Analysis of pan-genome content and its application in microbial identification

Lukjancenko, Oksana; Ussery, David; Larsen, Mette Voldby

Publication date:
2014

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Analysis of pan-genome content and its application in microbial identification

PhD Thesis

Oksana Lukjancenko

October, 2013
Contents

1 Introduction 1

1 Comparative genomics 3
   1.1 The pan-genome 4
   1.2 Sequence homology search methods 4
      1.2.1 Phylogenetic-tree based homology search 5
      1.2.2 Pairwise homology search 5
      1.2.3 Profile-based homology search 8
   1.3 InterPro 9

2 Microbial identification and characterization 11
   2.1 Methods for microbial identification 11
   2.2 Epidemiological insight into microbial characterization 12
## CONTENTS

### II Projects

#### 3 BLAST-based comparative genomics

| 3.1 Paper I. Comparative Genomics of *Bifidobacterium, Lactobacillus* and Related Probiotic Genera | 19 |
| 3.2 Paper II. Genome sequencing identifies two nearly unchanged strains of persistent *Listeria monocytogenes* isolated at two different fish processing plants sampled 6 years apart | 43 |

#### 4 HMM-based comparative genomics

| 4.1 Paper III. (Manuscript). PanFunPro: Pan-genome analysis based on Functional Profiles | 55 |
| 4.2 Paper IV. (Manuscript). Life’s Set of Core Genes, Revisited | 67 |
| 4.3 Paper V. (Manuscript). Chromosome-specific families in *Vibrio* genomes | 83 |

#### 5 Microbial Identification Using Whole Genome Sequences

| 5.1 Paper VI. Design of an Enterobacteriaceae Pan-genome Microarray Chip | 99 |
| 5.2 Paper VII. Genomic variation in *Salmonella enterica* core genes for epidemiological typing | 115 |
| 5.3 TaxonomyFinder web-server | 127 |
| 5.4 Paper VIII. (Manuscript). Benchmarking of Methods for Genomic Taxonomy | 131 |

### 6 Conclusions and Future prospects

### Bibliography

### A Supplementary Material
Preface

This thesis was prepared at the Center for Biological Sequence Analysis, Department of Systems Biology, the Technical University of Denmark, in fulfillment of the requirements for acquiring a Ph.D. degree. This PhD was funded by DTU and Center for Genomic Epidemiology.

The work was carried out at the Center for Biological Sequence Analysis and Center for Genomic Epidemiology under the supervision of David W. Ussery and Mette Voldby Larsen. This thesis describes methods for pan-genome analysis and application of pan-genome content for taxonomy prediction. The thesis consists of an introduction and a collection of VIII research papers written during the period October, 2010 - September, 2013.

Oksana Lukjancenko
Lyngby, October 2013
Abstract

With the rapid development of DNA sequencing technology it is today possible to sequence multiple genomes in a single day at a low cost with a single machine. This has resulted in several large-scale genomic projects, such as Ten Thousand Microbial Genomes (BGI) to explore microbial diversity in China, and understand its influence to the environment and humans; The Human Microbiome project (NIH) to find microorganisms in association with healthy and infected humans; and The 100K Genome Project (University of California, Davis, and FDA), which aims to sequence the genomes of 100,000 infectious microorganisms and eventually speed up the diagnosis of foodborne illnesses. This genomic data can give biologists many possibilities to improve knowledge of organismal evolution and complex genetic systems.

The general interest of this PhD thesis is how to obtain relevant information from growing amounts of genomic data and use this to answer important biological questions. More specifically, comparison of prokaryotic proteomes is used to determine possible sets of functions, essential to sustain microbial life; to extract and interpret similarities and variance in genomic content within different taxonomic groups or genomic structures; and to use the information of a specific proteome to predict which species it might belong to. Two different algorithms, BLAST and profile Hidden Markov Models (HMMs), are used to determine similarity between sequences and to address the questions in this thesis.

The first project, described in Chapter 3, is based on using protein Basic Local Alignment Search Tool (BLAST) comparisons for sequence-based homology searches. Paper I presents comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera.; and Paper II illustrates the use of in silico analyses for the characterization of two Listeria monocytogenes strains.

Chapter 4 describes the use of profile HMMs for comparative analysis using for sequence-based homology searches. Paper III introduces PanFunPro a new, profile HMM-based method for pan-genome analysis. Paper IV illustrates the application of PanFunPro to a set of more than 2000 genomes; this paper aims to define set of protein families, which are conserved among all the genomes. Papers V demonstrates comparative genomics analysis of proteomes,
belonging to *Vibrio* genus.

