA (Sorek & Cossart 2010). More conventionally, the operon organization is inferred computationally. Computational operon prediction can be based on the distance between the consecutive genes, the conservation of gene order, the functional relation of genes in the clusters, and/or the presence of sequence motifs such as promoters, transcription factor binding sites, and transcription termination sites (Brouwer et al. 2008).

In addition to CDSs, bacterial genomes contain other regions that code for functional elements, such as ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), clustered regularly interspaced short palindromic repeat (CRISPR) RNAs, and various small RNAs. Similarly as for CDSs, various computational approaches exist for the identification of these noncoding-RNA elements. For example, rRNA genes can be identified based on sequence similarity with known rRNAs (Lagesen et al. 2007), while the identification of tRNA genes is primarily based on the conserved structural profiles of tRNA molecules (Lowe & Eddy 1997, Laslett

& Canback 2004). CRISPR RNA, which cooperates with the associated proteins (CRISPR associated proteins; Cas) in the defense against invading DNA (Deveau et al. 2010, Barrangou & Horvath 2012, Makarova et al. 2011), is encoded by arrays that are composed of short (average ≈20-50 bp) direct repeats, interspaced by spacer sequences of defined length (Deveau et al. 2010). Because of their organization, the CRISPR arrays can be identified with repeat-finding algorithms or those built on these types of algorithms. For example, the highly popular PILER-CR CRISPR-finding algorithm builds upon repeat finding methods (Edgar 2007).

Bacterial genomes also often contain regions that have been obtained laterally from other organisms. These genomic islands (GIs) usually have a sequence composition different from that of the rest of the genome, and this composition difference is often used in their identification (Hsiao et al. 2005, Waack et al. 2006). Alternatively, GIs can be identified using comparative genomics (Langille et al. 2010). Moreover, search strategies similar to those used in the GI prediction can also be applied in the detection of plasmid-derived genomic regions (Zhou & Xu 2010). The detection of phages integrated in the bacterial genome (prophages) can rely on sequence comparisons to known phage and prophage genes, sequence composition analysis, and identification of potential attachment sites. For instance, the popular prophage finding programs Prophinder (Lima-Mendez et al. 2008) and PHAge Search Tool (PHAST) (Zhou et al. 2011) combine the afore-mentioned strategies to provide prophage predictions for bacterial genomes.

2.1.3.2 Functional Annotation

In addition to calling genetic features, realization of the full biological value of the sequenced genomes also requires the assignment of functions to the detected features. The cornerstone of this process can be considered to be the functional classification of CDSs (Lee et al. 2007). In addition, protein function can be studied at the system level and CDSs can be assembled into larger entities to examine the biological processes they conduct as an organizational unit.

The various functions attached to a sequence feature can be described in different ways. Usually, the functions of bacterial CDSs are given in description lines (DEs), which are informative descriptions of functional properties but suffer from the use of synonyms and ambiguous expressions (Friedberg 2006, Klimke et al. 2011). More uniform types of function descriptions also exist, such as the use of controlled vocabularies in various annotation systems, including TIGRFAM (Haft et al. 2003), SEED (Overbeek et al. 2005), clusters of orthologous groups of proteins (COGs) (Tatusov et al. 2003), and Enzyme Commission (EC) classification. The last of these is a hierarchical way of describing enzymatic activities of gene products. In addition, gene function can be described according to Gene Ontology (GO) terms (Ashburner et al. 2000), which consist of controlled vocabularies that describe the biological processes, cellular components, and molecular functions associated with the annotated protein. GO annotations are highly machine-readable and thus suitable for computational annotation; however, as they are focused on eukaryotic functions, the applicability of GOs in the annotation of bacterial genomes is limited.

The most typical approach in functional classification of the CDSs in a newly sequenced bacterial genome relies on the transfer of functional descriptions from a previously characterized protein with a similar sequence. This approach is premised on the correlation of functional similarity with sequence similarity (Lee et al. 2007, Friedberg 2006) and typically begins with searching the sequences of interest against a general protein sequence database with the help of pair-wise sequence similarity search algorithms (Altschul et al. 1997, Pearson & Lipman 1988, Koskinen & Holm 2012). Similarity searches are commonly performed at the level of protein instead of DNA sequences as this way the effects of the redundancy of the genetic code on the sensitivity of the search can be avoided. Moreover, if desired, the different substitution rates of amino acids can be taken into account (Koonin & Galperin 2003). The functional information associated with the most significant hit can then be transferred to the query sequence directly, as in the best-BLAST approach. However, this hit does not necessarily represent a genuine evolutionary counterpart nor provide the most informative functional description (Friedberg 2006). For this reason, more sophisticated methods do not transfer function based on a single hit but aggregate information from multiple hits (for example, see Martin et al. 2004, Hawkins et al. 2006). Basing the functional prediction on more than a single hit has been shown to both improve the accuracy of the functional description of the query sequence and to increase the number of functionally classified CDSs in a genome (Martin et al. 2004, Hawkins et al. 2006). Another approach to improve the accuracy is to transfer annotations only among orthologs, *i.e.*, proteins related by speciation, in contrast to paralogs that resulted from gene duplication and may have acquired new functions (Friedberg 2006, Eisen 1998, Sonnhammer & Koonin 2002).

In addition to the above-described forms of homology-based annotation transfer, sequences of interest can be searched for signature motifs and domains in protein signature databases, such as PFAM (Finn et al. 2014), Conserved Domain Database (CDD) (Marchler-Bauer et al. 2015), TigrFAM (Haft et al. 2003) or InterPro (Mitchell et al. 2015). Moreover, several specialized databases and annotation services are dedicated to the characterization of specific subsets of CDSs. Examples of such services include databases for bacterial virulence factors (Zhou et al. 2007, Chen et al. 2005), carbohydrate-active enzymes (Lombard et al. 2014), and proteolytic enzymes (Rawlings et al. 2014), as well as tools for the characterization of bacteriocins (van Heel et al. 2013) or prophage-like genes (Lima-Mendez et al. 2008). In addition to the widely popular homology-based annotation transfer, the genomic context of the CDS or the intrinsic properties of its product can be used in functional annotation (Friedberg 2006, Lee et al. 2007). For example, the sub-cellular localization of a CDS can be predicted based solely on its amino acid content information (Gardy & Brinkman 2006).

The integration of different annotation approaches can result in a higher number of functionally classified CDSs than what is obtainable with a single method. For example, the conventional homology-based annotation transfer has been reported to be able to assign functions to approximately 73% of CDSs in an average genome, whereas inclusion of methods based on gene context information and protein interaction data has been reported to increase this percentage to an average of 85% (Raes et al. 2007). Homology-based annotation transfer naturally fails to annotate gene products without adequate sequence similarity to annotated products. Moreover, database search tools can fail to detect the correct counterparts given that sequence similarity does not always imply functional similarity (Eisen 1998). It is also noteworthy that sequence databases have numerous erroneous annotations (Schnoes et al. 2009). The level of sequence identity between the query and annotation source may also affect the outcome. Although pairwise sequence identity as low as 30% has been deemed sufficient for function transfer (Devos & Valencia 2000), other studies have indicated that pairwise identities of 40-50% (Tian & Skolnick 2003, Rost 2002) or up to 75% (Rost et al., 2003) should be used. Thus, the quality of functional annotation is affected by various factors, including the choice of sequence database, the search method and parameters, as well as the functional information transferred (Rost et al. 2003, Clark & Radivojac 2011).

Functional annotation of bacterial genomes can also be extended to system wide analyses of cellular processes (Feist et al. 2009). The metabolic information of an organism, for example, can be represented as networks, comprising a holistic view of the chemical conversion reactions of an organism. Metabolic reconstructions usually rely on the annotation information attached to the CDSs (Feist et al. 2009, Francke et al. 2005). Typically, EC numbers describing the enzymatic reactions or transport commission (TC) identifiers (Saier 2000) are used. The identifiers are then assembled into a comprehensive collection by mapping them against information in metabolic databases or reference pathways, such as those available in the KEGG (KEGG: Kyoto Encyclopedia of Genes and Genomes) resource (Kanehisa et al. 2012). Similar to all of the computational annotation processes, metabolic reconstructions can be further refined using information derived from the literature or experimental characterizations (Francke et al. 2005, Feist et al. 2009).

2.1.4 Comparative Bacterial Genomics

While sequencing an individual genome has the potential to provide an immense knowledge on the particular isolate, the comparison of several genome sequences allows a more comprehensive view of the bacterial group of interest by enabling the examination of intra- and interspecies differences and similarities. Importantly, comparative genomics approaches are also intertwined in the annotation process of newly sequenced genomes, as illustrated in the previous sections.

The comparison of whole genome sequences provides a view of the similarities and differences between the organisms in question at the genome level. These whole genome comparisons can reveal genomic rearrangements, such as inversions or translocations; can be used to explore the conservation of synteny, *i.e.*, collinear localization of the shared genomic segments; and aid the elucidation of evolutionary relationships (Ventura et al. 2009, Binnewies et al. 2006, Land et al. 2015, Batzoglou 2005). The identification of regions of dissimilarity between genomes can also suggest the presence and location of GIs and other sequence features of external origin, such as prophage clusters and transposable elements (Batzoglou 2005, Land et al. 2015, Binnewies et al. 2006, Ventura et al. 2009). Similar genomic subregions and their locations and orders can be determined by aligning two or more genome sequences (Batzoglou 2005). Genome alignments can be performed pairwise but can also be extended to cover multiple genome sequences, and there are different approaches for the generation of the alignments (Batzoglou 2005, Darling et al. 2010). In addition to sequence alignments, whole genome comparisons can be generated using dot matrix analysis and visualized as dot plots (Krumsiek et al. 2007).

Comparative genomics analyses can also be performed at the protein-level. Orthologous grouping (Li et al. 2003, Kristensen et al. 2011) of CDSs is often performed to classify them according to their evolutionary origins (Sonnhammer & Koonin 2002). Orthologous grouping can rely on phylogeny-based strategies or those based on pairwise sequence comparison, such as BLAST (Kuzniar et al. 2008). As mentioned above (section 2.1.3.2), the identification of orthologs also has a major role in the functional classification of proteins as orthologous proteins can be assumed to share aspects of their functional characteristics or perform the same task (Friedberg 2006, Lee et al. 2007). Accordingly, orthologous grouping in conjunction with commensurable phenotype information provides a means to link a phenotype to a particular genotype. Such conclusions are possible because this type of analysis allows the identification of genes that are present only in isolates expressing the trait (Korbel et al. 2005).

The identification of orthologous groups shared by all of the isolates of a particular bacterial group, on the other hand, can be used to reveal the principle features of that group of organisms. The orthologous groups or genes possessed by all of the isolates form the core genome of these organisms and encode several of the properties and traits defining the group (Medini et al. 2005, Tettelin et al. 2005, Tettelin et al. 2008). In turn, the full complement of orthologous groups or genes in a group of bacteria forms the pan-genome of that group (Tettelin et al. 2008, Tettelin et al. 2005, Medini et al. 2005). The pan-genome encompasses the total genetic reservoir of the group and therefore also includes the core genome. The portion of the pan-genome excluding the core genome is termed the dispensable or accessory genome and comprises the orthologous groups or genes present in only some or one of the isolates (Medini et al. 2005, Tettelin et al. 2005, Tettelin et al. 2008). The accessory genome is responsible for the intraspecific diversity and codes for variable characteristics, which might be of use in adapting to new niches. Interestingly, accessory genes often reside in GIs, illustrating the role of

HGT in the adaptation of bacteria to new environments (Medini et al. 2005, Tettelin et al. 2008, Tettelin et al. 2005). As bacterial genomes often contain duplicated genes, the number of orthologous groups of an isolate is generally lower than the number of its CDSs.

The pan-genome data can also be used to assess how well the sequenced genomes capture the genetic repertoire of the particular species or bacterial group. The number of shared genes or novel genes found on sequential addition of each new sequenced genome can be extrapolated to estimate the size of the core or pan-genome of an infinite number of isolates, respectively (Tettelin et al. 2005, Medini et al. 2005, Tettelin et al. 2008). Moreover, if the size of the pan-genome grows as new genomes are added, the pan-genome is considered to be ‘open’. Highly diverse species occupying multiple environments, for example, would require a vast number of independent isolates to have their genomes defined before the complete genetic pool of the species would be characterized (Medini et al. 2005, Tettelin et al. 2008).

In addition, the identification of orthologous groups can help elucidate evolutionary events and relationships between organisms. A typical phylogenetic analysis of a set of sequences, for example, begins with the identification of homologous sequences. Following this step, a multiple sequence alignment is constructed and used in the inference of phylogeny via approaches based on parsimony, distance, or probabilistic models such as maximum likelihood (Anisimova et al. 2013). Omnipresent housekeeping genes or 16S rRNA genes are often used in the phylogenetic tree constructions, but trees based on genome-wide data are often considered to provide a more consistent and accurate picture of evolutionary events (Rokas et al. 2003).

2.1.5 Resolving Bacterial Transcriptomes via Sequencing

The manifestation of the bacterial phenotype is largely determined by the expression of the genes encoded in the bacterium’s genome. Thus, the examination of the RNA transcripts not only allows insights into the characteristics and functions of bacteria in a given condition, but also is useful in the detection of genes and verification of gene annotations (Sorek & Cossart 2010). The messenger RNA (mRNA) portion of the transcriptome is often of interest in gene expression studies as it reflects the phenotypic output of protein-coding genes. mRNA molecules, however, are the minority of all of the RNA molecules present in a given sample in given time. A typical bacterium, for example, has 0.05-0.1 pg of RNA in its cell, the majority of which consists of non-coding RNA molecules, particularly rRNA; the portion of mRNA rarely exceeds 4% (Skvortsov & Azhikina 2010).

The power of NGS technologies has been harnessed for the examination of transcriptomes, often termed RNA-seq (RNA-sequencing) or whole transcriptome shotgun sequencing. RNA-seq allows a more in-depth view of the transcriptome than is possible with hybridization-based techniques such as northern-blotting or microarrays (van Vliet 2010, Wang et al. 2009). A typical RNA-seq experiment (Figure 4) begins with the isolation of total RNA from the sample. The mRNAs can then be enriched prior to library preparation by depleting the rRNA and tRNA molecules (Skvortsov & Azhikina 2010, van Vliet 2010). After a complementary-DNA (cDNA) library is prepared, the library is sequenced in similar manner as the DNA libraries used in whole-genome sequencing (van Vliet 2010, Wang et al. 2009, section 2.1.2). Briefly, the library-inserted templates are sequenced from one or both ends to generate single-end or paired reads, respectively, and the generated reads are preprocessed before further analysis. The transcriptome can then be inferred from the reads either by performing a *de novo* assembly of reads or by using a reference genome or a set of reference genomes (van Vliet 2010).

In the case of a known reference genome, preprocessed reads are typically aligned to the reference sequence to infer their genomic origin. Different tools have been dedicated to this process, which is usually referred to as mapping (Trapnell & Salzberg 2009, Oshlack et al. 2010). The number of reads mapping to specific genomic regions, most often genes or CDSs, are then counted. The number of reads generated from a transcript is not directly proportional to the relative abundance of the transcript in the sample; many other factors affect the read counts, such as the sequencing depth, the choice of sequencing platform, the library preparation protocol, and the intrinsic properties of the transcripts (Mortazavi et al. 2008, Oshlack et al. 2010, Dillies et al. 2013, Oshlack & Wakefield 2009). To compare the expression levels within or between samples, the read counts must thus be normalized (Oshlack et al. 2010).

One of the earliest and most used normalization approaches is the reads per kilobase per million mapped reads (RPKM) (Mortazavi et al. 2008), which attempts to account for both the library size (the total number of mapped reads) and the gene length effects (random fragmentation generally results in more reads originating from bigger transcripts than smaller transcripts at the same expression level). The RPKM strategy can be considered to be useful for the normalization of read counts for the comparison of gene expression within the sample, with some limitations (Oshlack et al. 2010). However, the use of this normalization strategy in the comparison of gene expression levels between different samples can suffer from its poor suitability for weakly expressed genes and disregard of differences in the sampled RNA populations, such as differences in the GC-contents of different transcripts (Robinson & Oshlack 2010, Dillies et al. 2013). Compared samples can differ largely in their RNA population composition, and a large number of transcripts originating from just a few genes can account for the majority of the sequencing library (Oshlack et al. 2010). If not appropriately accounted for, these differences may skew the downstream differential expression analysis towards one condition. Thus, various, more sophisticated, normalization strategies have been developed for between-sample comparisons depending on the particular application (Robinson & Oshlack 2010, Oshlack et al. 2010, Dillies et al. 2013).

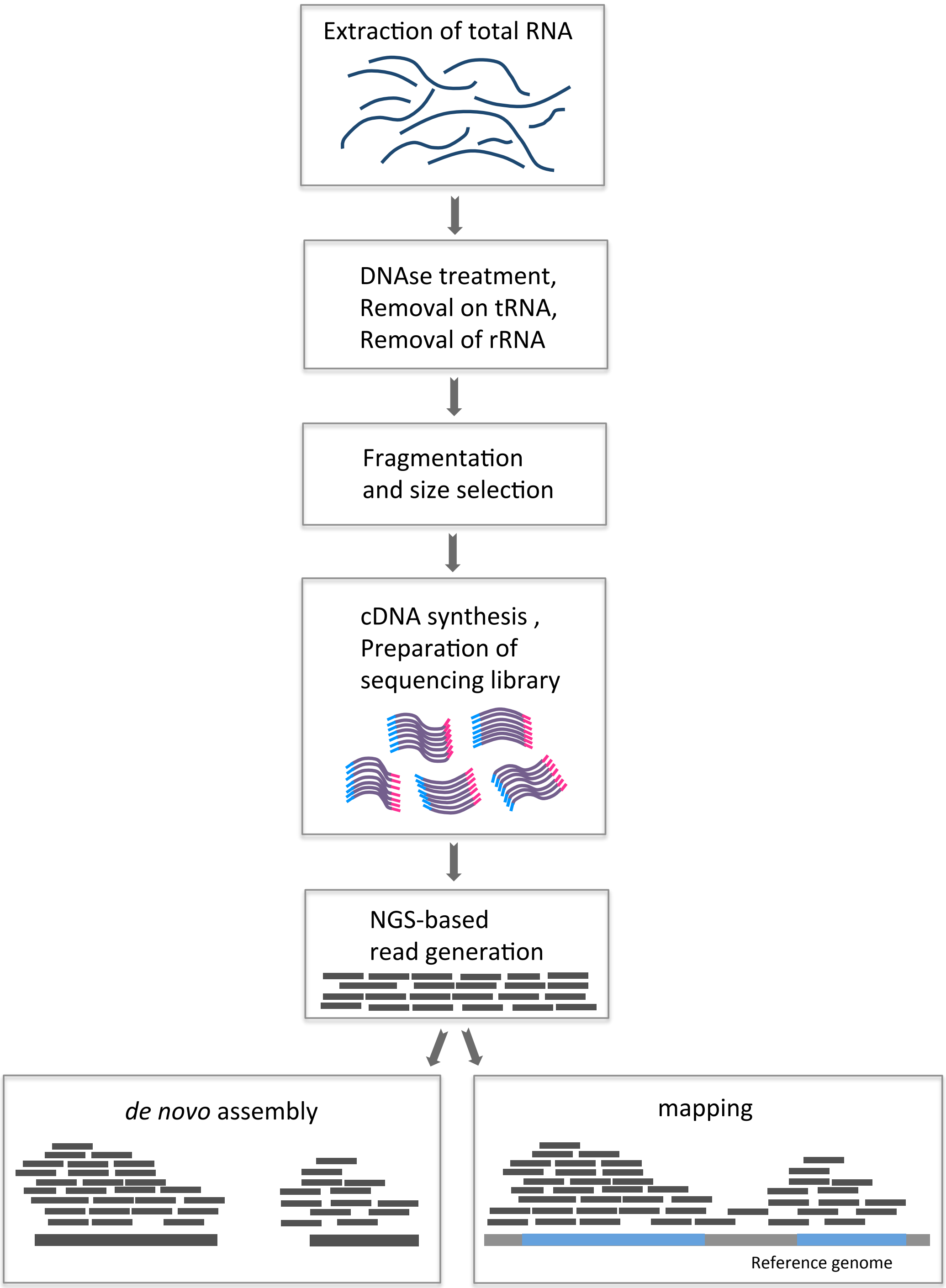


Figure 4 *Simplified representation of typical workflow of a RNA-seq project. The total RNA is isolated from the sample, the contaminating DNA is removed, and the mRNA is enriched, fragmented, and used in the cDNA sequencing library preparation. The cDNA libraries are then sequenced and the generated reads can be either be used in* de novo *assembly or mapped onto a reference genome. The light blue boxes in the reference genome represent the CDSs. The grey arrows indicate the direction of the workflow.*

After read count normalization, the genes that have significantly changed their expression between conditions can be identified in the differential expression analysis. Briefly, this analysis uses normalized read count tables in statistical testing between the samples under interest (Oshlack et al. 2010). Various strategies for this statistical testing exist, and, similar to the choice of the normalization method, the most suitable method should be chosen based on the experimental design (Soneson & Delorenzi 2013). The resulting list of genes with altered expression levels provides information on the transcriptional changes between samples, and deeper biological insights can be obtained by (i) looking at the transcriptional changes of sets of genes or (ii) performing enrichment or pathway analyses of the differentially expressed genes (Oshlack et al. 2010). Moreover, the expression patterns of genes can be used to assign genes to clusters (D'haeseleer 2005). This type of analysis can provide interesting clues about genes that are co-regulated, involved in the same cellular processes, or which underlie a bacterial trait of interest.

2.2 Lactobacilli

The genus *Lactobacillus* belongs to the family *Lactobacillaceae*, order *Lactobacillales*, class *Bacilli*, and phylum *Firmicutes* of the domain Bacteria (Garrity et al. 2007a, Hammes & Vogel 1995). Phylogenetically, it is intermixed with another genus in the same family, namely, *Pediococcus* (Salvetti et al. 2012). Lactobacilli – and pediococci – are also traditionally classified into a loosely defined functional group of lactic acid bacteria (LAB). This group represents a heterogeneous collection of bacteria united by an array of morphological, physiological, and metabolic properties, most notably the production of lactate as the main end product of carbohydrate fermentation (Hammes & Vogel 1995, Axelsson 2004). In addition to lactate production, the typifying characteristics of the genus *Lactobacillus* include Gram-positive staining, a lack of enzyme catalase, complex nutritional requirements, generally strictly fermentative metabolism, and low genomic GC content (generally below 50 mol%) (Axelsson 2004, Hammes & Vogel 1995). Lactobacilli are also anaerobic or aero-tolerant and grow well in acidic environments (Axelsson 2004, Hammes & Vogel 1995).