In the last project, described in Chapter 5, both BLAST- and profile HMM-based methods are employed to infer taxonomy group-specific gene families, which are used for microbial identification. Paper VI illustrates the use of specific genes for microarray chip design; Paper VII demonstrates the use of the *Salmonella enterica* core-genome content for epidemiological typing; and Paper VIII represents the application of PanFunPro approach for *in silico* taxonomy prediction.

In summary, this thesis presents three projects that have contributed to identification and characterization of microbial organisms, and open new possibilities for comparative genomics and epidemiology.
Dansk resume

Grundet en rivende udvikling indenfor DNA sekventerings teknologi er det i dag muligt at sekventere flere genomer på en enkelt dag ved brug af en enkelt maskine til en lav pris. Dette har først til igangsættelse af flere store genomprojekter som ”The Ten Thousand Microbial Genomes” (BGI), der har til formål at udforske mikrobiediversitet i Kina og forstå dens indflydelse på miljøet og på mennesker; ”The Human Microbiome Project” (NIH) der undersøger samspillet mellem mikroorganismer i syge og raske mennesker, og ”The 100K Genome Project” (University of California, Davis and FDA) der har til mål at sekventere genomerne fra 100.000 infektionse mikroorganismer og med tiden forkorte diagnostikstiden på fædevarebørne sygdomme. Akkumuleringen af genom data giver biologer mulighed for at ge deres viden om evolutionen af mikroorganismer og giver indblik i komplekse genetiske systemer.

Det generelle fokus for denne PhD afhandling er hvordan man kan opnå relevant information fra denne voksende mængde af data og hvordan denne information kan bruges til at besvare vigtige biologiske spørgsmål. Mere specifikt sammenlignes prokaryote proteomer for at estimere mulige protein funktioner, der er essentielle for at opretholde mikrobielt liv; for at identificere ligheder og forskelle i genetisk indhold mellem forskellige taksonomiske grupper; og for at bruge denne information til forudsigelse af den taksonomiske placering af ukendte arter. To forskellige algoritmer, BLAST og profil HMMs, bruges til at bestemme similaritet mellem sekvenser og til at adressere PhD studiets centrale spørgsmål.

Det første projekt i denne afhandling, beskrevet i kapitel 3, giver eksempler på komparativ genom analyse ved brug af BLAST til bestemmelse af sekvens homologi. Artikel I præsenterer en komparativ analyse af genomer fra Bifidobacterium, Lactobacillus og beslægtede probiotiske genera; og Artikel II illustrerer brugen af in silico analyse til karakterisering af to Listeria monocytogenes stammer.

Kapitel 4 giver eksempler på komparativ analyse ved brug af profil HMMs til bestemmelse af sekvens homologi. I Artikel III introduceres PanFunPro, en ny metode til analyse af pan-genomer baseret på profil HMM modeller. Artikel IV illustrerer hvordan PanFunPro kan bruges til at analysere flere end 2.000 vi
genomer, med henblik på identificering af proteiner der er konserverede i alle genomerne. Artikel V demonstrerer komparativ analyse af genomer fra Vibrio genus.

I det sidste projekt, beskrevet i kapitel 5, bruges både BLAST- og profil HMM baserede metoder til at udlede gen familier der er specifikke for bestemte taksonomiske grupper, og som kan bruges til mikrobiel identifikation. Artikel VI illustrerer brugen af specifikke gener i microarray design; Artikel VII demonstrerer brugen af kerne genomet fra Salmonella enterica til epidemiologisk klassificering; og Artikel VIII præsenterer brugen af PanFunPro i in silico taksonomi forudsigelse.

Som opsummering, denne afhandling præsenterer tre projekter, der har bidraget til identifikation og karakterisering af mikrobielle organismer, og som har bnet op for nye muligheder indenfor komparativ genom analyse og epidemiologi.
Papers included in the thesis


III Oksana Lukjancenko, Martin Christinsen Frølund Thomsen, Mette Voldby Larsen, David Wayne Ussery. PanFunPro: PAN-genome analysis based on FUNctional PROfiles. Manuscript ready for submission.

IV Oksana Lukjancenko and David Wayne Ussery. Life’s set of core genes, Revisited. Manuscript ready for re-submission.