Lactobacilli are ubiquitous in a wide variety of environments that are rich in carbohydrates, with favorable niches ranging from plants and dairy to host-associated habitats (Axelsson 2004, Hammes & Vogel 1995). Notably, lactobacilli are important in the fermentation of a range of food and feed products (cheeses, yoghurts, pickled vegetables, silage, fermented meets, *etc.*), in which their ability to efficiently acidify the surrounding environment is regarded as particularly beneficial as a low pH can inhibit the growth of undesirable bacteria (Giraffa et al. 2010). In addition, lactobacilli can enhance the flavor, texture, and nutritional value of the fermented food product (Giraffa et al.

2010, Leroy & De Vuyst 2004). The ability of some lactobacilli to produce health-enhancing ingredients, such as vitamins and bioactive peptides, has also gained interest, as has the potential of some strains to confer a health benefit to the consumer (Giraffa et al. 2010, Leroy & De Vuyst 2004). Thus, it is unsurprising that lactobacilli are added to functional food products and are claimed to improve the health and wellbeing of the consumer (Giraffa et al. 2010). Several *Lactobacillus* species also possess the generally regarded as safe (GRAS) status (Salvetti et al. 2012), *i.e.*, they have been recognized as safe by expert evaluations based on scientific procedures and/or history of safe use in foods (Burdock & Carabin 2004).

Lactobacilli are significant members of the normal host flora, being found in the mouth, gastrointestinal tract (GIT), and genitourinary tract (GUT) of humans and animals (Hammes & Vogel 1995, Walter 2008). A key determiner of the sites suitable for *Lactobacillus* colonization seems to be the type of epithelium that lines the alimentary canal and the GUT. High numbers of lactobacilli are generally associated with the presence of stratified squamous epithelium (Walter 2008). For instance, lactobacilli are abundant in the stratified squamous epithelium-lined human vagina (Gustafsson et al. 2011, Vasquez et al. 2002, El Aila et al. 2009, Ravel et al. 2011, Chaban et al. 2014, Gajer et al. 2012, Romero et al. 2014) and in the proximal parts of the GIT of various animals containing such epithelium, including rodents, pigs, horses, and birds (Walter 2008). Lactobacilli are consistently numerically insignificant in the human GIT that lacks a stratified squamous epithelium, and the lactobacilli that are encountered in the human GIT most likely originate from food or the more proximal parts of the alimentary canal (Walter 2008).

Traditionally, lactobacilli are divided into three subgroups based on their carbohydrate fermentation: homofermentative, facultatively heterofermentative, and obligately heterofermentative (Hammes & Vogel 1995, Axelsson 2004, Felis & Dellaglio 2007, Salvetti et al. 2012). First among these, the homofermentative lactobacilli ferment hexoses almost exclusively via the glycolytic Embden-Meyerhof-Parnas pathway (EMPP). Second, facultatively heterofermentative lactobacilli ferment hexoses via the EMPP but can also ferment pentoses via a pentose phosphate pathway (PPP). Third, obligately heterofermentative lactobacilli lack a functional EMPP and ferment both hexoses and pentoses via PPP. These different modes of fermentation can also be explained by the presence or absence of genes encoding the key enzymes of the EMPP and PPP, *i.e.*, fructose-1,6-bisphosphate aldolase and phosphoketolase, respectively (Axelsson 2004). Differences in these pathways are also manifested in the differences in the resulting fermentative products. When glucose is metabolized, the homofermentative lactobacilli produce more than 85% of their fermentative products as lactate, whereas both types of heterofermenters produce lactate, carbon dioxide, ethanol, and/or acetate in equimolar amounts (Axelsson 2004, Hammes & Vogel 1995).

Species-wise, the genus *Lactobacillus* is the largest genus in its order (Felis & Dellaglio 2007), and the most recent survey revealed the genus to comprise 152 species (Salvetti et al. 2012). A phylogenetic analysis of the 16S rRNA genes (Salvetti et al. 2012) identified 29 distinct phylogenetic clades within the genus. Notably, the phylogenetic placement of a species did not necessarily correlate with the distribution of metabolic or other phenotypic properties. For example, the largest phylogenetic subgroup, termed the *Lactobacillus delbrueckii* group after its type species, comprises species of varying genomic GC content (ranging from 33 to an atypically high 51 mol%) and producing either D- and L-isomers of lactate or both. Species in this group demonstrate also strictly homo-fermentative or heterofermentative (facultatively or obligately) modes of carbohydrate fermentation (Salvetti et al. 2012). In addition, the isolation sources and habitats of species in the *L. delbrueckii* group vary vastly and include the feces, GIT, and GUT of humans and various animals; different dairy products, such as cheeses, yoghurts, and fermented milks; and vegetable and plant-associated fermentations, such as waste-corn fermentations (Salvetti et al. 2012).

An increasing number of lactobacilli have been subjected to whole-genome sequencing (Figure 2B). The sequenced *Lactobacillus* genomes and the genomics-based insights into the characteristics of this diverse group of bacteria are discussed in the following subsections with a particular focus on vaginal lactobacilli.

2.2.1 *Lactobacillus* genomes

The first *Lactobacillus* genome sequence to be publically available was that of *Lactobacillus plantarum* WCFS1, published in 2003 (Kleerebezem et al. 2003). The analysis of the ~3.3 Mb circular chromosome of this strain revealed it to harbor over 3,000 CDSs coding for, *e.g.*, several carbohydrate metabolism-related transporters and enzymes, such as the components of both the EMPP and PPP, corroborating the classification of the species *L. plantarum* as a heterofermenter (Kleerebezem et al. 2003).

In the following few years, the genome sequences of several other *Lactobacillus* isolates, representing many biotechnologically and biomedically important species, were elucidated (Altermann et al. 2005, Pridmore et al. 2004, Chaillou et al. 2005, van de Guchte et al. 2006, Makarova et al. 2006). These sequences provided insights into the physiology of lactobacilli and revealed interesting aspects about the evolution and frequency of HGT in lactobacilli. For instance, genome sequencing and analysis of *Lactobacillus acidophilus* NCFM revealed several putative adhesion and host-interaction factors (Altermann et al. 2005), the functionality of some were validated in later studies (Buck et al. 2005, Konstantinov et al. 2008). The sequencing of the whole genome of *Lactobacillus johnsonii* NCC 533 shed light on the bacterium's adaptation to the intestinal environment (Pridmore et al. 2004), and determination of the whole-genome sequence of a sausage-isolated *Lactobacillus sakei* 23K revealed mechanisms important in the colonization of meat and the tolerance of cold (Chaillou et al. 2005). The genome of yoghurt-associated *L. delbrueckii* ssp.

bulgaricus ATCC 11842 suggested an on-going adaptation to a lactose and protein-rich milk environment through the loss of superfluous biosynthetic functions and possibly via protocooperation with another dairy LAB, *Streptococcus thermophilus* (van de Guchte et al. 2006). In addition, Makarova and co-workers (Makarova et al. 2006) produced genome sequences for four lactobacilli and five LAB isolates, highlighting the importance of genome reduction and HGT in the niche-adaptation of LAB.

After 2006, an increasing number of *Lactobacillus* genomes have been characterized using NGS technologies. By the end of 2009, for example, nearly 60 *Lactobacillus* genomes had been sequenced and published (Figure 2B), and as of April 2015, there are 384 genome entries in GenBank classified as *Lactobacillus*. In accordance with the general trend of bacterial genome sequencing (Figure 2A), the majority of the entries released since 2009 do not represent finished genome sequences (Figure 2B). The finished *Lactobacillus* genome sequences exhibit great variation in several characteristics, such as size (1.4-3.4 Mb) and GC-content (33–52 %), highlighting the very heterogeneous make-up of this genus.

2.2.2 Comparative Genomics of Lactobacilli

As the first *Lactobacillus* genome sequences were determined when DNA sequencing costs were over $0.01 per base pair (Collins et al. 2003), it is unsurprising that the first genome sequencing based studies of lactobacilli (*e.g.*, Kleerebezem et al. 2003, Pridmore et al. 2004, Altermann et al. 2005, Chaillou et al. 2005) were concentrated on characterizing a single representative of the particular species. Consequently, the first comparative genomic analyses of *Lactobacillus* were interspecies comparisons, revealing different species to differ vastly in both their gene content and genome structure (Chaillou et al. 2005, Pridmore et al. 2004, Boekhorst et al. 2004). Comparative genomics of the distantly related *L. plantarum* WCFS1 and *L. johnsonii* NCC, for example, exemplified limited genomic synteny (only 28 regions containing at least 7 genes demonstrated conserved gene order between the strains) and revealed that ~50% and ~30% of their predicted proteomes, respectively, have no sufficiently similar counterpart in the other strain's proteome (Boekhorst et al. 2004).

Genomic comparisons of well-characterized isolates of related or the same *Lactobacillus* species have also proved useful in exploring the genetic background of the observed phenotypic differences between the isolates. For instance, genome comparison of a bacteriocin-producing *Lactobacillus reuteri* strain with a non-producing strain of the species *Lactobacillus fermentum* allowed the identification of a GI encoding an antimicrobial compound (Morita et al. 2008) and comparative genomics of well- and poorly adherent *Lactobacillus rhamnosus* strains resulted in the discovery of a gene cluster encoding adhesion-mediating pili on the surface of the adherent strain (Kankainen et al. 2009).

A few *Lactobacillus* species have also been subjected to a more comprehensive characterization involving the elucidation of the genomic make-up of several independent isolates of the species (Table 2). The similarities and differences within the specific *Lactobacillus* species were determined by comparing a set of isolates against a selected reference or references (Cai et al. 2009, Siezen et al. 2010, Douillard et al. 2013) or by performing comparative genomics analyses on a set of whole-genome sequences (El Kafsi et al. 2014, Smokvina et al. 2013, Broadbent et al. 2012, Kant et al. 2014, Spinler et al. 2014). The reference-based comparisons of lactobacilli include analyses performed via comparative genomic hybridizations with the reference (Siezen et al. 2010, Cai et al. 2009) or by generating short sequence reads and mapping them onto the selected reference (Douillard et al. 2013). Although offering a very limited view of the scale of intraspecific genomic variation, these approaches permit the detection of the features conserved across the species and have, for example, revealed the ecologically versatile *Lactobacillus* species *L. casei, L. plantarum,* and *L. rhamnosus* to have core genomes that constitute a major part of the genetic complement of the selected reference strain (~70%, ~70%, and ~80%, respectively) (Cai et al. 2009, Siezen et al. 2010, Douillard et al. 2013)

Whole genome sequences of multiple isolates of the same species, on the other hand, have offered a more in-depth view of the intraspecific similarities and differences in lactobacilli than what is possible using reference-based comparisons that can naturally detect only the portion of the genome shared with the reference. Moreover, whole genome-based comparisons have also allowed the identification of conserved genes that were undetectable in comparative genomic hybridizations due to low sequence homology. For instance, sequencing-based analyses revealed *L. plantarum* strains to harbor a teichoic acid biosynthesis gene cluster originally missed in CGH analyses. This cluster was not detected in the first analysis because the gene cluster in the reference strain shared only 69-74% nucleotide identity with the clusters of the other strains (Siezen & van Hylckama Vlieg 2011). A few *Lactobacillus* species with multiple independent genome sequences have been subjected to whole-genome comparative analyses, including *L. casei* (Broadbent et al. 2012)*, L. paracasei* (Smokvina et al. 2013)*, L. delbrueckii* (El Kafsi et al. 2014)*, L. reuteri* (Nelson et al. 2010, Spinler et al. 2014)*,* and *L. rhamnosus* (Kant et al. 2014). When the sizes of the reported core genomes of these species are compared with the number of CDSs present in the first representative sequenced genome of each species (the CDS numbers retrieved from the GenBank database (Benson et al. 2013), May 2015), the species core genomes cover ~59% (*L. paracasei* and *L. delbrueckii*) to ~99% (*L. acidophilus*) of the CDSs of the selected reference. Interestingly, the *L. rhamnosus* core genome, deduced from the analysis of 13 whole-genome sequences (Kant et al. 2014), covered ~77% of the reference, whereas the core genome deduced in the reference-based comparative analysis of 101 isolates (Douillard et al. 2013) covered ~89%. Similarly, the core genome obtained by comparative analysis of 17 *L. casei* isolates (Broadbent et al. 2012) covered ~62% of

Table 2. *Selected comparative genomics studies of lactobacilli describing the pan- and/or core genomes of at least five individual isolates.*

Species	No. of genomes	Core genome	Pan-genome	Approach	Reference
L. acidophilus	34	1815	-	Comparative genomics	Bull et al., 2014
L. casei	22	1941	-	Reference based (CGH)	Cai et al., 2009
L. casei	17	1715	5935	Comparative genomics	Broadbent et al., 2012
L. delbrueckii	10	989	-	Comparative genomics	El Kafsi et al., 2014
L. paracasei	13	1800	4200	Comparative genomics	Smokvina et al., 2013
L. plantarum	42	2049	-	Reference based (CGH)	Siezen et al., 2010
L. reuteri	7	~1600	-	Comparative genomics	Nelson et al., 2010
L. reuteri	10	1230	3700	Comparative genomics	Spinler et al., 2014
L. rhamnosus	13	2095	4893	Comparative genomics	Kant et al., 2014
L. rhamnosus	101	2419	-	Reference based (mapping)	Douillard et al., 2013
L. casei group (3 species)	10	1682	-	Comparative genomics	Toh et al., 2013
Lactobacillus (5 species)	5	593	-	Comparative genomics	Canchaya et al., 2006
Lactobacillus (11 species)	12	141	-	Comparative genomics	Claesson et al., 2008
Lactobacillus (14 species)	20	383	~14000	Comparative genomics	Kant et al., 2011
Lactobacillus (14 species)	21	363	13069	Comparative genomics	Lukjancenko et al., 2012
Lactobacillus (27 species)	67	311	11047	Comparative genomics	Mendes-Soares et al., 2014

the reference, and the core genome of 22 *L. casei* isolates (obtained using reference-based approaches) (Cai et al. 2009) covered ~70% of the reference. These results indicate that reference-based approaches are less strict than whole-genome approaches in defining the genomic core of a species.

In addition to encoding basic cellular functions, the core genomes of *L. rhamnosus* (Kant et al. 2014), *L. casei* (Broadbent et al. 2012), and *L. paracasei* (Smokvina et al. 2013), for instance, encode several surface molecules that are possibly involved in the interactions with the surrounding environment and even have some core features in common (Kant et al. 2014). Moreover, multiple genome sequence based analyses have determined, that HGT and gene loss have played a major role in the evolution of *Lactobacillus* species (Broadbent et al. 2012, Smokvina et al. 2013, El Kafsi et al. 2014). An exception appears to be *L. acidophilus,* for which the core genome covers almost the entire gene repertoire of a single isolate (Bull et al. 2014).

In addition to describing the gene repertoires of particular *Lactobacillus* species, the pan- and core genomes of the entire genus *Lactobacillus* have been investigated several times (Table 2) (Claesson et al. 2008, Canchaya et al. 2006, Lukjancenko et al. 2012, Kant et al. 2011). These analyses have considered varying numbers of whole genome sequences, ranging from 5 (Canchaya et al. 2006) to 21 (Lukjancenko et al. 2012) representatives of different *Lactobacillus* species. These studies have resulted in the discovery of *Lactobacillus* core genomes ranging from 141 (Claesson et al. 2008) to 593 (Canchaya et al. 2006) genes, with the core genomes constituting only ~3% of the respective *Lactobacillus* pan-genomes (Table 2). Moreover, a recent study attempting to describe *Lactobacillus* genes specific for a vaginal niche (Mendes-Soares et al. 2014) reported the core and pan-genomes of altogether 67 isolates of various origins, with the core genome again forming only ~3% of the pan-genome (Table 2).

2.2.3 Sequencing-based Investigations of Vaginal Lactobacilli

The vaginas of healthy premenopausal women are nutrient-rich niches that harbor approximately an average of $10^{6–8}$ bacteria per gram of secretions (Danielsson et al. 2011). These vaginal commensal bacteria are considered to have a pivotal role in maintaining the urogenital health of the host (Boris & Barbés 2000, Witkin et al. 2007, Ma et al. 2012, Lamont et al. 2011, Petrova et al. 2015). Vaginal lactobacilli in particular are regarded to protect the GUT from aberrant conditions and to create an inhospitable environment for pathogenic bacteria via competitive exclusion and the production of lactate, biosurfactants, and hydrogen peroxide, among other factors (Ma et al. 2012, Boris & Barbés 2000, Witkin et al. 2007, Lamont et al. 2011, Petrova et al. 2015). Several insights into the biology of the vaginal lactobacilli have been derived from sequencing-based studies, as is discussed in more detail below.

In accordance with the culture-based identification of lactobacilli as the major components of normal vaginal flora (Gustafsson et al. 2011, Vasquez et al. 2002, El Aila et al. 2009), several culture-independent studies based on marker gene sequencing have detected lactobacilli in the vaginal flora of asymptomatic reproductive aged women (Figure 5) (Ravel et al. 2011, Chaban et al. 2014, Gajer et al. 2012, Romero et al. 2014). Studies based on 16S rRNA gene sequences (Romero et al. 2014, Gajer et al. 2012, Ravel et al. 2011) report that rather similar fractions of the analyzed vaginal floras are *Lactobacillus* dominated; the most abundant *Lactobacillus* species is *L. iners*, followed closely by *L. crispatus* (Figure 5). Analysis of the vaginal flora composition using the universal target region of the *cpn60*-gene (Chaban et al. 2014), on the other hand, has revealed *L. crispatus* to be the most abundant species in more vaginal samples than *L. iners* (Figure 5), a result that might reflect differences in detection capacities of the different methods and/or in the study populations.

The vaginal bacterial communities of asymptomatic reproductive aged women have further been classified into community types based on the composition and relative abundances of the detected species (Ravel et al. 2011). In their large-scale 16S rRNA gene sequence analysis-based study comprising nearly 400 women, Ravel and colleagues (Ravel et al. 2011) identified five main community types. Four of these community types were dominated by different *Lactobacillus* species (*L. crispatus, L. iners, L. jensenii*, and *L. gasseri*) and the fifth comprised a diverse array of facultatively or strictly anaerobic bacteria. This finding suggests that a healthy vaginal state can also exist without a prevalence of lactobacilli. The proportions of each community type were also reported to vary among women from different ethnic backgrounds; the *L. crispatus*-dominated community type was the most common type in white women, whereas *L. iners*-dominated communities were most common in Asian and the variable community type in black and Hispanic women. Together, *Lactobacillus*-dominated communities were found in ~90% and ~80% of white and Asian women, respectively, but only in ~60% of black and Hispanic women (Ravel et al. 2011). Notably, vaginal community types, including *Lactobacillus*-dominated types, have been identified in other marker gene-based studies; however, less abundant communities, namely, *L. gasseri* or *L. jensenii* have not always been identified (Chaban et al. 2014, Romero et al. 2014, Gajer et al. 2012).

A few genome sequences available in GenBank (Benson et al. 2013) belong to vaginal *Lactobacillus* isolates and have helped to elucidate the biology of vaginal lactobacilli. The first *L. iners* genome sequence, for instance, revealed the genome of this vaginal isolate to be only ~1.3 Mb in size and to encode a proteome with very limited biosynthetic abilities (Macklaim et al. 2011). Interestingly, the genome was reported to lack genes for most of the adhesion factors commonly found in other lactobacilli, such as mucus-binding proteins (Macklaim et al. 2011). Potential fibronectin- and fibrinogen -binding adhesins were, however, among the predicted adhesins (Macklaim et al. 2011), and fibronectin-binding was speculated to play a role in the persistence of *L. iners* in the vagina (McMillan et al. 2013).

Genome sequences of other urogenital *Lactobacillus* isolates, including also representatives of *L. iners* and the other most commonly found vaginal species (*L. crispatus, L. gasseri*, and *L. jensenii*), have also been produced by the Human Microbiome Project (HMP) (Nelson et al. 2010). The aim of the HMP is to provide reference genomes for further studies on health- and disease-associated microbes and microbiomes. As of May 2015, the HMP reference genome catalogue lists nearly 30 urogenital tract isolates of lactobacilli with genome sequence data submitted to GenBank (Benson et al. 2013). In addition, genomes of vaginal lactobacilli representing the less common vaginal species have been defined, including *L. plantarum* CMPG5300 (Malik et al. 2014), *L. helveticus* MTCC 5463 (Prajapati et al. 2011, Senan et al. 2015), and *L. pentosus* KCA1 (Anukam et al. 2013).

The abundance of *Lactobacillus* genome sequences has also prompted interest in the niche-specific traits of vaginal lactobacilli. However, a recent comparative genomics analysis (Mendes-Soares et al. 2014) found that the genomes chosen to represent the four most common vaginal *Lactobacillus* species (*L. crispatus, L. iners, L. jensenii*, and *L. gasseri*) did not share any features that were absent from other lactobacilli, suggesting that adaptation to the vaginal environment has not resulted in any vagina-specific gene set in lactobacilli (Mendes-Soares et al. 2014).

The different *Lactobacillus* species also seem to modulate the host's urogenital health differently. Interestingly, a longitudinal bacterial community analysis (Gajer et al. 2012) noted that the *L. crispatus*-dominated vaginal communities protect the vaginal normal flora from transitioning to a BV-associated community type, while such protection was not observed with the *L. iners*-dominated community type. The vaginal communities dominated by *L. crispatus* have also been observed to generally result in lower vaginal pH than communities dominated by other lactobacilli or bacteria (Ravel et al. 2011). Moreover, a metatranscriptomics analysis of vaginal communities (Macklaim et al. 2013) identified *L. crispatus* to be the most active species in the normal floras of asymptomatic women, whereas *L. iners* and *L. jensenii* were more active in the vaginas of women diagnosed with BV. Thus, it is unsurprising that colonization of the vagina by *L. crispatus* has been suggested to serve as a biomarker of a healthy vaginal state, while *L. iners*-dominated vaginal flora could mark a transitional state between healthy and BV-associated microbiotas (Petrova et al. 2015).

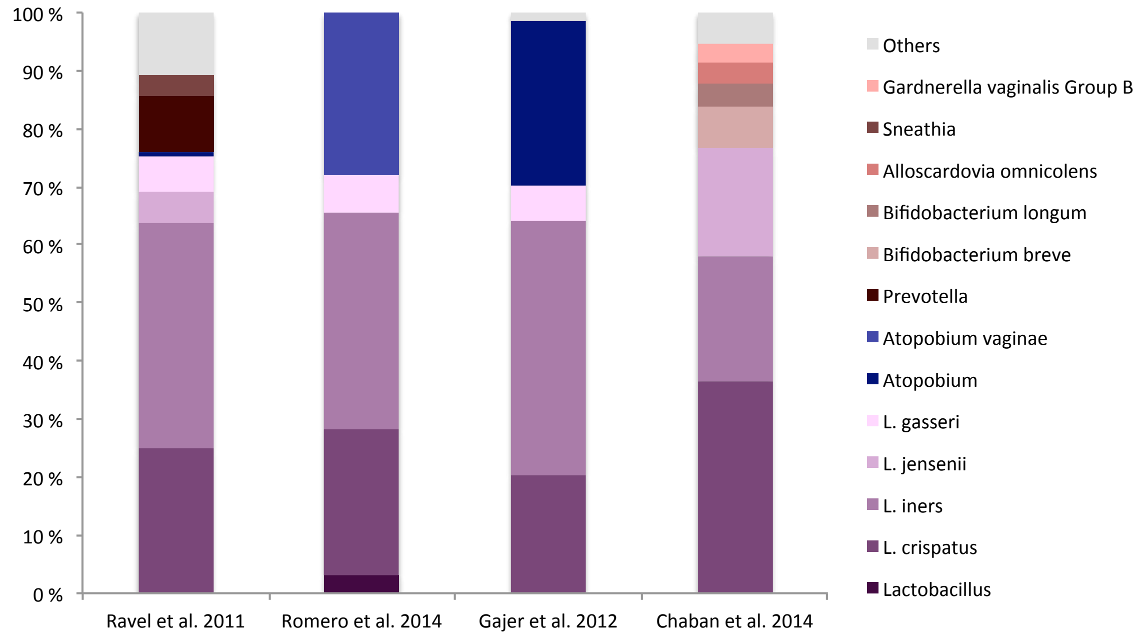


Figure 5 *Bacterial species and other taxa reported as the most abundant in the vaginas of healthy, non-pregnant, reproductive age women. Fractions for taxa reported as more abundant than any other single taxa in the normal floras of sampled women are given for four representative vaginal microbiome studies. The studies by Ravel et al., 2011, Gajer et al., 2012, and Romero et al., 2014 were based on 16S rRNA gene community profiling. The study by Chaban et al., 2014 was based on* cpn60 *community profiling. Pregnant subjects were excluded from the data reported by Romero et al., 2014. In the studies by Gajer et al., 2012, Romero et al., 2014, and Chaban et al., 2014, the test subjects were sampled multiple times, and the numbers given in the figure represent the taxa reported as the most abundant in a sample most often per each test subject. The studies were selected based on the availability of amplicon sequencing generated data of vaginal communities of non-pregnant, reproductive age and asymptomatic women.*

2.3 Propionibacteria

The genus *Propionibacterium* belongs to the family *Propionibacteriaceae*, order *Propionibacteriales* and class *Actinobacteria* of the high GC% phylum of Gram-positive bacteria (Actinobacteria) (Goodfellow 2012). The typical genomic GC content of propionibacteria ranges between 57 to 70 mol%, and the other typifying characteristics include generally complex nutritional requirements, anaerobic or aerotolerant lifestyle, and, usually, the possession of the enzyme catalase (Patrick & McDowell 2012). Metabolically, propionibacteria have a distinguishable ability to produce large quantities of propionate and acetate (Patrick & McDowell 2012).

The genus *Propionibacterium* has traditionally been divided into two groups according its main habitats: cutaneous and dairy (or classical) propionibacteria. Cutaneous propionibacteria inhabit the skin, GIT, and GUT of humans and animals, whereas dairy propionibacteria occupy silage and dairy products such as Swiss-type cheeses (Patrick & McDowell 2012, Poonam et al. 2012). Similar to lactobacilli, several members of the dairy group demonstrate a long history of safe application in food manufacturing and possess the GRAS status (Poonam et al. 2012); some are even considered to have health-promoting properties and are used in probiotics products (Cousin et al. 2011, Cousin et al. 2012, Poonam et al. 2012).

Following the traditional division, six of the twelve currently recognized *Propionibacterium* species fall into the cutaneous group, whereas four belong to the dairy group, and two do not fulfill the habitat-based criteria for either group (Patrick & McDowell 2012). The cutaneous species include, for example, (i) the skin colonizing *Propionibacterium acnes*, which has been associated with the formation of acne (Dessinioti & Katsambas 2010); and (ii) *Propionibacterium avidum*, implicated in abscess formation (Panagea et al. 2005, Janvier et al. 2013). The dairy group, on the other hand, includes several industrially important species, such as *Propionibacterium acidipropionici*, that has demonstrated a potential for the large-scale production of propionate (Poonam et al. 2012, Thierry et al. 2011, Wang et al. 2014). The dairy group also includes *P. freudenreichii*, which is important in the commercial production of B-vitamins and Swiss-type cheeses (Thierry et al. 2011, Poonam et al. 2012). Specifically, *P. freudenreichii* is the main *Propionibacterium* species used in the cheese industry. Its population density can reach over 10^9 cfu/g during ripening, and it has a main role in the formation of structure and flavor of Swiss-type cheeses (Thierry et al. 2011). The eyes of the Swiss-type cheese, for instance, are a result of extensive formation of CO_2 during the growth of *P. freudenreichii*, and several of the flavor compounds result from the fermentation of lactate produced by LAB, the catabolism of amino acids, and fat hydrolysis (Thierry et al. 2011).

The traditional division of propionibacteria into dairy and cutaneous groups, however, does not reflect the 16S rRNA gene sequence-based phylogenetic relationships within the genus (Patrick & McDowell, 2012). For example, the cutaneous species *Propionibacterium* *australiense* and *Propionibacterium acidifaciens* share a more recent common ancestor with dairy representative *P. freudenreichii* than with other cutaneous species. Moreover, *Propionibacterium microaerophilum*, not fitting either of the traditional group descriptions, has its closest relative in the dairy species *P. acidipropionici* (Patrick & McDowell 2012).

P. freudenreichii is also traditionally divided into two subspecies (ssp. *freudenreichii* and ssp. *shermanii*), distinguished on the basis of their ability or inability to reduce nitrate and ferment lactose. *P. freudenreichii* ssp. *freudenreichii* can reduce nitrate and is unable to ferment lactose, while the reverse is true for *P. freudenreichii* ssp. *shermanii* (Patrick & McDowell 2012). However, a recent study (de Freitas et al. 2015) argued against this subspecies division by demonstrating that some *P. freudenreichii* strains are able to both reduce nitrate and ferment lactose. In addition, the study reported that the nitrate reduction and lactose fermentation phenotypes of the strains did not correlate with the aroma compound production capabilities, suggesting the subspecies division to be of minor industrial relevance.

The research on propionibacteria has benefited greatly from advances in the sequencing technologies, enabling the characterization of increasing numbers of *Propionibacterium* genomes and transcriptomes. In the following subsections, the genomics and transcriptomics of propionibacteria are discussed in more detail.

2.3.1 *Propionibacterium* Genomes

The first *Propionibacterium* genome sequence to be publically available was that of *P. acnes* KPA171202, a member of the cutaneous group of propionibacteria (Bruggemann et al. 2004). The sequence revealed the genome to be ~2.6 Mb in size and to code for 2,333 CDSs. Among the predicted CDSs were several coding for putative virulence factors, such as sialidases, neuraminidases, and pore-forming factors (Bruggemann et al. 2004). Differing from the situation for lactobacilli, the genome of strain KPA171202 was the sole representative of propionibacteria for several years, with additional genome sequences beginning to accumulate only after 2009 (Figure 2C). The next six *Propionibacterium* genome entries following the release of the KPA171202 genome also described cutaneous propionibacteria, whereas the seventh described a dairy *Propionibacterium*. Five of the cutaneous *Propionibacterium* entries represented *P. acnes* and one represented *P. acidifaciens*, revealing the genome of this species to be an average of approximately 500 kb larger than the previously sequenced *P. acnes* genomes.

The first genome sequence of a dairy *Propionibacterium* was that of *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 (Falentin et al. 2010a). This ~2.7 Mb genome had a GC-content of 67% and was predicted to encode over 2,400 CDSs. This genome revealed important insights into the industrially significant characteristics of this cheese culture. The genome analysis of this strain, for instance, confirmed the bacterium to possess PPP and EMPP for the catabolism of a variety of carbon sources. Importantly, the CIRM-BIA1 genome also encoded a permease for L-lactate import, enzymes for the conversion of lactate into pyruvate, and Wood-Werkman cycle enzymes for the conversion of pyruvate into propionate (Falentin et al. 2010a). These results shed light on strain's ability to form the important cheese aroma compound propionate from the LAB-generated lactate during cheese manufacturing. The genome also revealed high biosynthetic abilities, such as pathways for the production of several amino acids and vitamin B12, which, together with a gene for a sodium/bile acid symporter and several stress response genes, suggested better adaptability to the intestinal rather than dairy environment (Falentin et al. 2010a).

As of May 2015, over 130 *Propionibacterium* genome entries are listed in GenBank, and this number can be expected to reach 144 by the end of the year (Figure 2C). The current *Propionibacterium* genome entries include both completed and fully annotated genome sequences as well as genomes of draft quality. The completed *Propionibacterium* genome sequences are all approximately ~2.5-3.7 Mb in size and have a GC content equal to or higher than 60%, which is in accordance with the genus characteristics (Patrick & McDowell 2012).

Despite the importance of dairy propionibacteria to the industry, the *Propionibacterium* genome sequences available in GenBank are dominated by representatives of the cutaneous propionibacteria; over 80% of the listed genome entries belong to the cutaneous group, whereas less than ten belong to the dairy group. Several of the genome entries representing cutaneous propionibacteria, particularly *P. acnes,* have been produced in the HMP project (Nelson et al. 2010). However, not all of the recognized cutaneous species are represented in the public genome entries, while each of the recognized dairy species (Patrick & McDowell 2012) is represented by at least one genome entry.

2.3.2 Comparative Genomics of Propionibacteria

The comparative genomics analyses of different propionibacteria have primarily been performed using complete *Propionibacterium* genomes sequences and/or predicted proteomes, *i.e.*, rather than the use of comparative genomic hybridizations. The first whole-genome sequence comparison of propionibacteria was performed in 2010 by Falentin and colleagues, who compared the sequenced genome of the *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 (Falentin et al. 2010a) to that of *P. acnes* KPA171202 (Bruggemann et al. 2004). Although this comparison revealed relatively high synteny between the genomes, the virulence-related genes of *P. acnes* were absent in the CIRM-BIA1 genome, supporting the GRAS status of *P. freudenreichii* (Falentin et al. 2010a).

A few other interspecies comparative genomics analyses have also been performed (Table 3). A comparative genomics analysis encompassing *P. acidipropionici* ATCC 4875 in addition to the aforementioned strains CIRM-

BIA1 and KPA171202 (Parizzi et al. 2012)revealed further similarities between the different species of propionibacteria. For instance, the representatives of *P. freudenreichii* and *P. acnes* were reported to have nearly 60% of their genes in common and to share ~65% and ~73% of their genes with *P. acidipropionici,* respectively. The greater similarity of *P. acidipropionici* with *P. acnes* than between *P. acidipropionici* and *P. freudenreichii* was also evident from genomic synteny analyses (Parizzi et al. 2012). This study also reported that the core and pan-genomes for the three species comprised 1,026 and 3,937 protein clusters, respectively (Table 3). Another interspecies comparative genomics analysis (Mak et al. 2013) focused on cutaneous propionibacteria. This study suggested distinct host-interaction strategies for each of the included three species. In particular, an array of unique surface proteins was present in *P. acnes*, while *P. avidum* and *P. granulosum* harbored exopolysaccharide (EPS) and pili-like structures, respectively. The selected strains were noted to possess a common core genome of 1,380 genes and a pan-genome of 3,025 genes (Table 3).

Possibly reflecting the abundance of publically available *P. acnes* genome sequences, the elucidation of intraspecific genomic differences and similarities of propionibacteria has primarily concerned *P. acnes* (Table 3). Comparison of five whole genome sequences representing different *P. acnes* phylotypes, for instance, has revealed several GIs harboring genes for a variety of putative virulence- and fitness-associated traits (Brzuszkiewicz et al. 2011). A comparison of 71 *P. acnes* genomes has further suggested the differences in virulence to stem partly from plasmids and GIs (Fitz-Gibbon et al. 2013). Moreover, comparative genomics of acne- and health-associated *P. acnes* strains has suggested that some previously identified virulence factors, such as hyaluronidase are not disease-associated (Lomholt & Kilian 2010). In addition, a comparative genomics analysis of 82 *P. acnes* genomes (Tomida et al. 2013) has revealed that the conserved sequence regions cover ~88% of an average *P. acnes* genome and that the common core genome of the strains includes 1,888 genes (Table 3). This

Table 3. *Selected comparative genomics studies of propionibacteria.*

Species	No. of genomes	Core genome	Pan-genome	Approach	Reference
P. freudenreichii ssp. *shermanii*; *P. acnes*	2	1108	-	Comparative genomics	Falentin et al., 2010a
P. freudenreichii ssp. *shermanii*; *P. acnes*; *P. acidipropionici*	3	1026	3937	Comparative genomics	Parizzi et al., 2012
P. acnes; *P. avidum*; *P. granulosum*	3	1380	3025	Comparative genomics	Mak et al.,2013
P. acnes	5	-	-	Comparative genomics	Brzuszkiewicz et al., 2011
P. acnes	71	-	-	Comparative genomics	Fitz-Gibbon et al., 2013
P. acnes	5	-	-	Comparative genomics	Lomholt & Kilian, 2010
P. acnes	82	1888	3136	Comparative genomics	Tomida et al., 2013
P. freudenreichii	21	1343	10962	Comparative genomics	Loux et al., 2015

study also reported the *P. acnes* pan-genome to be open and to expand slowly, with only a few new genes added per new genome, although the included genome sequences represent phylogenetically different linages. Each lineage was also observed to contain specific genomic elements and alterations in their noncore regions (Tomida et al. 2013).

A recent study (Loux et al. 2015) reported the core and pan-genomes for a dairy group member, *P. freudenreichii*. The clustering of the CDSs of 21 strains revealed the core genome to comprise 1,343 CDS-clusters and the pan-genome to include 10,962 CDS-clusters (Table 3). This study also succeeded in reconstructing complete pathways for the metabolism of five sugars (glucose, glycerol, mannose, galactose, and inositol) from the identified core genes. Interestingly, the abilities to degrade lactose and melibiose and to reduce nitrate were reported to be conferred by GIs harboring the necessary genes (Loux et al. 2015). Unfortunately, the newly sequenced genomes used in this study have not been submitted to GenBank (May 2015).

2.3.3 Transcriptomics of Propionibacteria

A few gene expression studies complement the genomic analyses of propionibacteria. The majority of these studies, particularly those concerning dairy propionibacteria, have been performed using approaches other than RNA-seq (Table 4). The expression levels of selected genes of *P. freudenreichii*, for example, have been quantified in microbiologically controlled model cheeses using real-time reverse transcription PCR (Falentin et al. 2010b). In addition, microarray-based analyses have been used to elucidate the global gene expression profiles of *P. freudenreichii* in conditions mimicking cheese ripening and cold storage (Dalmasso et al. 2012) and in the piglet GIT (Saraoui et al. 2013). Nevertheless, these studies have revealed *P. freudenreichii* to be able to adapt to various conditions encountered during cheese manufacturing as well as to the intestinal environment. The study by Falentin and colleagues (Falentin et al. 2010b) showed the activity of *P. freudenreichii* to increase during the cheese ripening process, and a more comprehensive gene expression analysis by Dalmasso and colleagues (Dalmasso et al. 2012) reported that nearly one quarter of the genes in the analyzed strain are differentially expressed between the warm and cold ripening conditions. Moreover, this study showed that several genes involved in the formation of cheese flavor compounds remained active even in the cold. The transcriptomic profile of *P. freudenreichii* in the colon (Saraoui et al. 2013) has suggested that this species can adapt its metabolism to match the carbohydrate availability in the intestine and indicates that the bacterium can grow in the piglet colon.

The transcriptomes of *P. acnes*, on the other hand, have been examined with both microarray and RNA-seq-based approaches, although not many studies concentrating on the activity and function of this bacterium exist (Table 4). Notably, comparative transcriptome analysis of two *P. acnes* strains grown under standard conditions (Brzuszkiewicz et al. 2011) suggested the differences in the virulence potentials of the strains to arise from differences not only in their genomes but also in the expression of various genes related to transport, metabolism, and virulence. Interestingly, the growth phase of the bacterium altered the expression of the genes coding for the virulence factors and indicated that virulence-associated genes were more active during exponential growth, whereas stress response genes were up-regulated during the stationary phase (Brzuszkiewicz et al. 2011). Another study (Lin et al. 2013) used RNA-seq-based approaches to characterize aspects of gene regulation in *P. acnes*. This study revealed individual transcription of tRNA-genes and the prevalence of translation initiation in the absence of a Shine-Dalgarno interaction in *P. acnes* (Lin et al. 2013)

Table 4. *Genome-wide gene expression studies of propionibacteria.*

Species	Strains	Sampling material	Sampling points	Approach	Reference
P. acnes	KPA171202, 266	Laboratory culture (37°C)	Exponential growth phase, stationary growth phase (KPA171202); mid and late exponential growth phase (KPA171202 and 266)	Microarray	Brzuszkiewicz et al., 2011
P. acnes	KPA171202	Laboratory culture (34°C)	1h after subculturing with or without potassium downshift	RNA-seq	Lin et al., 2013
P. freudenreichii ssp. *shermanii*	CIRM-BIA1	Laboratory culture mimicking cheese ripening	Exponential growth phase at 30°C (20h and 40h); stationary growth phase at 4°C (3d, 6d, and 9d)	Microarray	Dalmasso et al., 2012
P. freudenreichii ssp. *shermanii*	CIRM-BIA1	Culture in piglet intra-colonic dialysis tubing ; control laboratory culture	24h (intra-colonic culture); stationary phase (control)	Microarray	Saraoui et al., 2013

3 Aims of the Study

This thesis aimed to extend our understanding of the physiology, ecology, and function of two bacterial species of human relevance using sequencing- and the subsequent sequence analysis-based approaches. The objectives were (i) to provide annotated genomes for experimentally studied representatives of *L. crispatus*, a component of normal flora, and of the dairy culture *P. freudenreichii* and (ii) to discover the genetic underpinnings of the characteristic traits of these bacteria via functional genomic investigations.

With regards to *L. crispatus,* the scale and scope of the intraspecific variation was investigated utilizing pan- and core genome analysis approaches. Given that *L. crispatus*-dominated vaginal microbiotas are generally associated with a healthy vaginal state regardless of strain composition, specific focus was placed on uncovering *L. crispatus* core-encoded mechanisms supporting vaginal health.

The genomic investigation of *P. freudenreichii* was in turn complemented with transcriptome profiling to understand how the annotated genes and pathways are expressed at the different stages of cheese ripening and how the bacterium contributes to cheese properties during the industrial manufacture process.

In addition, automated protein function prediction approaches and strategies were compared to determine a practical annotation strategy for the genome-wide functional annotation of bacterial proteins and to aid the functional annotation of *L. crispatus* and *P. freudenreichii* protein-coding genes.

4 Materials and Methods

The materials and key methods are summarized in Table 5 and described in detail, with appropriate references, in the original publications. More information on the generation of the annotated whole-genome sequences of *L. crispatus* ST1 and *P. freudenreichii* ssp. *shermanii* JS can be found in Studies I and IV, respectively. The comparative genomics analyses of multiple *L. crispatus* isolates, including the pan- and core genome analyses, are described in Study III, and detailed description of the transcriptome profiling of strain JS during cheese ripening is presented in Study IV. The approaches used in the evaluation of the different annotation transfer methods are described in Study II.