V Oksana Lukjancenko, and David Wayne Ussery. Chromosome-specific families in Vibrio genomes. [Submitted]


VIII Mette Voldby Larsen, Salvatore Cosentino, Oksana Lukjancenko, Dhany Saputra, Simon Rasmussen, Henrik Hasman, Thomas Sicheritz Ponten,
Papers not included in the thesis


- Tammi Vesth, Asl Ozen, Sandra Andersen, Rolf Sommer Kaas, **Oksana Lukjancenko**, Jon Bohlin, Intawat Nookaew, Trudy Wassenaar and David W. Ussery. *Veillonella, Firmicutes disguised as Gram negatives*. [Submitted]

- Trudy Wassenaar and **Oksana Lukjancenko**. *Chapter 5. Comparative genomics of Lactobacillus and other LAB.* "Lactic Acid Bacteria – Biodiversity And Taxonomy", JOHN WILEY & SONS, LTD. [Accepted]
Acknowledgements

It has been a great pleasure to work as PhD student at Center for Biological Sequence analysis and Center for Genomic Epidemiology, surrounded by people who were always there ready to help me when I needed it, to motivate, advice, or just give a hug and chat. I thank to all the CBS and CGE people, and especially headers of these two centers, Søren Brunak, Ole Lund and Frank Aarestrup.

My special thanks to my supervisors David Ussery and Mette Voldby Larsen, for your support, encouragement, and believing in me. Thank you for motivating me and having alternative ideas, even though some of them sounded impossible to implement. Thank you for being patient with me and listening to my sometimes crazy stories. I am grateful to David Ussery’s scientific networks, which gave me an opportunity to visit multiple centers during my external stay, and to participate at the comparative genomics workshops in exotic places.

Many thanks to Ole Lund, Martin Thompsen, Simon Rasmussen, Thomas Ponten, Salvatore Cosentino, Marlene Hansen, Shinny, Rolf Kaas, and other people from Center for Genomic Epidemiology. It have been a pleasure to work in a relaxed and friendly environment, discuss microbial identification and epidemiological characterization aspects, and participate in Global Microbial Identifier meetings.

I am grateful to Tammi, Asli, Steve, Rachita, Kristine, and Helen. It was a pleasure to share the office and work in Comparative Microbial Genomics group with some of you; discuss programming and other challenges while drinking the morning coffee; or just chatting. Thank you for listening, supporting and helping when I needed it.

Thanks to CBS office administration: Lone, Dorthe, Marlene, Karina, Martin, for your help. Also I would like to acknowledge CBS system administrators. Special thanks to John Damm Sørensen for being helpful with the issues related to the queueing system, and responding fast to the other multiple questions.
I would also like to mention people that I met during my external stay in USA and Canada. I am grateful to Tatiana Tatusova, for giving me an opportunity to work in National Center for Biotechnology Information (NCBI); Lance Price and Paul Keim for Tgen visit in Flagstaff; and Fiona Brinkman for hosting me in Simon Fraser University, Vancouver and introducing me to the details of IslandViewer. Additionally, I say thank you to representatives of "NCBI Russian mafia" and Fiona Brinkman’s Lab PhD students for all the moments we share, discussing scientific problems, hiking, and tasting all possible combinations of vegetarian lunch.

Thanks to all my friends in Denmark, especially Madara, Mireia, Tomas, Kasia, Ignas, Josef, Andreas, Dennis, Kim, Mathias, and Louise. I am glad I met you and am grateful for all the moments we spent together. Special gratitude to Bent, who has been very good and supportive friend; and for multiple walks and chats we had.

I am grateful to my friends in Lithuania, who are always waiting for me to come back.

My very special gratitude to my mom, dad and sister for always been supportive, believing in me and motivating me to reach my goals. This PhD would not have been possible without you. And finally thanks to Anders Ravn for being a special person for me.

During the three years in CBS I have met many nice people. Thanks to Edita, Ali, Dhanny, Ida, Leon, Emil, Piotr, Agata, Jens, Anne, Andrea, Mette, Jacob for nice chats in the CBS kitchen, cake club and for afterwork activities.