Table 5. *Materials and key methods used in this thesis study.*

	Study I	Study II	Study III	Study IV
Experimental procedures				
Bacterial strains	*L. crispatus* ST1	-	*L. crispatus* EX533959VC06, *G. vaginalis* 101	*P. freudenreichii* ssp. *shermanii* JS
Growth conditions and biological samples	de Man, Rogosa, and Sharpe medium	-	-	Propionimedium, cheese (sampled at 12 d into warm ripening and at 7 d into cold ripening)
Adhesion assay	-	-	Bacterial adhesion to HeLa cells with/without pretreatment with Fab fragments	-
Sequencing	454 GS FLX, ABI 3730 Big Dye (Genome sequencing)	-	-	454 GS FLX and GS FLX Titanium, ABI 3130xl, SOLiD 4, PacBio RS II (Genome Sequencing); SOLiD 5500XL (Transcriptome sequencing); MiSeq (Amplicon sequencing)
Evaluation of annotation transfer methods and strategies				
Annotation transfer optimization	-	Sequence identity, Query ocverage, Annotation source (DE, GO, DE+GO)	-	-
Result comparison	-	Manual review, algorithmic	-	-
Genome data analysis				
Public genome entries	18 complete *Lactobacillus* genomes	-	10 *L. crispatus genomes*, 33 *G. vaginalis* genomes	29 *Propionibacterium* genomes
Genome assembly	Overlap layout consensus approach	-	-	Overlap layout consensus approach

Structural annotation	*ab initio* methods (CDSs); Evidence based methods (CDSs; rRNA); Covariance models based methods (tRNA); Specialized annotation systems (CRISPR, GIs, Intrisic terminators); manual curation (CDSs)	-	Specialized annotation systems (CRISPR, GIs, plasmids)	Evidence based methods (CDSs; rRNA); Covariance models based methods (tRNA); Specialized annotation systems (CRISPR, GIs)
Functional annotation	Pairwise sequence comparisons (General function); Domain annotation methods (Adhesins); Specialized databases and annotation systems (Enzymes, transporters, prophage, bacteriocins, cellular location), manual curation (General function)	-	Pairwise sequence comparisons (General function); Domain annotation methods (Adhesins, Cas); Specialized databases and annotation systems (Enzymes, prophage, bacteriocins); Manual curation (Prophage)	Pairwise sequence comparisons (General function); Domain annotation methods (protein domains); Specialized databases and annotation systems (Enzymes, prophage, bacteriocins)
Comparative genomics	Whole genome comparisons, Orthologous grouping, Phylogentic tree reconstructions	-	Whole genome comparisons, Orthologous grouping, Phylogentic tree reconstructions	Whole genome comparisons, Orthologous grouping, Phylogentic tree reconstructions
Metabolic reconstuctions	Reference-based metabolic reconstructions, literature search	-	Reference-based metabolic reconstructions, literature search	Reference-based metabolic reconstructions, literature search
Pan- and core genome analysis	-	-	Exponential decaying function, Power law	-
Transcriptome data analysis				
RNA-seq data processing and alignment	-	-	-	Preprocessing, Burrowa-Wheeler transform based read mapping, Feature summation, Trimmed Mean of M values normalization
Statistical analysis of RNA-seq data	-	-	-	Emperical Bayes, Ordinary T-test, Muliple hypthesis correction, gene set enrcihment analysis

5 Results and Discussion

This thesis aimed to advance our understanding of the physiology and biology of two bacterial species using sequencing- and sequence analysis-based approaches. The first species investigated, *L. crispatus*, belongs to the *L. delbrueckii* group of lactobacilli (Felis & Dellaglio 2007, Salvetti et al. 2012). It is a prominent member of the healthy human vaginal flora (Ravel et al. 2011, Chaban et al. 2014, Gajer et al. 2012, Romero et al. 2014, El Aila et al. 2009, Gustafsson et al. 2011, Vasquez et al. 2002) and encountered in the GIT of many animals (Abbas Hilmi et al. 2007, De Angelis et al. 2006, Yuki et al. 2000). To advance our understanding of the physiology of this species and shed light on its adhesion and host-microbe interaction mechanisms, the genome sequence of chicken-isolated *L. crispatus* ST1, previously shown to adhere strongly to chicken epithelial tissues as well as human vaginal and buccal cells (Edelman et al. 2002, Edelman et al. 2012), was characterized and annotated in Study I. The genome of strain ST1 was then compared to and analyzed together with the publically available genome sequences of vaginal *L. crispatus* isolates to provide a global view of the gene content and mutual characteristics of this species (Study III). Notably, this comparative genomics study proved effective in exposing adhesion-associated and other factors that were conserved at the species level, providing an intriguing explanation for the beneficial effects the species is considered to exert in the vagina.

The second species investigated in this thesis was *P. freudenreichii,* an important cheese culture also under study for its immunomodulatory properties (Thierry et al. 2011, Poonam et al. 2012). *P. freudenreichii,* ssp. *shermanii* JS, used in the manufacture of commercial cheeses, was subjected to whole-genome sequencing in Study IV, complementing the existing genome-derived information of this important dairy species. Study IV also extended the characterization of this bacterium to cover its function during cheese ripening. This aim was achieved by describing the global gene-expression of strain JS at different ripening stages. Importantly, this study provided the first genome-wide insights into the activity of *P. freudenreichii* during industrial cheese ripening. As strain JS was noted to be highly similar to previously genomically characterized *P. freudenreichii* isolates, the transcriptomics data derived in Study IV are also relevant for the functional characterization of the cheese ripening mechanisms of other *P. freudenreichii* isolates.

As in all genome projects, accurate protein function prediction formed the foundation of most of the functional genomic analyses performed in this thesis. Thus, in Study IV, different strategies for annotation transfer were tested, and their suitability for assigning DEs for bacterial protein sequences was evaluated. In the following subsections, the main findings of the annotation transfer evaluation as well as the results of all of the other studies included in this thesis are summarized and discussed in more detail.

5.1 Genomics of *L. crispatus* ST1

Strain ST1 exhibits several intriguing properties, including (i) strong adherence to a chicken cell line (Spivey et al. 2014), various areas of the chicken alimentary canal (Edelman et al. 2002), and human epithelial cells (Edelman et al. 2012); (ii) an ability to inhibit the adhesion of avian pathogenic *Escherichia coli* (Edelman et al. 2003); and (iii) a potential to interact with the human proteolytic plasminogen/plasmin system via secreted proteins (Hurmalainen et al. 2007). To better understand the molecular mechanisms underlying these traits and to gain insights into the physiology of strain ST1, it was subjected to whole genome sequencing (Study I). The ultimate goal was to obtain a finished whole-genome assembly; however, due to a particularly challenging repeat region within a gene encoding a *Lactobacillus* epithelium adhesin (LEA) (Edelman et al. 2012), one gap of approximately 590 bp in length remained in the final assembly of the ST1 genome. Nevertheless, the final genome assembly of strain ST1 was the first single scaffold assembly for *L. crispatus* to be publically available, providing a full list of genomic features of this bacterium and an almost complete reference assembly point for subsequent WGS and comparative genomics studies of this species. Particularly, the assembly has aided the production of other *L. crispatus* genome assemblies (Power et al. 2013) and served as a point of reference against which the scaffolds in *L. crispatus* draft genomes could be organized in Study III. Moreover, the genome assembly of strain ST1 has facilitated the elucidation of *Lactobacillus* biology in other comparative genomics analyses (Kant et al. 2011, Lukjancenko et al. 2012, Mendes-Soares et al. 2014).

The final genome assembly of strain ST1 obtained in Study I comprised a single circular chromosome of 2.04 Mb in size, having an overall GC content of ~37% (Table 6), which falls within the typical GC range (35-38%) reported for *L. crispatus* (Salvetti et al. 2012). Similar genomic GC contents have been reported for the closely related species *L. helveticus* (~38%) (Callanan et al. 2008) and *L. acidophilus* (~35%) (Altermann et al. 2005). The genome of strain ST1 was highly collinear with the genomes of these close relatives from the *L. delbrueckii* group (Figure 6) as well as with those of other *L. crispatus* isolates, as shown in Study III.

Table 6. *General characteristics of the sequenced genomes.*

	L. crispatus ST1	*P. freudenerichii* ssp. *shermanii* JS
Genome size (Mb)	2.04	2.68
GC%	36.90	67.23
CDSs	2024	2377
tRNA	64	45
rRNA clusters	4	2
Bacteriocin-like molecules	2	0
CRISPR/Cas system	Type I	Type I
Prophage like clusters	0	1
GIs	10	11

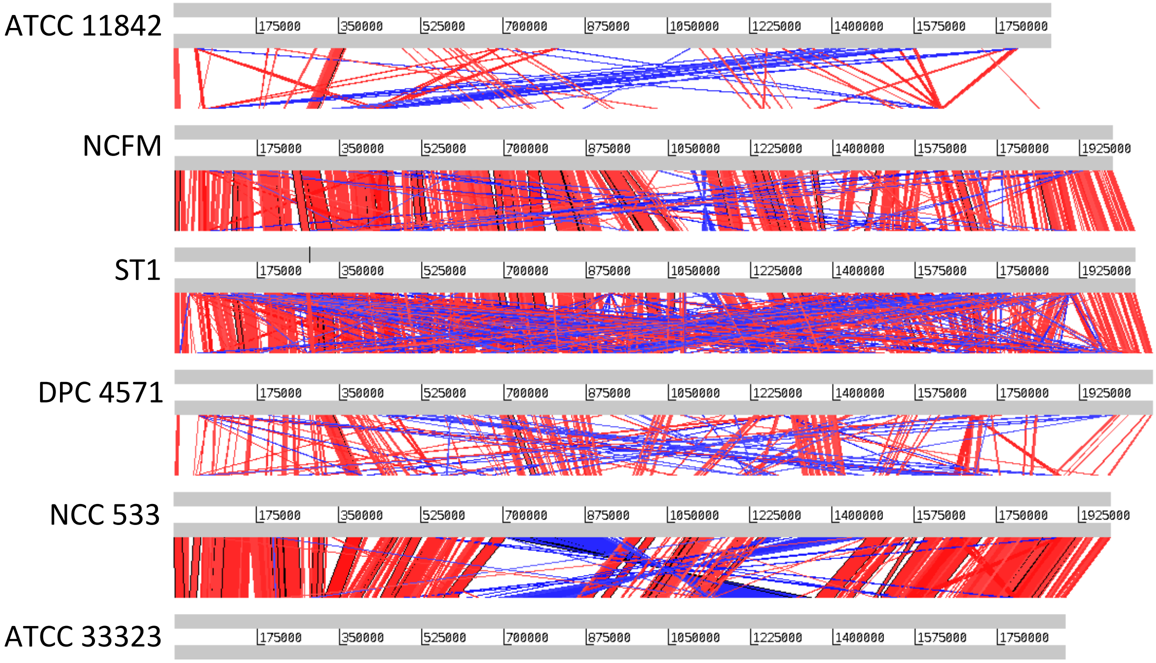


Figure 6 *Whole-genome alignments of selected lactobacilli of the* L. delbrueckii *subgroup. Genomes of* L. delbrueckii *subsp.* bulgaricus *ATCC 11842 (van de Guchte et al. 2006),* L. acidophilus *NCFM (Altermann et al. 2005),* L. crispatus *ST1 (Study I),* L. helveticus *DPC 457 (Callanan et al. 2008),* L. johnsonii *NCC 533 (Pridmore et al. 2004), and* L. gasseri *ATCC 33323 (Azcarate-Peril et al. 2008) were compared with BLASTN and visualized with the Artemis Comparison Tool (ACT). Red and blue vertical bands represent the forward and reverse matches (bit score ≥ 500), respectively.*

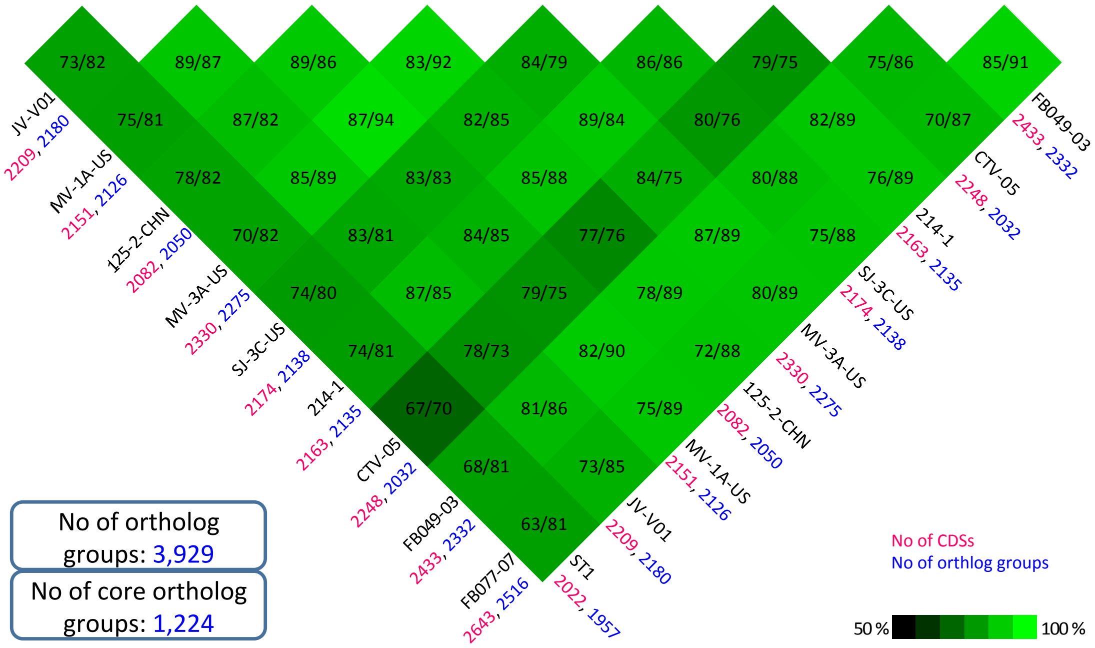
Annotation of the genome assembly of strain ST1 in Study I yielded various new insights into the characteristics of this strain, several of which extended to the entire species *L. crispatus* in Study III. Both EMPP and PPP were annotated based on EC number assignments (Studies I, III). Notably, the presence of both EMPP and PPP pathways suggested a facultatively heterofermentative nature of carbohydrate metabolism and is contradictory to the current classification of *L. crispatus* as homofermentative (Salvetti et al. 2012, Felis & Dellaglio 2007). Enzymes for the *de novo* synthesis or inter-conversion of eight amino acids were also found, revealing biosynthetic capabilities similar those of other previously analyzed related lactobacilli, including *L. helveticus* DPC 457 (Callanan et al. 2008), *L johnsonii* NCC 533 (Pridmore et al. 2004), and *L. acidophilus* NCFM (Altermann et al. 2005). The genome was also predicted to encode a putative bile hydrolase, which could be beneficial in the colonization of the chicken GIT and underlie the bile tolerance observed for the Rif^R variant of strain ST1 (Spivey et al. 2014).

Among the predicted genome features of strain ST1 were also CRISPR arrays and the associated *cas* genes, a gene cluster for EPS biosynthesis, and genes implicated in the production of antimicrobial compounds and adhesion-associated molecules (Studies I, III). These results are discussed in more detail in the following section together with the results obtained from a comparative genomics analysis of ten *L. crispatus* isolates (Study III).

5.2 Comparative Genomics of *L. crispatus*

The accumulation of *L. crispatus* genome data in public databases allowed for an investigation of intraspecific similarities and differences as well as a determination of the scale and scope of the pan- and core genomic potential of *L. crispatus* in Study III. The genome sequences of a total of ten *L. crispatus* isolates were investigated, including the genome of strain ST1 (Study I) and the draft genomes of nine vaginal isolates produced by the HMP (Nelson et al. 2010) and submitted to GenBank before January 2013. All of the analyzed strains exhibited extensive genomic similarity and synteny (Figure 1 in Study III), with the strains sharing on average ~90% of their genome sequence with at least one other included *L. crispatus* genome. This conservation within *L. crispatus* is in line with the previously observed genome synteny of lactobacilli closely related to *L. crispatus*. Specifically, the closely related *L. acidophilus* is notable for its intraspecific uniformity at the genome sequence level (Bull et al. 2014), and general conservation of overall gene order has been observed between representatives of *L. helveticus*, *L. acidophilus*, *L. johnsonii*, and *L. gasseri* (Callanan et al. 2008); Study I suggested this similarity extend to the species *L. crispatus* (Figure 6).

In accordance with the extensive intraspecific genome similarity, orthologous grouping of the predicted proteomes revealed the different *L. crispatus* isolates to share a majority of their proteinaceous features with at least one other isolate. On average, ~57% of the ortholog groups of a given strain were present in all of the strains (Figure 7, Figure 2D in Study III). Specifically, the pan- and core genomes of the ten analyzed *L. crispatus* isolates comprised 3,929 and 1,224 ortholog groups, respectively, with the core forming ~31% of the pan-genome of these ten isolates. This proportion of core groups is similar to those reported for *L. casei* (~29%) (Broadbent et al. 2012) and *L. reuteri* (~33%) (Spinler et al. 2014) and is slightly smaller than that of *L. paracasei* (~43%) (Smokvina et al. 2013), which, in turn, is similar to that of core CDSs relative to all reported CDSs for *L. rhamnosus* (~43%) (Kant et al. 2014). The comparison of the different pan- and core genome analyses, however, is complicated by the differences in the methods, parameters, and number of strains included (Table 2). These effects are exemplified, for example, in the different core genome sizes reported for *L. rhamnosus* and *L. casei*. Reference-based analyses (Douillard et al. 2013, Cai et al. 2009) have identified the core genomes of these species to be over ~300 and ~200 features larger, respectively, than the core genomes obtained via whole-genome sequence data and orthologous grouping analyses (Kant et al. 2014, Broadbent et al. 2012) (Table 2). These differences are particularly interesting given that (i) the aforementioned reference-based analyses comprised more isolates than the orthologous grouping-based studies (101 versus 13 *L. rhamnosus* isolates and 22 versus 17 *L. casei* isolates) and (ii) the size of the core genome should decrease rather than increase as more genomes are analyzed (Medini et al. 2005).



	FB049-03 2433, 2332	CTV-05 2248, 2032	214-1 2163, 2135	SJ-3C-US 2174, 2138	MV-3A-US 2330, 2275	125-2-CHN 2082, 2050	MV-1A-US 2151, 2126	JV-V01 2209, 2180	ST1 2022, 1957
JV-V01 2209, 2180									73/82
MV-1A-US 2151, 2126								89/87	75/81
125-2-CHN 2082, 2050							89/86	87/82	78/82
MV-3A-US 2330, 2275						83/92	87/94	85/89	70/82
SJ-3C-US 2174, 2138					84/79	82/85	83/83	83/81	74/80
214-1 2163, 2135				86/86	89/84	85/88	84/85	87/85	74/81
CTV-05 2248, 2032			79/75	80/76	84/75	77/76	79/75	78/73	67/70
FB049-03 2433, 2332		75/86	82/89	80/88	87/89	78/89	82/90	81/86	68/81
FB077-07 2643, 2516	85/91	70/87	76/89	75/88	80/89	72/88	75/89	73/85	63/81

No of ortholog groups: 3,929
No of core ortholog groups: 1,224
No of CDSs
No of orthlog groups
50 %
100 %

Figure 7 *Shared ortholog gene groups between different* L. crispatus *strains included in Study III. The number of predicted CDSs (red) and OrthoMCL defined ortholog groups (blue) is given for each strain below its name. In the matrix cells, the percentage of the shared groups between the strains is given with respect to all of the ortholog groups assigned to the strain; the first and second numbers are the percentages for the strains on the left and right sides of the matrix, respectively.*

Extrapolation of the *L. crispatus* pan-genome data further revealed that the size of the pan-genome was far from saturated and would continue to grow with the addition of new isolates (Figure 2A in Study III). Interestingly, comparative genomics analyses of *L. reuteri* have highlighted the effect of host origin on the genetic composition of different isolates, a feature that is also reflected in the number of new genes each isolate adds to the pan-genome (Frese et al. 2011, Nelson et al. 2010, Spinler et al. 2014). Thus, the overrepresentation of vaginal isolates and the presence of only one animal isolate in the *L. crispatus* data set used in Study III (which comprised all *L. crispatus* genome sequences that were publically available at the time of the study) might have a produced a much smaller pan-genome than would be obtained with more equal niche-distributions of the isolates. Although the ten isolates far from captured the entire genetic variation within the species, they led to a highly precise definition of the conserved features in *L. crispatus*. Regression analysis of the *L. crispatus* core genome data estimated the core genome of an unlimited number of *L. crispatus* isolates to level at approximately 1,116± 58 ortholog groups (Figure 2B in Study III). This estimate revealed the core genome of the ten isolates (1,224 groups) to be a rather close approximation of the final species core genome and provided a solid foundation for the investigation of the basic aspects of *L. crispatus* biology.

With regards to metabolism, the majority of the enzymes identified in the ten *L. crispatus* isolates were part of the core, and metabolic pathway analysis revealed the isolates to have highly similar fermentative and biosynthetic abilities (Additional files 8 and 9 in Study III). Notably, metabolic pathway analysis indicated that both the EMPP and PPP, already identified in the genome analysis of strain ST1 (Study I), were encoded in the core genome. This result highlights the discrepancy between the predicted fermentative character and the classification of *L. crispatus* as a homofermentative *Lactobacillus,* which would be expected to ferment hexoses via EMPP and to not possess PPP for pentose fermentation (Felis & Dellaglio 2007, Salvetti et al. 2012, Axelsson 2004, Hammes & Vogel 1995). Overall, the results of the metabolic pathway analyses in Study III emphasized the similarities between the isolates and supported the dependency of external nutrients, which a typical feature for lactobacilli (Hammes & Vogel 1995, Axelsson 2004).

As the colonization of the vagina by *L. crispatus* is predominantly associated with the healthy vaginal state (Ravel et al. 2011, Chaban et al. 2014, Gajer et al. 2012, Romero et al. 2014, El Aila et al. 2009, Gustafsson et al. 2011, Vasquez et al. 2002) and inversely associated with a variety of aberrant conditions such as BV (Verstraelen et al. 2009, Fredricks et al. 2007, Macklaim et al. 2013), the comparative genomics analysis of the ten *L. crispatus* isolates in Study III also focused on the molecular mechanisms by which the species could promote vaginal health. Notably, several of the identified features implicated in the production of antimicrobial substances or governing host-interactions were encoded by the core genome of the ten isolates (Study III). For instance, the core genome encoded two hydrogen peroxide-producing enzymes as well as two L-lactate dehydrogenases and one D-lactate dehydrogenase for the production of lactate; notably, both lactate and hydrogen peroxide are regarded as inhibitory towards harmful bacteria (Boris & Barbés 2000, Witkin et al. 2007). Furthermore, an additional L-lactate dehydrogenase-encoding gene was detected in nine isolates. For comparison, the genome of a previously sequenced isolate of another vaginal species, *L. iners,* has been annotated to contain only one L-lactate dehydrogenase encoding gene and no genes coding D-lactate dehydrogenase (Macklaim et al. 2011). The vast enzymatic repertoire of *L. crispatus* implicated in lactate production is particularly interesting as vaginal communities dominated by *L. crispatus* are associated with lower vaginal pH than are vaginal communities dominated by other bacteria, such as *L. iners* (Ravel et al. 2011).

Screening the predicted proteomes of the ten *L. crispatus* isolates for adhesion-associated PFAM domains revealed several putative adhesins belonging to 21 ortholog groups, seven of which also belonged to the core genome of the ten isolates (Table 3 and additional file 10 in Study III). Among these core adhesins was a putative fibronectin-binding protein, FbpA, which has also been detected in the genome analyses of, *e.g*, the GIT-associated *L. acidophilus* NCFM (Altermann et al. 2005) and the vaginal *L. iners* (Macklaim et al. 2011). In contrast, the search failed to identify the previously characterized adhesin LEA, which has been shown to mediate specific binding to both crop epithelium and epithelial cells of the human vagina

(Edelman et al. 2012). Thus, LEA protein might mediate adhesion via some as yet uncharacterized domain.

However, LEA was conserved in all of the ten strains, and when the predicted *L. crispatus* ortholog groups were screened for potential counterparts to core-encoded virulence factors of *Gardnerella vaginalis*, a species often associated with BV (Fredricks et al. 2005, Shipitsyna et al. 2013), LEA was among the reported matches. Moreover, LEA was predicted to be an antagonist to a pilus component of *G. vaginalis* and thus hypothesized to participate in the previously reported (Castro et al. 2013) ability of *L. crispatus* to reduce the adhesion of *G. vaginalis* to a human cell line. To test this hypothesis and to further support to the species-wise presence of LEA, the adhesion capacities of a BV-associated *G. vaginalis* isolate and a vaginal *L. crispatus* isolate not included in the pan- and core genome analysis were tested with and without pre-treatment with anti-LEA Fab fragments (Study III). Notably, the anti-LEA Fab fragments significantly reduced the adhesion of both of the isolates to the human vaginal cell line, whereas such inhibition was not evident following pre-treatment with control Fab fragments (Figure 5 in Study III). These results suggest that core-encoded LEA is a key mediator in the *L. crispatus* colonization and in the competitive exclusion of *G. vaginalis* in the vagina.

Despite the suggested prominence of core features to the biology of *L. crispatus* and a high degree of similarity between the isolates (Figure 7, Figures 1 and 2D in Study III), a few interesting genomic diversity regions were found in Study III. Similar to other *Lactobacillus* species (Berger et al. 2007, Siezen & van Hylckama Vlieg 2011, Raftis et al. 2011), the analyzed *L. crispatus* strains harbored highly variably *eps* gene clusters for the production of EPS, differing especially in their glycosyltransferase gene contents (Figure 4 in Study III). These differences presumably lead to variation in the sugar contents and structures of EPS. A few diversity regions also indicated clear differences between the vaginal isolates and the chicken isolate. Notably, all of the vaginal isolates included in Study III appeared to have a Type II CRISPR/Cas system, whereas a Type I CRISPR/Cas system was identified in the genome of strain ST1 (Table 2 and Figure 3 in Study III). Notably, strain ST1 was the only strain for which no prophage-like gene clusters were reported in the Prophinder analysis (Table I and Figure I in Study III). These findings might reflect possible differences in viral loads in the different life environments and are in accordance with the prevalence of lysogeny in vaginal *L. crispatus* (Damelin et al. 2011). Interestingly, analysis of the spacer sequence content of the CRISPR arrays revealed striking similarities between the spacers regions of the vaginal isolates, indicating either past encounters with common invading nucleoids or a recent mutual ancestor. The search for bacteriocin-like molecules also revealed interesting differences between the vaginal isolates and the chicken isolate. Although each isolate was predicted to encode a rather similar set of bacteriolysins, which lyse sensitive cells by catalyzing cell wall hydrolysis (Cotter et al. 2005), the vaginal isolates also contained genes that are implicated in the production of class II bacteriocins, which are small heat-stable peptides that insert themselves in the membrane of the target cell, induce membrane permeabilization, and result in cell death (Cotter et al. 2005) (Table 4 in Study III). Due to the limited availability of genome sequences of *L. crispatus* animal isolates, the niche-specific differences could not be addressed. Such issues remain to be resolved in the future in comparative genomics studies that include a large repertoire of both animal and human isolates.

5.3 Genomics of *P. freudenreichii* ssp. *shermanii* JS

P. freudenreichii ssp. *shermanii* JS has long been used in industrial cheese manufacturing but has also been considered as a bioprotective culture for use in fermented milk products (quark and yoghurt) and sourdough breads (Suomalainen & Mäyrä-Mäkinen 1999), as well as for probiotic preparations (Kajander et al. 2005, Kukkonen et al. 2007, Kukkonen et al. 2008). When consumed, strain JS survives transit through the GIT (Suomalainen et al. 2008, Hatakka et al. 2008) and can reduce serum levels of an inflammatory marker, C-reactive protein (Kekkonen et al. 2008). Administration of strain JS together with other potentially beneficial bacteria and prebiotic oligosaccharides has also been reported to reduce the incidence of respiratory infections (Kukkonen et al. 2008) and atopic allergies during the first two years of life (Kukkonen et al. 2007). Consumption of strain JS in a multispecies mixture has also been reported to alleviate the symptoms of irritable bowel syndrome (Kajander et al. 2005). In this thesis, this industrially relevant *Propionibacterium* strain was subjected to genomic analysis (Study IV). Using multiple sequencing approaches (Table 5), the genome of the strain JS was sequenced to approximately 691× coverage and closed, providing foundations for further analyses of the strain and generating a valuable addition to the pool of *P. freudenreichii* genome sequences.

The general characteristics and main annotation results of the final genome assembly of strain JS are summarized in Table 6 and described more in depth in Study IV. Briefly, the complete genome sequence of strain JS consists of a single circular chromosome of ~2.68 Mb, which is slightly larger than those of the dairy strains *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 (~2.62 Mb) (Falentin et al. 2010a) and *P. freudenreichii* ssp. *freudenreichii* ITG P20 (~2.59 Mb) (Le Marechal et al. 2015), the only *P. freudenreichii* strains with publically available annotated genome sequences at the time of Study IV. These three strains also demonstrated almost identical overall genomic GC contents (~67%), which fit well with the characteristic GC contents of both *P. freudenreichii* ssp. *shermanii* (64-67%) and ssp. *freudenreichii* (67%) (Patrick & McDowell 2012).

Study IV also highlighted further similarities between strain JS and the other genomically characterized *P. freudenreichii* strains. Whole-genome comparisons showed the genome sequences of both CIRM-BIA1 and ITG P20 to cover over 90% of the sequence of strain JS and to be highly similar to strain JS at the nucleotide level (Figure 3 and Table 1 in Study IV). Despite the high overall similarity and the presence of several synteny blocks, genomic rearrangements were also evident in whole-genome alignments between strain JS and strains CIRM-BIA1 and ITG P20 (Figure 3 in Study IV). Interestingly, the rearrangements likely resulted in past HGT events given that several putative GIs were identified at or near the non-syntenic regions (Figure 3 in Study IV). In accordance with earlier comparisons between representatives of *P. freudenreichii* and *P. acnes* (Falentin et al. 2010a), only weak synteny and similarity at the nucleotide level was observed between strain JS and a representative *P. acnes* strain (Figure 3 and Table 1 in Study IV).

Orthologous grouping of the protein-coding genes of strain JS and 29 other *Propionibacterium* strains representing 10 different species further exemplified the similarity between strain JS and the two other *P. freudenreichii* strains. As expected, strain JS shared most of its ortholog groups with the *P. freudenreichii* strains CIRM-BIA1 (~89%) and ITG P20 (~87%) (Figure 8), whereas only ~55 to ~56% was shared with the different *P. acnes* strains included in the study (Figure 2B in Study IV). The strains that were most dissimilar to strain JS were found among the cutaneous species *P. granulosum* and *P. acidifaciens.* Specifically, strain JS shared ~50% of its ortholog groups with these strains (Figure 2B in Study IV). The common core of all 30 strains analyzed comprised 546 ortholog groups, representing nearly 25% of the ortholog groups of strain JS. The core genome of the 30 strains was notably smaller than that obtained previously for *P. acidipropionici* ATCC4875, *P. acnes*

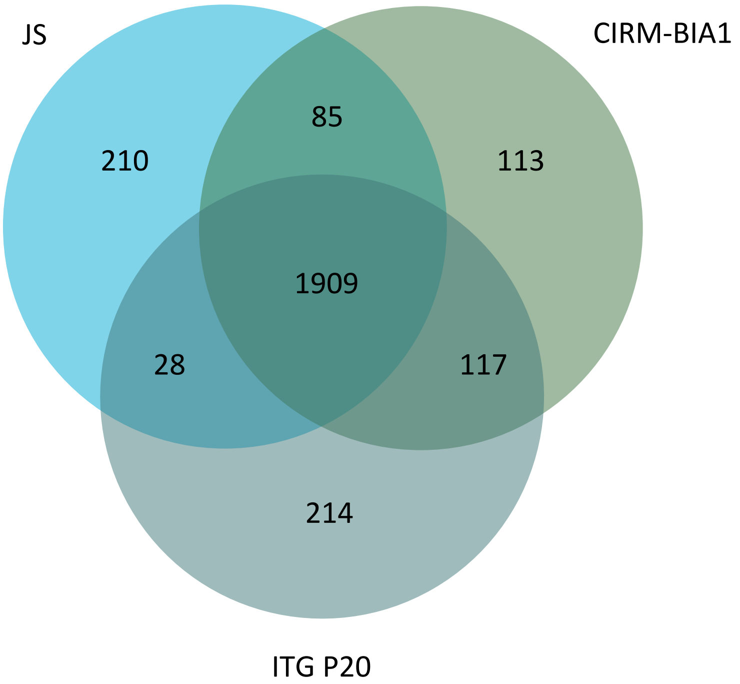


Figure 8 *Venn diagram showing the number of ortholog groups shared by the* P. freudenreichii *ssp.* shermanii *strains JS (Study IV) and CIRM-BIA1 (Falentin et al. 2010a) and* P. freudenreichii *ssp.* freudenreichii *strain ITG P20 (Le Marechal et al. 2015).*

KPA171202, and *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 (Table 3) (Parizzi et al. 2012), demonstrating the effect of dataset size on the size of the conserved core genome. Further validating that differences in core genome estimates primarily result from differences in dataset sizes, the orthologous grouping in Study IV found that strains ATCC4875, KPA171202, and CIRM-BIA1 had a similar number of ortholog groups in common as in a previous study (1,178 ortholog groups in Study IV versus 1,026 in Parizzi et al., 2012).

The genome analysis of strain JS also revealed the genetic background of several industrially relevant properties, such as the development of flavor and biosynthesis of vitamin B12 (Study IV). Similar to the dairy propionibacteria *P. freudenreichii* CIRM-BIA1 (Falentin et al. 2010a) and *P. acidipropionici* ATCC 4875 (Parizzi et al. 2012), strain JS possessed enzymes for the conversion of lactate into pyruvate and a Wood-Werkman cycle for the subsequent fermentation of pyruvate into propionate, a central flavor compound in Swiss-type cheeses (Figure 5 in Study IV). Moreover, pathways for the production of other flavor compounds such as acetate, diacetyl, and acetoin were also annotated (Figure 5 in Study IV). Unsurprisingly, all of the afore-mentioned enzymes had orthologous counterparts in the *P. freudenreichii* ssp. *shermanii* CIRM-BIA1, and the majority of these enzymes were also conserved in *P. freudenreichii* ssp. *freudenreichii* ITG P20. In addition, *P. freudenreichii* is considered to contribute to the aroma development of cheese via amino-acid catabolism and lipolysis (Thierry et al. 2011). Accordingly, strain JS was predicted to harbor several potential proteases and peptidases as well as enzymes for the degradation of various amino acids including serine, alanine, threonine, aspartate, leucine, valine, and isoleucine (Study IV). The catabolism of branched chain amino acids (leucine, isoleucine, and valine) is regarded as particularly important in the production of volatile short branched-chain fatty acids (Poonam et al. 2012). The genome of strain JS also encoded several putative esterases/lipases, one of which matched a previously characterized esterase A (Suoniemi & Tynkkynen 2002). The esterase A-encoding gene was absent from the other *P. freudenreichii* strains analyzed, but the amino acid metabolism-related enzymes were primarily conserved across the different *P. freudenreichii* strains included in Study IV.

Moreover, orthologous grouping of the predicted proteomes of the selected *Propionibacterium* strains in Study IV revealed strain JS to harbor the vitamin B12 biosynthesis pathway previously described in the *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 (Falentin et al. 2010a) and also indicated this pathway to be conserved in the *P. freudenreichii* ssp. *freudenreichii* ITG P20. Similar to strain CIRM-BIA1 (Falentin et al. 2010a), strain JS was also predicted to encode a bile/sodium symporter, which could underlie the observed ability of strain JS to withstand exposure to bile (Suomalainen et al. 2008) and suggests an adaptation to the GIT.

Taken together, the genomic investigation conducted in Study IV highlighted similarities between the different *P. freudenreichii* isolates and suggested a shared evolutionary origin for several industrially relevant properties.

5.4 Activity of *P. freudenreichii* ssp. *shermanii* JS during Cheese Ripening

A deeper understanding of the role of *P. freudenreichii* in cheese ripening was further expanded in this thesis (Study IV) by profiling the transcriptomes of industrial cheese samples obtained at two different stages of ripening process. Specifically, using RNA-seq, the transcriptomes were determined at the warm ripening (+20°C) stage and the subsequent cold ripening (+5°C) stage. This analysis added to earlier transcriptome-level information of the activity of *P. freudenreichii* during cheese ripening as these previous efforts focused on a very small subset of genes (Falentin et al. 2010b) or used conditions that mimicked the ripening process (Dalmasso et al. 2012). The results of the transcriptome profiling are discussed in more detail below and in Study IV.

Following previously established RPKM thresholds (Hackett et al. 2012), ~93% of CDS of strain JS could be identified as being expressed during cheese ripening (average RPKM > 0.125 across the warm or cold samples). Approximately 98% of the expressed CDSs were also highly expressed in either or both of the ripening stages (average RPKM > 10); a greater number of these CDSs demonstrated high expression levels in the cold than in the warm stage (2,128 versus 1,865 CDSs, respectively). The differential gene expression analysis performed in Study IV allowed a deeper understanding of the activities of strain JS at different stages of cheese ripening, revealing that nearly 15% of the CDSs were significantly differentially expressed (at least 2-fold difference) between the two time points. Interestingly, this fraction of genes is about ten percentage point lower than what has been previously reported for *P. freudenreichii* after transfer from warm (+30°C) to cold (+4°C) laboratory conditions that mimicked cheese ripening (Dalmasso et al. 2012). This difference could, in part, be explained by the different experimental set ups. The controlled laboratory conditions used in the afore mentioned study may have enabled the detection of expression differences that were not observed in the more variable industrial setting in Study IV. Over 60% of the differentially expressed CDSs of strain JS were up-regulated in the cold compared to the warm ripening stage, a result that is contradictory to the afore-mentioned global gene expression analysis, which suggested that the majority of the differentially expressed genes of *P. freudenreichii* are down-regulated in the cold (Dalmasso et al. 2012).

Despite the more prominent gene up-regulation in the cold ripening stage, the differential gene expression analysis suggested that strain JS grows and contributes to the flavor development primarily at the warm stage of cheese ripening (Study IV). For example, several CDSs up-regulated in the warm were associated with metabolic functions and aroma-formation, while the cold stage appeared to induce the expression of a notable number of CDSs associated with mobile elements or lacking a functional annotation. The enrichment analysis of the differentially expressed genes in different COG categories and KEGG reference pathways further revealed that the CDSs up-regulated at the warm stage were often associated with (i) energy production and conversion (COG category C); (ii) amino acid transport and metabolism (COG category E) (Figure 6 in Study IV); and (iii) metabolic pathways describing the global and large-scale metabolic functions (Supplemental file 2 in Study IV). These results indicated overall more active metabolism and growth during the warm than the cold ripening period. These findings are in accordance with previous studies (Dalmasso et al. 2012, Falentin et al. 2010b), suggesting that *P. freudenreichii* is in the exponential phase of growth during the warm stage of cheese ripening and that the activity of the bacterium's cell machinery and carbon metabolism is slower in the cold. The CDSs of strain JS that were up-regulated in the cold, on the other hand, were enriched in the arginine and proline metabolism reference pathway and in replication, recombination, and repair functions (COG category L) (Figure 6 in Study IV). The latter enrichment was largely driven by the cold-inducible mobile elements.

With regards to the formation of aroma compounds in the cheese, a few enzymes of the Wood-Werkman cycle and the associated TCA exhibited significantly higher expression in the warm than in the cold (Figure 5 in Study IV), which is in accordance with the production of majority of propionate in the cheese during the warm ripening period (Thierry et al. 2005). Acetate production, which is coupled to propionate production (Piveteau 1999), has also been reported to be more prominent in the warm than in the cold (Thierry et al. 2005). Accordingly, the enzyme catalyzing the first step of the formation of acetate from pyruvate was up-regulated in the warm ripening stage (Figure 5 in Study IV). The importance of the warm ripening period in flavor-formation was also highlighted by the higher expression of both subunits of the acetolactate synthase, the first enzyme of the pathway that leads to formation of the flavor compounds diacetyl and acetoin (Figure 5 in Study IV).

Although a more prominent flavor-compound producer in the warm than in the cold, *P. freudenreichii* has been suggested to contribute to cheese characteristics also in the cold (Thierry et al. 2005, Dalmasso et al. 2012). Thus, it was unsurprising that strain JS was observed to express several of its flavor-related CDSs at more or less similar levels between the two time points (Study IV). Notably, many CDSs involved in amino acid catabolism were up-regulated during the cold ripening stage. As the CDSs responsible for glutamate-glutamine interconversions were also down-regulated in the cold, the differential gene expression analysis suggested a broad reprogramming of amino-acid metabolism during this ripening stage. Some of the cold-induced amino acid metabolism-related genes were reported to be up-regulated at cold conditions in a previous gene expression analysis (Dalmasso et al. 2012) that examined *P. freudenreichii* ssp. *shermanii* CIRM-BIA1 in conditions mimicking cheese ripening; other such genes demonstrated different expression patterns between the studies.

The discrepancies between the transcriptome profiles obtained in Study IV and in the previous laboratory study (Dalmasso et al. 2012) most likely arise from the differences in the experimental set-ups, such as the use of different strains, settings, and analysis strategies. These differences highlight the value of industrial cheese samples in elucidating cheese ripening in an industrially relevant setting.

5.5 Annotation Transfer Evaluation

As exemplified in Studies I, III, and IV, accurate description of the biological processes of CDSs can greatly advance the understanding of a given organism. Producing accurate description for bacterial CDSs is, however, complicated by the poor availability and applicability of GO descriptions for bacterial genes as they are primarily used for describing eukaryotic functions (Ashburner et al. 2000, Klimke et al. 2011). Another hurdle is the propagation of inconsistent and erroneous annotations in the databases (Lee et al. 2007, Friedberg 2006, Schnoes et al. 2009). Thus, when the genome sequence of strain ST1 was first annotated (Study I), a variety of methods were applied (Table 5), and the annotation results were subjected to manual curation to determine the most suitable DE for each of the CDSs predicted to be encoded in the genome. In this evaluation, a special focus was placed on assessing the consistency and accuracy of DEs provided by the different methods as these factors were considered to be of relevance for subsequent genome projects. Furthermore, the ability of different strategies to reproduce correct DEs for a larger test set of well-annotated bacterial protein sequences was evaluated (Study II).

In Study II, various protein function prediction tools were applied to assign biological information to the CDSs of *L. crispatus* ST1. These methods included an *in*-house developed approach BLANNOTATOR (Study II), a fully automated RAST (Rapid Annotation using Subsystem Technology) service (Aziz et al. 2008), and the simple best-BLAST approach (Table 7). The methods were able to produce acceptable DEs for ~85% (BLANNOTATOR), ~58% (RAST), and ~69% (best-BLAST) of the predicted ST1 proteins with decipherable function (Table 7). Slightly different sets of query sequences remained without an assigned informative function when using these methods (Study II). These results support the use of multiple methods in the functional annotation of predicted proteomes, an approach that was also adopted in Studies III and IV. Notably, the application of a variety of different methods in the re-annotation of the CDSs of the ten *L. crispatus* isolates in Study III increased the overall number of CDSs with an assigned function by ~41% and standardized the quality of the annotation information for each of the isolates. These results underscore the importance of such approach in comparative genomic studies. In addition, when the predicted CDSs of strain JS were annotated in Study IV, BLANNOTATOR assigned DEs to practically all predicted proteins with homologous counterparts in the databases, whereas RAST assigned DEs to only ~71% (Table 7). The lower performance of RAST in the functional annotation of strains ST1 and JS compared to that of BLANNOTATOR

Table 7. *Performance of selected protein function prediction methods. ST1 and JS dataset columns indicate the percentage of acceptable DEs the given methods assigned to ST1 and JS protein sequences with decipherable functions, respectively. The SWISS-PROT dataset column shows the mean modified Levenshtein distance (mLD) between the correct and predicted DEs after excluding BLAST hits with sequence identity below 40% and coverage below 60%, as described in Study II.*

Approach	Description	ST1 dataset (%)	JS dataset (%)	SWISS-PROT dataset (mean mLD)
RAST	Relies on manually curated subsystems (*i.e.* collections of functional roles mapped onto genes across multiple genomes) and on protein families (FIGfams) largely	59	71	-
BLANNOTATOR	Assigns DEs to the query sequences from DE and GO information clustered database hits	85	100	0.42
Best-BLAST	DE of the most significant BLAST hit is transferred to the query	69	-	0.53
Top informative	Top informative BLAST match is transferred to the query	-	-	0.61
Most frequent	Most common DE among the BLAST hits is transferred to the query	-	-	0.46
Highest cumulative score	DE associated with the highest cumulative BLAST bit score is transferred to the query	-	-	0.46
Word score	Word-based scoring scheme	-	-	0.45

probably reflects the preference of RAST's annotation strategy to avoid false positive predictions at the cost of sensitivity (Aziz et al. 2008, Meyer et al. 2009).

In this thesis, the power of different strategies for selecting the most suitable DE from a list of BLAST hits was also examined (Study II). Namely, the ability of BLANNOTATOR and five other BLAST-based annotation strategies (Table 7) to reproduce correct DEs for a set of 3,090 well-annotated bacterial protein sequences was evaluated. A database preceding the functional characterization of the query proteins was used in this test. Notably, all of the strategies assigned reasonably accurate DEs to the queries. However, the strategies basing their prediction on information from multiple BLAST hits performed marginally better than others, with BLANNOTATOR being the best in the task (Figure 4 in Study II, Table 7). These findings are in accordance with previous observations that relying on multiple hits in functional annotation improves the accuracy of the functional description of the query sequence and increases the number of functionally classified CDSs in a genome (Martin et al. 2004, Hawkins et al. 2006).

To aid the parameter choices in the homology searches, the effect of sequence identity and coverage on the annotation quality was also examined in Study II. Notably, strict thresholds and the subsequent reduction in the pool of sequences applicable for annotation transfer negatively affected the mean annotation quality (Figure 3 and Additional file 1 in Study II). Identity thresholds seemed to have a bigger effect on the results than coverage thresholds. A notable portion of the suitable DEs in the test data were lost at a ~60% identity threshold, whereas an identity of nearly 40% was often sufficient for good annotation quality. With regards to coverage, unfavorable effects on the annotation quality were observed when nearly 100% coverage over the subject was required. These data thus argue against the use of strict identity thresholds in functional annotation and are in accordance with previous observations that sequence identity as low as 30% can be sufficient for annotation transfer (Devos & Valencia 2000).

The evaluations performed in Study II also highlighted the difficulties involved in the benchmarking. Reflecting the poor availability of the GO annotations to bacterial proteins (Ashburner et al. 2000, Klimke et al. 2011), the DEs were found to be a better and more comprehensive source of annotation information and were preferred in describing the functions of bacterial protein sequences. However, the DEs are “non-standardized” and suffer somewhat from synonymous and ambiguous expressions, which complicate the comparison of different annotations. Manual curation, such as was performed in the functional annotation of the predicted proteome of strain ST1 in Study II, can relatively easily identify highly dissimilar yet synonymous DEs as matching and highly similar DEs describing different functions as different; however, this process is laborious and time-consuming. Therefore, Study II used a faster and more scalable computational assessment to evaluate the different strategies for selecting the most suitable DE from a list of BLAST hits (Table 7) and in the examination of the effect of sequence identity and coverage on annotation quality. This approach, in which the fraction of character changes between the words in DEs was measured computationally, was effective in producing quantifiable and unbiased information on the similarities of the different DEs. However, it incorrectly identified the highly dissimilar yet synonymous DEs as different and the highly similar DEs describing different functions as synonymous. These observations underscore the impact the evaluation metric has on assessment, which was highlighted in a recent systematic assessment of the ability of 54 methods to correctly assign GO annotations to a set of protein sequences (Radivojac et al. 2013).

6 Conclusions

The sequencing-based analysis of bacterial genomes and transcriptomes has enabled a previously unimaginable view of the characteristics and functions of bacteria (Forde & O'Toole 2013, Hall 2007, Medini et al. 2008). In this thesis, novel insights into the physiology of *L. crispatus* and *P. freudenreichii* were obtained by applying modern sequencing methods and down-stream analyses. While the annotation of genome sequences of *L. crispatus* ST1 (Study I) and *P. freudenreichii* ssp. *shermanii* JS (Study IV) revealed the genetic blueprints of these individual strains and provided solid foundations for further investigations, the subsequent functional genomics studies provided in depth knowledge on how *L. crispatus* can support vaginal health (Studies III) and how *P. freudenreichii* adapts to changing conditions (Study IV). Another key aspect of this thesis was to understand how well automated annotation transfer methods perform in bacterial protein function prediction. Importantly, the different methods appeared to be almost equal in quality to manual curation and were observed to assign microbiologically informative functions to a vast majority of proteins, greatly facilitating genome annotation (Study II).

Specifically, comparative genomics of closely related bacterial strains proved to be a particularly valuable approach in describing genes responsible for the characteristic traits of a group of bacteria, as exemplified in Study III. Although the bacterial components of strain ST1 were successfully characterized in Study I, the comparative analysis relying on distantly related lactobacilli provided only a limited view of the species-wide significance and health promoting potential of the newly annotated features. In contrast, the comparative genomics analysis of strain ST1 and nine vaginal *L. crispatus* isolates in Study III revealed that the genomic backbone of *L. crispatus* includes several genes of potential importance to vaginal health. Included in this list of core genes was a previously characterized adhesin LEA (Edelman et al. 2012). In Study III, this adhesin was experimentally shown to be a probable antagonist to BV-associated *G. vaginalis*, suggesting that health-promoting features are universally present in the species rather than specific to individual *L. crispatus* strains. Study III also highlighted that a rather limited set of genomes was sufficient for defining the conserved core features of a species, as has been previously reported for other *Lactobacillus* sp. core genome models (Smokvina et al. 2013, Nelson et al. 2010, Broadbent et al. 2012). In contrast, substantially more genomes are required to fully capture the genomic variation within a species and for resolving gene-niche relationships. Particularly, more non-vaginal *L. crispatus* isolates should be subjected to genome sequencing to allow investigations of the potential niche-specific subgroups and features of this species. Similarly to the initial comparative genomic analysis of strain ST1, comparative genomics analysis of strain JS was complicated by the limited availability of *P. freudenreichii* genomes at the time of the genome analysis (Study IV). However, the increasing amount of genome data available for *P.*

freudenreichii should soon allow a means to infer the genomic basis and species-wide distribution of industrially relevant traits of *P. freudenreichii*, similar to the successful identification of possible key microbe-host interaction factors of *L. crispatus* in Study III.

While genomics approaches provides a means to catalogue the genomic building blocks of bacteria, transcriptomics describes the extent and complexity of transcriptomes and allows the quantification of the expression levels of transcripts under different conditions. In Study IV, RNA-sequencing based transcriptome profiling was used to advance understanding of the role of strain JS during cheese ripening. This strategy allowed linking the different enzymes and pathways identified in the genome analysis with specific stages of industrial cheese ripening. This analysis also provided a previously unrealized view to the dynamic nature of the flavor-forming mechanisms of strain JS. The transcriptome profiling revealed this strain to be more metabolically active and to produce more propionate at the warm than cold stage of cheese ripening. These results complemented prior microarray based gene expression studies of *P. freudenreichii* (Dalmasso et al. 2012, Saraoui et al. 2013) and provided an interesting starting point for the industrial optimization of cheese ripening. In the future, selection of the most optimal *P. freudenreichii* strain for a particular industrial application could be aided by transcriptome profiling and be based on the activity of genes governing the desired metabolic conversions. Similarly, transcriptome profiling could provide an attractive means to investigate the role of the interaction factors of *L. crispatus* in different environments and conditions and could be used to validate the function of *lea* in protecting the vagina against pathogenic microorganisms.

Taken together, the sequencing- and sequence analysis-based approaches applied in this thesis were powerful tools in deciphering and studying bacteria and greatly extended our knowledge of the physiology of *L. crispatus* and *P. freudenreichii*. Specifically, genome sequencing and comparative genomics were highly valuable in characterizing the genomic building blocks underlying phenotypic traits, while genome-wide transcriptome profiling elucidated the importance of different genomic features in different conditions, providing a view of the dynamic nature of different processes. The studies also generated various new hypotheses on how lactobacilli can support vaginal health and how propionibacteria contribute to the cheese ripening, providing important steps forward in understanding the role of these bacteria in the studied environments. It should, however, be noted that while sequencing is powerful tool in discovery science, it is most powerful when coupled with carefully designed and hypothesis-driven studies.

Acknowledgements

This thesis work was mainly carried out at DNA sequencing and genomics laboratory at Institute of Biotechnology, University of Helsinki, with the financial support from Viikki Doctoral Programme in Molecular Biosciences (VGSB) and Integrative Life Science Doctoral Program (ILS).

I am grateful to everyone involved in this process. I am most grateful to my thesis supervisors Docent Petri Auvinen and Professor Liisa Holm for all the help, support, and guidance they have provided over the years. I thank my thesis advisory committee members Professors Kaarina Sivonen and Benita Westerlund-Wikström for the helpful discussions and valuable advice regarding this thesis project.

Assistant Professor Mirko Rossi and Professor Per Saris are thanked for helpful suggestions and constructive critiques on this thesis during the preliminary examination process.

All the co-authors and collaborators in Finland and abroad are greatly thanked for their contribution; without you this would not have been possible.

I thank all my co-workers at the DNA sequencing and genomics laboratory for the numerous scientific and non-scientific discussions. Finally, I owe my deepest gratitude to my family for their presence and support.

Teija Ojala

Helsinki, December 2015

References

Abbas Hilmi, H. T., Surakka, A., Apajalahti, J. & Saris, P. E. 2007, "Identification of the most abundant lactobacillus species in the crop of 1- and 5-week-old broiler chickens", *Appl. Environ. Microbiol.,* vol. 73, no. 24, pp. 7867-7873.

Altermann, E., Russell, W. M., Azcarate-Peril, M. A., Barrangou, R., Buck, B. L., McAuliffe, O., et al. 2005, "Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 102, no. 11, pp. 3906-3912.

Altschul, S. F., Madden, T.L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. 1997, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.,* vol. 25, no. 17, pp. 3389-3402.

Anisimova, M., Liberles, D.A., Philippe, H., Provan, J., Pupko, T. & von Haeseler, A. 2013, "State-of the art methodologies dictate new standards for phylogenetic analysis", *BMC Evol. Biol.,* vol. 13, pp. 161-2148-13-161.

Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S. & Smith, R. D. 2008, "Proteogenomics: needs and roles to be filled by proteomics in genome annotation", *Brief Funct Genomic Proteomic,* vol. 7, no. 1, pp. 50-62.

Anukam, K. C., Macklaim, J.M., Gloor, G. B., Reid, G., Boekhorst, J., Renckens, B., et al. 2013, "Genome sequence of *Lactobacillus pentosus* KCA1: vaginal isolate from a healthy premenopausal woman", *PLoS One,* vol. 8, no. 3, pp. e59239.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. 2000, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", *Nat. Genet.,* vol. 25, no. 1, pp. 25-29.

Axelsson, L. 2004, "Lactic acid bacteria: Classification and physiology" in *Lactic acid bacteria: microbiological and functional aspects*, eds. S. Salminen, A. von Wright & A. Ouwehand, 3rd edn, Marcel Dekker Inc., New York, pp. 1-66.

Azcarate-Peril, M. A., Altermann, E., Goh, Y. J., Tallon, R., Sanozky-Dawes, R. B., Pfeiler, E. A., et al. 2008, "Analysis of the genome sequence of *Lactobacillus gasseri* ATCC 33323 reveals the molecular basis of an autochthonous intestinal organism", *Appl. Environ. Microbiol.,* vol. 74, no. 15, pp. 4610-4625.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. 2008, "The RAST Server: rapid annotations using subsystems technology", *BMC Genomics,* vol. 9, pp. 75-2164-9-75.

Barrangou, R. & Horvath, P. 2012, "CRISPR: new horizons in phage resistance and strain identification", *Annu. Rev. Food Sci. Technol.,* vol. 3, pp. 143-162.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. 2013, "NCBI GEO: archive for functional genomics data sets--update", *Nucleic Acids Res.,* vol. 41, no. Database issue, pp. D991-5.

Bartels, D., Kespohl, S., Albaum, S., Druke, T., Goesmann, A., Herold, J., et al. 2005, "BACCardI--a tool for the validation of genomic assemblies, assisting genome finishing and intergenome comparison", *Bioinformatics,* vol. 21, no. 7, pp. 853-859.

Batzoglou, S. 2005, "The many faces of sequence alignment", *Brief Bioinform,* vol. 6, no. 1, pp. 6-22.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. 2013, "GenBank", *Nucleic Acids Res.,* vol. 41, no. Database issue, pp. D36-42.

Bentley, D. R. 2006, "Whole-genome re-sequencing", *Curr. Opin. Genet. Dev.,* vol. 16, no. 6, pp. 545-552.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. 2008, "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature,* vol. 456, no. 7218, pp. 53-59.

Bentley, S. 2009, "Sequencing the species pan-genome", *Nat. Rev. Microbiol.,* vol. 7, no. 4, pp. 258-259.

Berger, B., Pridmore, R. D., Barretto, C., Delmas-Julien, F., Schreiber, K., Arigoni, F., et al. 2007, "Similarity and differences in the *Lactobacillus acidophilus* group identified by polyphasic analysis and comparative genomics", *J. Bacteriol.,* vol. 189, no. 4, pp. 1311-1321.

Binnewies, T. T., Motro, Y., Hallin, P. F., Lund, O., Dunn, D., La, T., et al. 2006, "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries", *Funct. Integr. Genomics,* vol. 6, no. 3, pp. 165-185.

Boekhorst, J., Siezen, R. J., Zwahlen, M. C., Vilanova, D., Pridmore, R. D., Mercenier, A., et al. 2004, "The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content", *Microbiology,* vol. 150, no. Pt 11, pp. 3601-3611.

Boris, S. & Barbés, C. 2000, "Role played by lactobacilli in controlling the population of vaginal pathogens", *Microbes Infect.,* vol. 2, no. 5, pp. 543-546.

Broadbent, J. R., Neeno-Eckwall, E. C., Stahl, B., Tandee, K., Cai, H., Morovic, W., et al. 2012, "Analysis of the *Lactobacillus casei* supragenome and its influence in species evolution and lifestyle adaptation", *BMC Genomics,* vol. 13, pp. 533-2164-13-533.

Brouwer, R. W., Kuipers, O. P. & van Hijum, S. A. 2008, "The relative value of operon predictions", *Brief Bioinform,* vol. 9, no. 5, pp. 367-375.

Bruggemann, H., Henne, A., Hoster, F., Liesegang, H., Wiezer, A., Strittmatter, A., et al. 2004, "The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin", *Science,* vol. 305, no. 5684, pp. 671-673.

Brzuszkiewicz, E., Weiner, J., Wollherr, A., Thurmer, A., Hupeden, J., Lomholt, H. B., et al. 2011, "Comparative genomics and transcriptomics of *Propionibacterium acnes*", *PLoS One,* vol. 6, no. 6, pp. e21581.

Buck, B. L., Altermann, E., Svingerud, T. & Klaenhammer, T. R. 2005, "Functional analysis of putative adhesion factors in *Lactobacillus acidophilus* NCFM", *Appl. Environ. Microbiol.,* vol. 71, no. 12, pp. 8344-8351.

Bull, M. J., Jolley, K. A., Bray, J. E., Aerts, M., Vandamme, P., Maiden, M.C., et al. 2014, "The domestication of the probiotic bacterium *Lactobacillus acidophilus*", *Sci. Rep.,* vol. 4, pp. 7202.

Burdock, G. A. & Carabin, I. G. 2004, "Generally recognized as safe (GRAS): history and description", *Toxicol. Lett.,* vol. 150, no. 1, pp. 3-18.

Cai, H., Thompson, R., Budinich, M. F., Broadbent, J. R. & Steele, J. L. 2009, "Genome sequence and comparative genome analysis of *Lactobacillus casei*: insights into their niche-associated evolution", *Genome Biol. Evol.,* vol. 1, pp. 239-257.

Callanan, M., Kaleta, P., O'Callaghan, J., O'Sullivan, O., Jordan, K., McAuliffe, O., et al. 2008, "Genome sequence of *Lactobacillus helveticus*, an organism distinguished by selective gene loss and insertion sequence element expansion", *J. Bacteriol.,* vol. 190, no. 2, pp. 727-735.

Canchaya, C., Claesson, M.J., Fitzgerald, G. F., van Sinderen, D. & O'Toole, P.W. 2006, "Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species", *Microbiology,* vol. 152, no. Pt 11, pp. 3185-3196.

Canfield, D. E., Glazer, A. N. & Falkowski, P. G. 2010, "The evolution and future of Earth's nitrogen cycle", *Science,* vol. 330, no. 6001, pp. 192-196.

Caplice, E. & Fitzgerald, G. F. 1999, "Food fermentations: role of microorganisms in food production and preservation", *Int. J. Food Microbiol.,* vol. 50, no. 1-2, pp. 131-149.

Casjens, S. 1998, "The diverse and dynamic structure of bacterial genomes", *Annu. Rev. Genet.,* vol. 32, pp. 339-377.

Castro, J., Henriques, A., Machado, A., Henriques, M., Jefferson, K. K. & Cerca, N. 2013, "Reciprocal interference between *Lactobacillus* spp. and *Gardnerella vaginalis* on initial adherence to epithelial cells", *Int.J.Med.Sci.,* vol. 10, no. 9, pp. 1193-1198.

Chaban, B., Links, M. G., Jayaprakash, T. P., Wagner, E. C., Bourque, D. K., Lohn, Z., et al. 2014, "Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle", *Microbiome,* vol. 2, pp. 23-2618-2-23. eCollection 2014.

Chaillou, S., Champomier-Verges, M. C., Cornet, M., Crutz-Le Coq, A. M., Dudez, A. M., Martin, V., et al. 2005, "The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K", *Nat. Biotechnol.,* vol. 23, no. 12, pp. 1527-1533.

Chain, P. S., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., et al. 2009, "Genomics. Genome project standards in a new era of sequencing", *Science,* vol. 326, no. 5950, pp. 236-237.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., et al. 2005, "VFDB: a reference database for bacterial virulence factors", *Nucleic Acids Res.,* vol. 33, no. Database issue, pp. D325-8.

Cho, S., Cho, Y., Lee, S., Kim, J., Yum, H., Kim, S. C., et al. 2013, "Current challenges in bacterial transcriptomics", *Genomics Inform.,* vol. 11, no. 2, pp. 76-82.

Claesson, M. J., van Sinderen, D. & O'Toole, P.W. 2008, "*Lactobacillus* phylogenomics--towards a reclassification of the genus", *Int. J. Syst. Evol. Microbiol.,* vol. 58, no. Pt 12, pp. 2945-2954.

Clark, W. T. & Radivojac, **P.** 2011, "Analysis of protein function and its prediction from amino acid sequence", *Proteins,* vol. 79, no. 7, pp. 2086-2096.

Collins, F. S., Morgan, M. & Patrinos, A. 2003, "The Human Genome Project: lessons from large-scale biology", *Science,* vol. 300, no. 5617, pp. 286-290.

Cotter, P. D., Hill, C. & Ross, R. P. 2005, "Bacteriocins: developing innate immunity for food", *Nat Rev Micro,* vol. 3, no. 10, pp. 777-788.

Cousin, F. J., Deutsch, S., Perez Chaia, A., Foligne, B. & Jan, G. 2012, "Interactions between probiotic dairy propionibacteria and the intestinal epithelium", *Current Immunology Reviews,* vol. 8, no. 3, pp. 216-226.

Cousin, F. J., Mater, D. D. G., Foligne, B. & Jan, G. 2011, "Dairy propionibacteria as human probiotics: A review of recent evidence", *Dairy Science & Technology,* vol. 91, no. 1, pp. 1-26.

Crossman, L. C., Gould, V. C., Dow, J. M., Vernikos, G. S., Okazaki, A., Sebaihia, M., et al. 2008, "The complete genome, comparative and functional analysis of *Stenotrophomonas maltophilia* reveals an organism heavily shielded by drug resistance determinants", *Genome Biol.,* vol. 9, no. 4, pp. R74-2008-9-4-r74.

Dalmasso, M., Aubert, J., Briard-Bion, V., Chuat, V., Deutsch, S. M., Even, S., et al. 2012, "A temporal-omic study of *Propionibacterium freudenreichii* CIRM-BIA1 adaptation strategies in conditions mimicking cheese ripening in the cold", *PLoS One,* vol. 7, no. 1, pp. e29083.

Damelin, L. H., Paximadis, M., Mavri-Damelin, D., Birkhead, M., Lewis, D. A. & Tiemessen, C. T. 2011, "Identification of predominant culturable vaginal *Lactobacillus* species and associated bacteriophages from women with and without vaginal discharge syndrome in South Africa", *J. Med. Microbiol.,* vol. 60, no. Pt 2, pp. 180-183.

Danielsson, D., Teigen, P. K. & Moi, H. 2011, "The genital econiche: focus on microbiota and bacterial vaginosis", *Ann. N. Y. Acad. Sci.,* vol. 1230, pp. 48-58.

Darling, A. E., Mau, B. & Perna, N.T. 2010, "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement", *PLoS One,* vol. 5, no. 6, pp. e11147.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. 2014, "Diet rapidly and reproducibly alters the human gut microbiome", *Nature,* vol. 505, no. 7484, pp. 559-563.

de Angelis, M., Siragusa, S., Berloco, M., Caputo, L., Settanni, L., Alfonsi, G., et al. 2006, "Selection of potential probiotic lactobacilli from pig feces to be used as additives in pelleted feeding", *Res. Microbiol.,* vol. 157, no. 8, pp. 792-801.

de Freitas, R., Madec, M., Chuat, V., Maillard, M. B., Abeiion Mukdsi, M. C., Falentin, H., et al. 2015, "New insights about phenotypic heterogeneity within *Propionibacterium freudenreichii* argue against its division into subspecies", *Dairy Science & Technology,* , pp. 1-13.

Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. 2007, "Identifying bacterial genes and endosymbiont DNA with Glimmer", *Bioinformatics,* vol. 23, no. 6, pp. 673-679.

Dessinioti, C. & Katsambas, A. D. 2010, "The role of *Propionibacterium acnes* in acne pathogenesis: facts and controversies", *Clin. Dermatol.,* vol. 28, no. 1, pp. 2-7.

Deveau, H., Garneau, J. E. & Moineau, S. 2010, "CRISPR/Cas system and its role in phage-bacteria interactions", *Annu. Rev. Microbiol.,* vol. 64, pp. 475-493.

Devos, D. & Valencia, A. 2000, "Practical limits of function prediction", *Proteins,* vol. 41, no. 1, pp. 98-107.

D'haeseleer, P. 2005, "How does gene expression clustering work?", *Nat. Biotechnol.,* vol. 23, no. 12, pp. 1499-1501.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. 2013, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis", *Brief Bioinform,* vol. 14, no. 6, pp. 671-683.

Douillard, F. P., Ribbera, A., Kant, R., Pietilä, T. E., Järvinen, H. M., Messing, M., et al. 2013, "Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG", *PLoS Genet.,* vol. 9, no. 8, pp. e1003683.

Edelman, S., Leskelä, S., Ron, E., Apajalahti, J. & Korhonen, T. K. 2003, "In vitro adhesion of an avian pathogenic *Escherichia coli* O78 strain to surfaces of the chicken intestinal tract and to ileal mucus", *Vet. Microbiol.,* vol. 91, no. 1, pp. 41-56.

Edelman, S., Westerlund-Wikström, B., Leskelä, S., Kettunen, H., Rautonen, N., Apajalahti, J., et al. 2002, "In vitro adhesion specificity of indigenous Lactobacilli within the avian intestinal tract", *Appl. Environ. Microbiol.,* vol. 68, no. 10, pp. 5155-5159.

Edelman, S. M., Lehti, T. A., Kainulainen, V., Antikainen, J., Kylväjä, R., Baumann, M., et al. 2012, "Identification of a high-molecular-mass *Lactobacillus* epithelium adhesin (LEA) of *Lactobacillus crispatus* ST1 that binds to stratified squamous epithelium", *Microbiology,* vol. 158, no. Pt 7, pp. 1713-1722.

Edgar, R. C. 2007, "PILER-CR: fast and accurate identification of CRISPR repeats", *BMC Bioinformatics,* vol. 8, pp. 18.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. 2009, "Real-time DNA sequencing from single polymerase molecules", *Science,* vol. 323, no. 5910, pp. 133-138.

Eisen, J.A. 1998, "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis", *Genome Res.,* vol. 8, no. 3, pp. 163-167.

El Aila, N.A., Tency, I., Claeys, G., Verstraelen, H., Saerens, B., Santiago, G. L., et al. 2009, "Identification and genotyping of bacteria from paired vaginal and rectal samples from pregnant women indicates similarity between vaginal and rectal microflora", *BMC Infect. Dis.,* vol. 9, pp. 167-2334-9-167.

El Kafsi, H., Binesse, J., Loux, V., Buratti, J., Boudebbouze, S., Dervyn, R., et al. 2014, "*Lactobacillus delbrueckii* ssp. *lactis* and ssp. *bulgaricus*: a chronicle of evolution in action", *BMC Genomics,* vol. 15, pp. 407-2164-15-407.

Falentin, H., Deutsch, S.M., Jan, G., Loux, V., Thierry, A., Parayre, S., et al. 2010a, "The complete genome of *Propionibacterium freudenreichii* CIRM-BIA1, a hardy actinobacterium with food and probiotic applications", *PLoS One,* vol. 5, no. 7, pp. e11748.

Falentin, H., Postollec, F., Parayre, S., Henaff, N., Le Bivic, P., Richoux, R., et al. 2010b, "Specific metabolic activity of ripening bacteria quantified by real-time reverse transcription PCR throughout Emmental cheese manufacture", *Int. J. Food Microbiol.,* vol. 144, no. 1, pp. 10-19.

Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. O. 2009, "Reconstruction of biochemical networks in microorganisms", *Nat. Rev. Microbiol.,* vol. 7, no. 2, pp. 129-143.

Felis, G. E. & Dellaglio, F. 2007, "Taxonomy of Lactobacilli and Bifidobacteria", *Curr. Issues Intest Microbiol.,* vol. 8, no. 2, pp. 44-61.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. 2014, "Pfam: the protein families database", *Nucleic Acids Res.,* vol. 42, no. Database issue, pp. D222-30.

Fitz-Gibbon, S., Tomida, S., Chiu, B. H., Nguyen, L., Du, C., Liu, M., et al. 2013, "*Propionibacterium acnes* strain populations in the human skin microbiome associated with acne", *J. Invest. Dermatol.,* vol. 133, no. 9, pp. 2152-2160.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. 1995, "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd", *Science,* vol. 269, no. 5223, pp. 496-512.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. 2010, "Direct detection of DNA methylation during single-molecule, real-time sequencing", *Nat. Methods,* vol. 7, no. 6, pp. 461-465.

Forde, B. M. & O'Toole, P. W. 2013, "Next-generation sequencing technologies and their impact on microbial genomics", *Brief Funct. Genomics,* vol. 12, no. 5, pp. 440-453.

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A., et al. 2013, "Country-specific antibiotic use practices impact the human gut resistome", *Genome Res.,* vol. 23, no. 7, pp. 1163-1169.

França, L.T., Carrilho, E. & Kist, T. B. 2002, "A review of DNA sequencing techniques", *Q. Rev. Biophys.,* vol. 35, no. 02, pp. 169-200.

Francke, C., Siezen, R. J. & Teusink, B. 2005, "Reconstructing the metabolic network of a bacterium from its genome", *Trends Microbiol.,* vol. 13, no. 11, pp. 550-558.

Fredricks, D. N., Fiedler, T. L. & Marrazzo, J. M. 2005, "Molecular identification of bacteria associated with bacterial vaginosis", *N. Engl. J. Med.,* vol. 353, no. 18, pp. 1899-1911.

Fredricks, D. N., Fiedler, T. L., Thomas, K.K., Oakley, B. B. & Marrazzo, J. M. 2007, "Targeted PCR for detection of vaginal bacteria associated with bacterial vaginosis", *J. Clin. Microbiol.,* vol. 45, no. 10, pp. 3270-3276.

Frese, S. A., Benson, A. K., Tannock, G. W., Loach, D. M., Kim, J., Zhang, M., et al. 2011, "The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*", *PLoS Genet.,* vol. 7, no. 2, pp. e1001314.

Friedberg, I. 2006, "Automated protein function prediction--the genomic challenge", *Brief Bioinform,* vol. 7, no. 3, pp. 225-242.

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schutte, U. M., Zhong, X., et al. 2012, "Temporal dynamics of the human vaginal microbiota", *Sci. Transl. Med.,* vol. 4, no. 132, pp. 132ra52.

Gardy, J. L. & Brinkman, F. S. 2006, "Methods for predicting bacterial protein subcellular localization", *Nat. Rev. Microbiol.,* vol. 4, no. 10, pp. 741-751.

Garrity, G., M., Lilburn, T., Cole, J. R., Harrison, S., H., Euzeby, J. & Tindall, B. J. 2007a, "Part 9 – The *Bacteria*: Phylum *Firmicutes*: Class "*Bacilli*" in *Taxonomic Outline of the Bacteria and Archaea, Release 7.7*, ed. G. Garrity M., Michigan State University Board of Trustees.

Garrity, G. M., Lilburn, T., Cole, J. R., Harrison, S. H., Euzeby, J. & Tindall, B. J. 2007b, *Taxonomic Outline of the Bacteria and Archaea, Release 7.7,* Michigan State University Board of Trustees.

Giraffa, G., Chanishvili, N. & Widyastuti, Y. 2010, "Importance of lactobacilli in food and feed biotechnology", *Res. Microbiol.,* vol. 161, no. 6, pp. 480-487.

Goodfellow, M. 2012, "Phylum XXVI *Actinobacteria* phyl. nov." in *Bergey's manual of systematic bacteriology. Vol. 5, the actinobacteria*, eds. M. Goodfellow, P. Kampfer, H.J. Busse, et al., 2nd ed. edn, Springer, New York, pp. 33.

Gustafsson, R. J., Ahrne, S., Jeppsson, B., Benoni, C., Olsson, C., Stjernquist, M., et al. 2011, "The *Lactobacillus* flora in vagina and rectum of fertile and postmenopausal healthy Swedish women", *BMC Womens Health,* vol. 11, no. 1, pp. 17-6874-11-17.

Haaber, J., Friberg, C., McCreary, M., Lin, R., Cohen, S. N. & Ingmer, H. 2015, "Reversible Antibiotic Tolerance Induced in *Staphylococcus aureus* by Concurrent Drug Exposure", *MBio,* vol. 6, no. 1, pp. 10.1128/mBio.02268-14.

Hackett, N. R., Butler, M. W., Shaykhiev, R., Salit, J., Omberg, L., Rodriguez-Flores, J. L., et al. 2012, "RNA-Seq quantification of the human small airway epithelium transcriptome", *BMC Genomics,* vol. 13, pp. 82-2164-13-82.

Haft, D. H., Selengut, J. D. & White, O. 2003, "The TIGRFAMs database of protein families", *Nucleic Acids Res.,* vol. 31, no. 1, pp. 371-373.

Hall, N. 2007, "Advanced sequencing technologies and their wider impact in microbiology", *J. Exp. Biol.,* vol. 210, no. Pt 9, pp. 1518-1525.

Hammes, W. P. & Vogel, R. F. 1995, "The genus *Lactobacillus*" in *The genera of lactic acid bacteria*, eds. Wood, B.,J.,B. & W. Holzapfel H., Blackie Academic & Professional, Glasgow, pp. 19-54.

Hatakka, K., Holma, R., El-Nezami, H., Suomalainen, T., Kuisma, M., Saxelin, M., et al. 2008, "The influence of *Lactobacillus rhamnosus* LC705 together with *Propionibacterium freudenreichii* ssp. *shermanii* JS on potentially carcinogenic bacterial activity in human colon", *Int. J. Food Microbiol.,* vol. 128, no. 2, pp. 406-410.

Hawkins, T., Luban, S. & Kihara, D. 2006, "Enhanced automated function prediction using distantly related sequences and contextual association by PFP", *Protein Sci.,* vol. 15, no. 6, pp. 1550-1556.

Hooper, L.V., Littman, D. R. & Macpherson, A. J. 2012, "Interactions between the microbiota and the immune system", *Science,* vol. 336, no. 6086, pp. 1268-1273.

Hsiao, W. W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B. & Brinkman, F. S. 2005, "Evidence of a large novel gene pool associated with prokaryotic genomic islands", *PLoS Genet.,* vol. 1, no. 5, pp. e62.

Hurmalainen, V., Edelman, S., Antikainen, J., Baumann, M., Lahteenmäki, K. & Korhonen, T. K. 2007, "Extracellular proteins of *Lactobacillus crispatus* enhance activation of human plasminogen", *Microbiology,* vol. 153, no. Pt 4, pp. 1112-1122.

Hutchison, C. A. 3rd. 2007, "DNA sequencing: bench to bedside and beyond", *Nucleic Acids Res.,* vol. 35, no. 18, pp. 6227-6237.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. 2010, "Prodigal: prokaryotic gene recognition and translation initiation site identification", *BMC Bioinformatics,* vol. 11, pp. 119-2105-11-119.

Ip, C. L., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., et al. 2015, "MinION Analysis and Reference Consortium: Phase 1 data release and analysis", *F1000Research,* vol. 4.

Janvier, F., Delacour, H., Larreche, S., Abdalla, S., Aubert, P. & Merens, A. 2013, "Abdominal wall and intra-peritoneal abscess by *Propionibacterium avidum* as a complication of abdominal parietoplasty", *Pathol. Biol. (Paris),* vol. 61, no. 5, pp. 223-225.

Jorth, P., Trivedi, U., Rumbaugh, K. & Whiteley, M. 2013, "Probing bacterial metabolism during infection using high-resolution transcriptomics", *J. Bacteriol.,* vol. 195, no. 22, pp. 4991-4998.

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N. & Whiteley, M. 2014, "Metatranscriptomics of the human oral microbiome during health and disease", *MBio,* vol. 5, no. 2, pp. e01012-14.

Kajander, K., Hatakka, K., Poussa, T., Farkkila, M. & Korpela, R. 2005, "A probiotic mixture alleviates symptoms in irritable bowel syndrome patients: a controlled 6-month intervention", *Aliment. Pharmacol. Ther.,* vol. 22, no. 5, pp. 387-394.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. 2012, "KEGG for integration and interpretation of large-scale molecular data sets", *Nucleic Acids Res.,* vol. 40, no. Database issue, pp. D109-14.

Kankainen, M., Paulin, L., Tynkkynen, S., von Ossowski, I., Reunanen, J., Partanen, P., et al. 2009, "Comparative genomic analysis of *Lactobacillus rhamnosus* GG reveals pili containing a human-mucus binding protein", *Proc .Natl. Acad. Sci. U.S.A.,* vol. 106, no. 40, pp. 17193-17198.

Kant, R., Blom, J., Palva, A., Siezen, R. J. & de Vos, W. M. 2011, "Comparative genomics of *Lactobacillus*", *Microb. Biotechnol.,* vol. 4, no. 3, pp. 323-332.

Kant, R., Rintahaka, J., Yu, X., Sigvart-Mattila, P., Paulin, L., Mecklin, J. P., et al. 2014, "A comparative pan-genome perspective of niche-adaptable cell-surface protein phenotypes in *Lactobacillus rhamnosus*", *PLoS One,* vol. 9, no. 7, pp. e102762.

Kekkonen, R. A., Lummela, N., Karjalainen, H., Latvala, S., Tynkkynen, S., Järvenpää, S., et al. 2008, "Probiotic intervention has strain-specific anti-inflammatory effects in healthy adults", *World J. Gastroenterol.,* vol. 14, no. 13, pp. 2029-2036.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., et al. 2003, "Complete genome sequence of *Lactobacillus plantarum* WCFS1", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 100, no. 4, pp. 1990-1995.

Klimke, W., O'Donovan, C., White, O., Brister, J.R., Clark, K., Fedorov, B., et al. 2011, "Solving the Problem: Genome Annotation Standards before the Data Deluge", *Stand. Genomic Sci.,* vol. 5, no. 1, pp. 168-193.

Konstantinov, S. R., Smidt, H., de Vos, W. M., Bruijns, S. C., Singh, S. K., Valence, F., et al. 2008, "S layer protein A of *Lactobacillus acidophilus* NCFM regulates immature dendritic cell and T cell functions", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 105, no. 49, pp. 19474-19479.

Koonin, E. V. & Galperin, M. Y. 2003, "Principles and Methods of Sequence Analysis" in *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics* Kluwer Academic, Boston.

Korbel, J. O., Doerks, T., Jensen, L. J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S. D., et al. 2005, "Systematic association of genes to phenotypes by genome and literature mining", *PLoS Biol.,* vol. 3, no. 5, pp. e134.

Koskinen, J. P. & Holm, L. 2012, "SANS: high-throughput retrieval of protein sequences allowing 50% mismatches", *Bioinformatics,* vol. 28, no. 18, pp. i438-i443.

Kristensen, D. M., Wolf, Y. I., Mushegian, A. R. & Koonin, E. V. 2011, "Computational methods for Gene Orthology inference", *Brief Bioinform,* vol. 12, no. 5, pp. 379-391.

Krumsiek, J., Arnold, R. & Rattei, T. 2007, "Gepard: a rapid and sensitive tool for creating dotplots on genome scale", *Bioinformatics,* vol. 23, no. 8, pp. 1026-1028.

Kube, M., Chernikova, T. N., Al-Ramahi, Y., Beloqui, A., Lopez-Cortez, N., Guazzaroni, M. E., et al. 2013, "Genome sequence and functional genomic analysis of the oil-degrading bacterium *Oleispira antarctica*", *Nat. Commun.,* vol. 4, pp. 2156.

Kukkonen, K., Savilahti, E., Haahtela, T., Juntunen-Backman, K., Korpela, R., Poussa, T., et al. 2008, "Long-term safety and impact on infection rates of postnatal probiotic and prebiotic (synbiotic) treatment: randomized, double-blind, placebo-controlled trial", *Pediatrics,* vol. 122, no. 1, pp. 8-12.

Kukkonen, K., Savilahti, E., Haahtela, T., Juntunen-Backman, K., Korpela, R., Poussa, T., et al. 2007, "Probiotics and prebiotic galacto-oligosaccharides in the prevention of allergic diseases: a randomized, double-blind, placebo-controlled trial", *J. Allergy Clin. Immunol.,* vol. 119, no. 1, pp. 192-198.

Kuo, C. H., Moran, N. A. & Ochman, H. 2009, "The consequences of genetic drift for bacterial genome complexity", *Genome Res.,* vol. 19, no. 8, pp. 1450-1454.

Kuzniar, A., van Ham, R. C., Pongor, S. & Leunissen, J. A. 2008, "The quest for orthologs: finding the corresponding gene across genomes", *Trends Genet.,* vol. 24, no. 11, pp. 539-551.

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T. & Ussery, D. W. 2007, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes", *Nucleic Acids Res.,* vol. 35, no. 9, pp. 3100-3108.

Lamont, R. F., Sobel, J .D., Akins, R. A., Hassan, S. S., Chaiworapongsa, T., Kusanovic, J. P., et al. 2011, "The vaginal microbiome: new information about genital tract flora using molecular based techniques", *BJOG,* vol. 118, no. 5, pp. 533-549.

Land, M., Hauser, L., Jun, S.R., Nookaew, I., Leuze, M. R., Ahn, T. H., et al. 2015, "Insights from 20 years of bacterial genome sequencing", *Funct. Integr. Genomics,* vol. 15, no. 2, pp. 141-161.

Langille, M. G., Hsiao, W. W. & Brinkman, F. S. 2010, "Detecting genomic islands using bioinformatics approaches", *Nat. Rev. Microbiol.,* vol. 8, no. 5, pp. 373-382.

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., et al. 2015, "Patterns of genome evolution that have accompanied host adaptation in *Salmonella*", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 112, no. 3, pp. 863-868.

Larsen, T. S. & Krogh, A. 2003, "EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance", *BMC Bioinformatics,* vol. 4, pp. 21.

Laslett, D. & Canback, B. 2004, "ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences", *Nucleic Acids Res.,* vol. 32, no. 1, pp. 11-16.

Le Marechal, C., Peton, V., Ple, C., Vroland, C., Jardin, J., Briard-Bion, V., et al. 2015, "Surface proteins of *Propionibacterium freudenreichii* are involved in its anti-inflammatory properties", *J.Proteomics,* vol. 113, pp. 447-461.

Lee, D., Redfern, O. & Orengo, C. 2007, "Predicting protein function from sequence and structure", *Nat. Rev. Mol. Cell Biol.,* vol. 8, no. 12, pp. 995-1005.

Leroy, F. & De Vuyst, L. 2004, "Lactic acid bacteria as functional starter cultures for the food fermentation industry", *Trends Food Sci. Technol.,* vol. 15, no. 2, pp. 67-78.

Letunic, I. & Bork, P. 2007, "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation", *Bioinformatics,* vol. 23, no. 1, pp. 127-128.

Li, L., Stoeckert, C. J. Jr & Roos, D. S. 2003, "OrthoMCL: identification of ortholog groups for eukaryotic genomes", *Genome Res.,* vol. 13, no. 9, pp. 2178-2189.

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. 2008, "Prophinder: a computational tool for prophage prediction in prokaryotic genomes", *Bioinformatics,* vol. 24, no. 6, pp. 863-865.

Lin, Y. F., A, D. R., Guan, S., Mamanova, L. & McDowall, K. J. 2013, "A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in *Propionibacterium acnes*, a major contributor to wide-spread human disease", *BMC Genomics,* vol. 14, pp. 620-2164-14-620.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. 2012, "Comparison of next-generation sequencing systems", *BioMed Research International,* vol. 2012.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., et al. 2012, "High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity", *Nat. Rev. Microbiol.,* vol. 10, no. 9, pp. 599-606.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. 2014, "The carbohydrate-active enzymes database (CAZy) in 2013", *Nucleic Acids Res.,* vol. 42, no. Database issue, pp. D490-5.

Lomholt, H. B. & Kilian, M. 2010, "Population genetic analysis of *Propionibacterium acnes* identifies a subpopulation and epidemic clones associated with acne", *PLoS One,* vol. 5, no. 8, pp. e12277.

Loux, V., Mariadassou, M., Almeida, S., Chiapello, H., Hammani, A., Buratti, J., et al. 2015, "Mutations and genomic islands can explain the strain dependency of sugar utilization in 21 strains of *Propionibacterium freudenreichii*", *BMC Genomics,* vol. 16, no. 1, pp. 296-015-1467-7.

Lowe, T. M. & Eddy, S. R. 1997, "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence", *Nucleic Acids Res.,* vol. 25, no. 5, pp. 955-964.

Lukjancenko, O., Ussery, D. W. & Wassenaar, T. M. 2012, "Comparative genomics of *Bifidobacterium*, *Lactobacillus* and related probiotic genera", *Microb. Ecol.,* vol. 63, no. 3, pp. 651-673.

Ma, B., Forney, L. J. & Ravel, J. 2012, "Vaginal microbiome: rethinking health and disease", *Annu. Rev. Microbiol.,* vol. 66, pp. 371-389.

Macklaim, J. M., Gloor, G. B., Anukam, K. C., Cribby, S. & Reid, G. 2011, "At the crossroads of vaginal health and disease, the genome sequence of *Lactobacillus iners* AB-1", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 108 Suppl 1, pp. 4688-4695.

Macklaim, J., Fernandes, A., Di Bella, J., Hammond, J., Reid, G. & Gloor, G. 2013, "Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis", *Microbiome,* vol. 1, no. 1, pp. 12.

MacLean, D., Jones, J. D. & Studholme, D. J. 2009, "Application of 'next-generation' sequencing technologies to microbial genetics", *Nat. Rev. Microbiol.,* vol. 7, no. 4, pp. 287-296.

Mak, T. N., Schmid, M., Brzuszkiewicz, E., Zeng, G., Meyer, R., Sfanos, K. S., et al. 2013, "Comparative genomics reveals distinct host-interacting traits of three major human-associated propionibacteria", *BMC Genomics,* vol. 14, pp. 640-2164-14-640.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., et al. 2006, "Comparative genomics of the lactic acid bacteria", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 103, no. 42, pp. 15611-15616.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., et al. 2011, "Evolution and classification of the CRISPR-Cas systems", *Nat. Rev. Microbiol.,* vol. 9, no. 6, pp. 467-477.

Malik, S., Siezen, R.J., Renckens, B., Vaneechoutte, M., Vanderleyden, J. & Lebeer, S. 2014, "Draft Genome Sequence of *Lactobacillus plantarum* CMPG5300, a Human Vaginal Isolate", *Genome Announc,* vol. 2, no. 6, pp. 10.1128/genomeA.01149-14.

Mandlik, A., Livny, J., Robins, W. P., Ritchie, J. M., Mekalanos, J. J. & Waldor, M. K. 2011, "RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression", *Cell. Host Microbe,* vol. 10, no. 2, pp. 165-174.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. 2015, "CDD: NCBI's conserved domain database", *Nucleic Acids Res.,* vol. 43, no. Database issue, pp. D222-6.

Mardis, E. R. 2008, "Next-generation DNA sequencing methods", *Annu. Rev. Genomics Hum. Genet.,* vol. 9, pp. 387-402.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. 2005, "Genome sequencing in microfabricated high-density picolitre reactors", *Nature,* vol. 437, no. 7057, pp. 376-380.

Martin, D. M., Berriman, M. & Barton, G. J. 2004, "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes", *BMC Bioinformatics,* vol. 5, pp. 178.

Maurice, C. F., Haiser, H. J. & Turnbaugh, P. J. 2013, "Xenobiotics shape the physiology and gene expression of the active human gut microbiome", *Cell,* vol. 152, no. 1-2, pp. 39-50.

Maynard, C. L., Elson, C. O., Hatton, R. D. & Weaver, C. T. 2012, "Reciprocal interactions of the intestinal microbiota and immune system", *Nature,* vol. 489, no. 7415, pp. 231-241.

McFarland, L. V. 2000, "Normal flora: diversity and functions", *Microb. Ecol. Health Dis.,* vol. 12, no. 4, pp. 193-207.

McMillan, A., Macklaim, J. M., Burton, J. P. & Reid, G. 2013, "Adhesion of *Lactobacillus iners* AB-1 to Human fibronectin: a key mediator for persistence in the vagina?", *Reprod.Sci.,* vol. 20, no. 7, pp. 791-796.

Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. 2005, "The microbial pan-genome", *Curr. Opin. Genet. Dev.,* vol. 15, no. 6, pp. 589-594.

Medini, D., Serruto, D., Parkhill, J., Relman, D. A., Donati, C., Moxon, R., et al. 2008, "Microbiology in the post-genomic era", *Nat. Rev. Microbiol.,* vol. 6, no. 6, pp. 419-430.

Mendes-Soares, H., Suzuki, H., Hickey, R. J. & Forney, L. J. 2014, "Comparative functional genomics of *Lactobacillus* spp. reveals possible mechanisms for specialization of vaginal lactobacilli to their environment", *J. Bacteriol.,* vol. 196, no. 7, pp. 1458-1470.

Metzker, M. L. 2005, "Emerging technologies in DNA sequencing", *Genome Res.,* vol. 15, no. 12, pp. 1767-1776.

Meyer, F., Overbeek, R. & Rodriguez, A. 2009, "FIGfams: yet another set of protein families", *Nucleic Acids Res.,* vol. 37, no. 20, pp. 6643-6654.

Mikheyev, A. S. & Tin, M. M. 2014, "A first look at the Oxford Nanopore MinION sequencer", *Mol. Ecol. Resour.,* vol. 14, no. 6, pp. 1097-1102.

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. 2015, "The InterPro protein families database: the classification resource after 15 years", *Nucleic Acids Res.,* vol. 43, no. Database issue, pp. D213-21.

Morita, H., Toh, H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., et al. 2008, "Comparative genome analysis of *Lactobacillus reuteri* and *Lactobacillus fermentum* reveal a genomic island for reuterin and cobalamin production", *DNA Res.,* vol. 15, no. 3, pp. 151-161.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. 2008, "Mapping and quantifying mammalian transcriptomes by RNA-Seq", *Nat.Methods,* vol. 5, no. 7, pp. 621-628.

Nagarajan, N. & Pop, M. 2013, "Sequence assembly demystified", *Nat.Rev.Genet.,* vol. 14, no. 3, pp. 157-167.

Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., et al. 2010, "A catalog of reference genomes from the human microbiome", *Science,* vol. 328, no. 5981, pp. 994-999.

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Barnabas, B., Kong, H. H., et al. 2014, "Biogeography and individuality shape function in the human skin metagenome", *Nature,* vol. 514, no. 7520, pp. 59-64.

Oshlack, A., Robinson, M. D. & Young, M. D. 2010, "From RNA-seq reads to differential expression results", *Genome Biol.,* vol. 11, no. 12, pp. 220. Epub 2010 Dec 22.

Oshlack, A. & Wakefield, M. J. 2009, "Transcript length bias in RNA-seq data confounds systems biology", *Biol. Direct,* vol. 4, pp. 14-6150-4-14.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. 2005, "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes", *Nucleic Acids Res.,* vol. 33, no. 17, pp. 5691-5702.

Pallen, M. J. & Wren, B. W. 2007, "Bacterial pathogenomics", *Nature,* vol. 449, no. 7164, pp. 835-842.

Panagea, S., Corkill, J. E., Hershman, M. J. & Parry, C. M. 2005, "Breast abscess caused by *Propionibacterium avidum* following breast reduction surgery: case report and review of the literature", *J. Infect.,* vol. 51, no. 5, pp. e253-5.

Parizzi, L. P., Grassi, M. C., Llerena, L. A., Carazzolle, M. F., Queiroz, V. L., Lunardi, I., et al. 2012, "The genome sequence of *Propionibacterium acidipropionici* provides insights into its biotechnological and industrial potential", *BMC Genomics,* vol. 13, pp. 562-2164-13-562.

Patrick, S. & McDowell, A. 2012, "Genus I. *Propionibacterium*" in *Bergey's manual of systematic bacteriology. Vol. 5, the actinobacteria*, eds. M. Goodfellow, P. Kämpfer, H. Busse, et al., 2nd ED. edn, Springer New York Dordrecht Heidelberg London, pp. 1138-1154.

Pearson, W. R. & Lipman, D. J. 1988, "Improved tools for biological sequence comparison", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 85, no. 8, pp. 2444-2448.

Petrova, M. I., Lievens, E., Malik, S., Imholz, N. & Lebeer, S. 2015, "*Lactobacillus* species as biomarkers and agents that can promote various aspects of vaginal health", *Front. Physiol.,* vol. 6, pp. 81.

Piveteau, P. 1999, "Metabolism of lactate and sugars by dairy propionibacteria: a review", *Le Lait,* vol. 79, no. 1, pp. 23-41.

Poonam, Pophaly, S. D., Tomar, S. K., De, S. & Singh, R. 2012, "Multifaceted attributes of dairy propionibacteria: a review", *World J. Microbiol. Biotechnol.,* vol. 28, no. 11, pp. 3081-3095.

Pop, M. 2009, "Genome assembly reborn: recent computational challenges", *Brief Bioinform,* vol. 10, no. 4, pp. 354-366.

Pop, M., Phillippy, A., Delcher, A. L. & Salzberg, S. L. 2004, "Comparative genome assembly", *Brief Bioinform,* vol. 5, no. 3, pp. 237-248.

Poptsova, M. S. & Gogarten, J. P. 2010, "Using comparative genome analysis to identify problems in annotated microbial genomes", *Microbiology,* vol. 156, no. Pt 7, pp. 1909-1917.

Power, S. E., Harris, H. M., Bottacini, F., Ross, R. P., O'Toole, P. W. & Fitzgerald, G. F. 2013, "Draft Genome Sequence of *Lactobacillus crispatus* EM-LC1, an Isolate with Antimicrobial Activity Cultured from an Elderly Subject", *Genome Announc,* vol. 1, no. 6, pp. 10.1128/genomeA.01070-13.

Prajapati, J. B., Khedkar, C. D., Chitra, J., Suja, S., Mishra, V., Sreeja, V., et al. 2011, "Whole-genome shotgun sequencing of an Indian-origin *Lactobacillus helveticus* strain, MTCC 5463, with probiotic potential", *J.Bacteriol.,* vol. 193, no. 16, pp. 4282-4283.

Pridmore, R. D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A. C., et al. 2004, "The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 101, no. 8, pp. 2512-2517.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. 2010, "A human gut microbial gene catalogue established by metagenomic sequencing", *Nature,* vol. 464, no. 7285, pp. 59-65.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. 2013, "A large-scale evaluation of computational protein function prediction", *Nat. Methods,* vol. 10, no. 3, pp. 221-227.

Raes, J., Harrington, E. D., Singh, A. H. & Bork, P. 2007, "Protein function space: viewing the limits or limited by our view?", *Curr. Opin. Struct. Biol.,* vol. 17, no. 3, pp. 362-369.

Raftis, E. J., Salvetti, E., Torriani, S., Felis, G. E. & O'Toole, P. W. 2011, "Genomic diversity of *Lactobacillus salivarius*", *Appl. Environ. Microbiol.,* vol. 77, no. 3, pp. 954-965.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., et al. 2011, "Vaginal microbiome of reproductive-age women", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 108 Suppl 1, pp. 4680-4687.

Rawlings, N. D., Waller, M., Barrett, A. J. & Bateman, A. 2014, "MEROPS: the database of proteolytic enzymes, their substrates and inhibitors", *Nucleic Acids Res.,* vol. 42, no. Database issue, pp. D503-9.

Reed, J. L., Famili, I., Thiele, I. & Palsson, B.O. 2006, "Towards multidimensional genome annotation", *Nat. Rev. Genet.,* vol. 7, no. 2, pp. 130-141.

Reuter, J. A., Spacek, D. V. & Snyder, M.P. 2015, "High-throughput sequencing technologies", *Mol. Cell,* vol. 58, no. 4, pp. 586-597.

Reuter, S., Connor, T. R., Barquist, L., Walker, D., Feltwell, T., Harris, S. R., et al. 2014, "Parallel independent evolution of pathogenicity within the genus *Yersinia*", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 111, no. 18, pp. 6768-6773.

Robinson, M. D. & Oshlack, A. 2010, "A scaling normalization method for differential expression analysis of RNA-seq data", *Genome Biol.,* vol. 11, no. 3, pp. R25-2010-11-3-r25. Epub 2010 Mar 2.

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. 2003, "Genome-scale approaches to resolving incongruence in molecular phylogenies", *Nature,* vol. 425, no. 6960, pp. 798-804.

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., et al. 2014, "The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women", *Microbiome,* vol. 2, no. 1, pp. 4-2618-2-4.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. & Nyren, P. 1996, "Real-time DNA sequencing using detection of pyrophosphate release", *Anal. Biochem.,* vol. 242, no. 1, pp. 84-89.

Ronaghi, M., Uhlen, M. & Nyren, P. 1998, "A sequencing method based on real-time pyrophosphate", *Science,* vol. 281, no. 5375, pp. 363, 365.

Rost, B. 2002, "Enzyme function less conserved than anticipated", *J. Mol. Biol.,* vol. 318, no. 2, pp. 595-608.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. & Ofran, Y. 2003, "Automatic prediction of protein function", *Cell Mol. Life Sci.,* vol. 60, no. 12, pp. 2637-2650.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. 2011, "An integrated semiconductor device enabling non-optical genome sequencing", *Nature,* vol. 475, no. 7356, pp. 348-352.

Saier, M. H. Jr 2000, "A functional-phylogenetic classification system for transmembrane solute transporters", *Microbiol. Mol. Biol. Rev.,* vol. 64, no. 2, pp. 354-411.

Salvetti, E., Torriani, S. & Felis, G. 2012, "The Genus *Lactobacillus*: A Taxonomic Update", *Probiotics and Antimicrobial Proteins,* vol. 4, no. 4, pp. 217-226.

Samad, A., Huff, E. F., Cai, W. & Schwartz, D. C. 1995, "Optical mapping: a novel, single-molecule approach to genomic analysis", *Genome Res.,* vol. 5, no. 1, pp. 1-4.

Sanger, F., Nicklen, S. & Coulson, A. R. 1977, "DNA sequencing with chain-terminating inhibitors", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 74, no. 12, pp. 5463-5467.

Saraoui, T., Parayre, S., Guernec, G., Loux, V., Montfort, J., Le Cam, A., et al. 2013, "A unique in vivo experimental approach reveals metabolic adaptation of the probiotic *Propionibacterium freudenreichii* to the colon environment", *BMC Genomics,* vol. 14, pp. 911-2164-14-911.

Schadt, E. E., Turner, S. & Kasarskis, A. 2010, "A window into third-generation sequencing", *Hum. Mol. Genet.,* vol. 19, no. R2, pp. R227-40.

Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. 2009, "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies", *PLoS Comput. Biol.,* vol. 5, no. 12, pp. e1000605.

Schuster, S. C. 2008, "Next-generation sequencing transforms today's biology", *Nat. Methods,* vol. 5, no. 1, pp. 16-18.

Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., et al. 2000, "RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array", *Nat. Biotechnol.,* vol. 18, no. 12, pp. 1262-1268.

Senan, S., Prajapati, J. B. & Joshi, C. G. 2015, "Whole-genome based validation of the adaptive properties of Indian origin probiotic *Lactobacillus helveticus* MTCC 5463", *J. Sci. Food Agric.,* vol. 95, no. 2, pp. 321-328.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., et al. 2010, "The primary transcriptome of the major human pathogen *Helicobacter pylori*", *Nature,* vol. 464, no. 7286, pp. 250-255.

Sheikh, M. & Erlich, Y. 2012, "Base-Calling for Bioinformaticians" in *Bioinformatics for High Throughput Sequencing*, eds. N. Rodriguez-Ezpeleta, M. Hackenberg & A.M. Aransay, Springer New York, pp. 67-83.

Shendure, J. & Ji, H. 2008, "Next-generation DNA sequencing", *Nat. Biotechnol.,* vol. 26, no. 10, pp. 1135-1145.

Shendure, J. & Lieberman Aiden, E. 2012, "The expanding scope of DNA sequencing", *Nat. Biotechnol.,* vol. 30, no. 11, pp. 1084-1094.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., et al. 2005, "Accurate multiplex polony sequencing of an evolved bacterial genome", *Science,* vol. 309, no. 5741, pp. 1728-1732.

Shipitsyna, E., Roos, A., Datcu, R., Hallen, A., Fredlund, H., Jensen, J. S., et al. 2013, "Composition of the vaginal microbiota in women of reproductive age--sensitive and specific molecular diagnosis of bacterial vaginosis is possible?", *PLoS One,* vol. 8, no. 4, pp. e60670.

Shively, J. M., English, R. S., Baker, S. H. & Cannon, G. C. 2001, "Carbon cycling: the prokaryotic contribution", *Curr. Opin. Microbiol.,* vol. 4, no. 3, pp. 301-306.

Siezen, R. J., Tzeneva, V. A., Castioni, A., Wels, M., Phan, H. T., Rademaker, J. L., et al. 2010, "Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches", *Environ. Microbiol.,* vol. 12, no. 3, pp. 758-773.

Siezen, R. J. & van Hylckama Vlieg, J. E. 2011, "Genomic diversity and versatility of *Lactobacillus plantarum*, a natural metabolic engineer", *Microb. Cell. Fact.,* vol. 10 Suppl 1, pp. S3. Epub 2011 Aug 30.

Simon, C. & Daniel, R. 2011, "Metagenomic analyses: past and future trends", *Appl. Environ. Microbiol.,* vol. 77, no. 4, pp. 1153-1161.

Skvortsov, T. A. & Azhikina, T. L. 2010, "A review of the transcriptome analysis of bacterial pathogens in vivo: Problems and solutions", *Russian Journal of Bioorganic Chemistry,* vol. 36, no. 5, pp. 550-559.

Smid, E. J. & Hugenholtz, J. 2010, "Functional genomics for food fermentation processes", *Annu. Rev. Food Sci. Technol.,* vol. 1, pp. 497-519.

Smokvina, T., Wels, M., Polka, J., Chervaux, C., Brisse, S., Boekhorst, J., et al. 2013, "*Lactobacillus paracasei* comparative genomics: towards species pan-genome definition and exploitation of diversity", *PLoS One,* vol. 8, no. 7, pp. e68731.

Soneson, C. & Delorenzi, M. 2013, "A comparison of methods for differential expression analysis of RNA-seq data", *BMC Bioinformatics,* vol. 14, pp. 91-2105-14-91.

Sonnhammer, E. L. & Koonin, E. V. 2002, "Orthology, paralogy and proposed classification for paralog subtypes", *Trends Genet.,* vol. 18, no. 12, pp. 619-620.

Sorek, R. & Cossart, P. 2010, "Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity", *Nat. Rev. Genet.,* vol. 11, no. 1, pp. 9-16.

Spinler, J. K., Sontakke, A., Hollister, E. B., Venable, S. F., Oh, P. L., Balderas, M. A., et al. 2014, "From prediction to function using evolutionary genomics: human-specific ecotypes of *Lactobacillus reuteri* have diverse probiotic functions", *Genome Biol. Evol.,* vol. 6, no. 7, pp. 1772-1789.

Spivey, M. A., Dunn-Horrocks, S. L. & Duong, T. 2014, "Epithelial cell adhesion and gastrointestinal colonization of *Lactobacillus* in poultry", *Poult. Sci.,* vol. 93, no. 11, pp. 2910-2919.

Staden, R. 1979, "A strategy of DNA sequencing employing computer programs", *Nucleic Acids Res.,* vol. 6, no. 7, pp. 2601-2610.

Stein, L. 2001, "Genome annotation: from sequence to biology", *Nat. Rev. Genet.,* vol. 2, no. 7, pp. 493-503.

Suomalainen, T., Sigvart-Mattila, P., Mättö, J. & Tynkkynen, S. 2008, "In vitro and in vivo gastrointestinal survival, antibiotic susceptibility and genetic identification of *Propionibacterium freudenreichii* ssp. *shermanii* JS", *Int. Dairy J.,* vol. 18, no. 3, pp. 271-278.

Suomalainen, T. H. & Mäyrä-Mäkinen, A. M. 1999, "Propionic acid bacteria as protective cultures in fermented milks and breads", *Le Lait,* vol. 79, no. 1, pp. 165-174.

Suoniemi, A. & Tynkkynen, S. 2002, "Cloning and characterization of an esterase from *Propionibacterium freudenreichii* ssp. *shermanii*", *Le Lait,* vol. 82, no. 1, pp. 81-89.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. 2003, "The COG database: an updated version includes eukaryotes", *BMC Bioinformatics,* vol. 4, pp. 41.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. 2005, "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 102, no. 39, pp. 13950-13955.

Tettelin, H., Riley, D., Cattuto, C. & Medini, D. 2008, "Comparative genomics: the bacterial pan-genome", *Curr. Opin. Microbiol.,* vol. 11, no. 5, pp. 472-477.

Thierry, A., Maillard, M., Richoux, R., Kerjean, J. & Lortal, S. 2005, "*Propionibacterium freudenreichii* strains quantitatively affect production of volatile compounds in Swiss cheese", *Le Lait,* vol. 85, no. 1-2, pp. 57-74.

Thierry, A., Deutsch, S. M., Falentin, H., Dalmasso, M., Cousin, F. J. & Jan, G. 2011, "New insights into physiology and metabolism of *Propionibacterium freudenreichii*", *Int. J. Food Microbiol.,* vol. 149, no. 1, pp. 19-27.

Tian, W. & Skolnick, J. 2003, "How well is enzyme function conserved as a function of pairwise sequence identity?", *J. Mol. Biol.,* vol. 333, no. 4, pp. 863-882.

Tomida, S., Nguyen, L., Chiu, B. H., Liu, J., Sodergren, E., Weinstock, G. M., et al. 2013, "Pan-genome and comparative genome analyses of *Propionibacterium acnes* reveal its genomic diversity in the healthy and diseased human skin microbiome", *MBio,* vol. 4, no. 3, pp. e00003-13.

Trapnell, C. & Salzberg, S. L. 2009, "How to map billions of short reads onto genomes", *Nat. Biotechnol.,* vol. 27, no. 5, pp. 455-457.

van de Guchte, M., Penaud, S., Grimaldi, C., Barbe, V., Bryson, K., Nicolas, P., et al. 2006, "The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 103, no. 24, pp. 9274-9279.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. 2014, "Ten years of next-generation sequencing technology", *Trends Genet.,* vol. 30, no. 9, pp. 418-426.

van Heel, A. J., de Jong, A., Montalban-Lopez, M., Kok, J. & Kuipers, O. P. 2013, "BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides", *Nucleic Acids Res.,* vol. 41, no. Web Server issue, pp. W448-53.

van Vliet, A. H. 2010, "Next generation sequencing of microbial transcriptomes: challenges and opportunities", *FEMS Microbiol.Lett.,* vol. 302, no. 1, pp. 1-7.

Vasquez, A., Jakobsson, T., Ahrne, S., Forsum, U. & Molin, G. 2002, "Vaginal lactobacillus flora of healthy Swedish women", *J. Clin. Microbiol.,* vol. 40, no. 8, pp. 2746-2749.

Ventura, M., O'Flaherty, S., Claesson, M. J., Turroni, F., Klaenhammer, T. R., van Sinderen, D., et al. 2009, "Genome-scale analyses of health-promoting bacteria: probiogenomics", *Nat. Rev. Microbiol.,* vol. 7, no. 1, pp. 61-71.

Verstraelen, H., Verhelst, R., Claeys, G., De Backer, E., Temmerman, M. & Vaneechoutte, M. 2009, "Longitudinal analysis of the vaginal microflora in pregnancy suggests that *L. crispatus* promotes the stability of the normal vaginal microflora and that *L. gasseri* and/or *L. iners* are more conducive to the occurrence of abnormal vaginal microflora", *BMC Microbiol.,* vol. 9, pp. 116-2180-9-116.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., et al. 2006, "Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models", *BMC Bioinformatics,* vol. 7, pp. 142.

Walter, J. 2008, "Ecological role of lactobacilli in the gastrointestinal tract: implications for fundamental and biomedical research", *Appl. Environ. Microbiol.,* vol. 74, no. 16, pp. 4985-4996.

Wang, Z., Gerstein, M. & Snyder, M. 2009, "RNA-Seq: a revolutionary tool for transcriptomics", *Nat. Rev. Genet.,* vol. 10, no. 1, pp. 57-63.

Wang, Z., Jin, Y. & Yang, S. T. 2014, "High cell density propionic acid fermentation with an acid tolerant strain of *Propionibacterium acidipropionici*", *Biotechnol.Bioeng.,* vol. 112, pp. 502–511.

Warren, A. S., Archuleta, J., Feng, W. C. & Setubal, J. C. 2010, "Missing genes in the annotation of prokaryotic genomes", *BMC Bioinformatics,* vol. 11, pp. 131.

Whitman, W. B., Coleman, D. C. & Wiebe, W. J. 1998, "Prokaryotes: the unseen majority", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 95, no. 12, pp. 6578-6583.

Wilson, M. 2008, *Bacteriology of humans: an ecological perspective,* Blackwell Publishing Ltd.

Witkin, S. S., Linhares, I. M. & Giraldo, P. 2007, "Bacterial flora of the female genital tract: function and immune regulation", *Best Pract. Res. Clin. Obstet. Gynaecol.,* vol. 21, no. 3, pp. 347-354.

Yoder-Himes, D. R., Chain, P. S., Zhu, Y., Wurtzel, O., Rubin, E. M., Tiedje, J. M., et al. 2009, "Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing", *Proc. Natl. Acad. Sci. U. S. A.,* vol. 106, no. 10, pp. 3976-3981.

Yuki, N., Shimazaki, T., Kushiro, A., Watanabe, K., Uchida, K., Yuyama, T., et al. 2000, "Colonization of the stratified squamous epithelium of the nonsecreting area of horse stomach by lactobacilli", *Appl. Environ. Microbiol.,* vol. 66, no. 11, pp. 5030-5034.

Zhou, C.E., Smith, J., Lam, M., Zemla, A., Dyer, M. D. & Slezak, T. 2007, "MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications", *Nucleic Acids Res.,* vol. 35, no. Database issue, pp. D391-4.

Zhou, F. & Xu, Y. 2010, "cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data", *Bioinformatics,* vol. 26, no. 16, pp. 2051-2052.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. 2011, "PHAST: a fast phage search tool", *Nucleic Acids Res.,* vol. 39, no. Web Server issue, pp. W347-52.