My very special gratitude to my family, mom, dad and sister for always been supportive, believing in me and motivating me to reach my goals. This PhD would not have been possible without you. And finally thanks to Anders Ravn for being a special person for me.
Part I

Introduction
Chapter 1

Comparative genomics

Sequencing of the complete genome of *Haemophilus influenzae* in 1995 pioneered a new era in genome sciences. Eight years later the number of complete sequences had increased to a hundred genomes. This number had doubled to about two-hundred genomes by the year 2005, and has further increased about twenty-fold by last year. Today, thousands of genomes are being sequenced worldwide. Some research groups are sequencing multiple strains from the same species to explore environmental adaptation and to determine the pan-genome of closely related organisms; others use bacterial sequencing information from diverse taxonomic groups to examine microbial variety. While sequencing becomes faster and cheaper, this rate of genomic data generation poses significant challenges for comparative genomics, such as speed and complexity of analysis, data quality assessment, along with result visualization and interpretation. Multiple approaches have been invented to face and overcome these challenges.

This chapter will briefly introduce the concept of a pan-genome, and several methods that are used in comparative microbial genomics. These methods are applied in search similarities and differences between multiple sets of prokaryotic genomes later in this thesis, as well as some challenges of prokaryotic pan-genome analysis.
1.1 The pan-genome

The concept of a bacterial species pan-genome was first introduced by Tettelin et al. in 2005 and defined as a repertoire of genetic sequences found in a given bacterial species [1]. Later, it was re-defined and stated that pan-genome consists of core genes, shared among all genomes in given taxonomic group; and a pool of dispensable genes, which can be present in several strains or specific to single organism (ORFans) [2].

The focus of pan-genome analysis is to compare the variance in proteomes between strains. The pan-genome size and content reflects the ability of a species to gain or loose genes. Multiple proteins can have significant similarity, thus the concept of protein equivalence - homology, should be considered. Homologs are categorized into two types: orthologs, genes diverged through speciation event from common ancestor, and paralogs, genes diverged through duplication event [3, 4]. Identification of homologous sequences sets is present in almost every comparative genomics study and is fundamental in understanding microbial diversity and evolutionary processes [5]. Furthermore, they are used to establish core and accessory genomes, assign functional annotation to the proteins of novel genome using previous knowledge of well-studied ones, and predict the size of the protein families. The core genome gives insight into functional potential, relations between organisms, genes necessary for distinct environmental niches, and pathogenicity; as a consequence core genes can be used as therapeutic and environmental markers for additional characterization and in determining the likely source of diseases, or in synthetic biology [6, 7, 8].

1.2 Sequence homology search methods

In general, determining homologs is a challenging problem. Many various approaches and associated databases were invented to determine orthology and paralogy. Homology search algorithms can be generally classified into: tree-based, pairwise similarity search, and profile-based [9, 10].
1.2.1 Phylogenetic-tree based homology search

Phylogenetic tree-based approaches rely on the evolutionary relationships between homologous genes in one or multiple organisms. Tree construction usually starts with a multiple sequence alignment and is further implemented by either distance-based, Unweighted Pair Group Method with Arithmetic Mean (UPGMA) \[1\] and neighbour-joining \[2\]; or character-based, Maximum parsimony \[3\], Maximum likelihood \[4\], and Bayesian statistics \[5\], algorithms. The main advantage of tree-based methods is their sensitivity. They are able to model the evolution of the whole group of genes at once, using the content information from the multiple alignment. However, trees are generally computationally expensive when the dataset is too large. Tree construction performance depends on the accuracy of multiple sequence alignments, which cannot be assured when a larger number of sequences is introduced, or when dealing with multi-domain proteins. Also, they are sensitive to the number of gaps in the alignment, which can lead to the reduced number of information, from which the model of evolution will be created in the tree. Therefore, automated phylogenetic trees construction is commonly a challenge in comparative genomics \[5\].

1.2.2 Pairwise homology search

Pairwise similarity search is an alternative way to assess sequence homology. The backbone of pairwise search is to compare query sequence to the sequence in the database, and to obtain the score that indicates the likelihood of matching to occur. This procedure is repeated for each sequence in the database and the best-match relationships are recorded. Pairwise comparison can be implemented using optimal alignment or heuristic alignment algorithms \[10\].

Optimal sequence alignment algorithms

Optimal alignment algorithms, such as Needleman-Wunsch \[16\] and Smith-Waterman \[17\], use dynamic programming for sequence alignment. The Needleman-Wunsch algorithm performs global sequence alignment, which assumes that two sequences are similar over the entire sequence length. This method is relevant
to the sequences that are roughly the same size and are expected to be similar over the entire length. However, the Needleman-Wunsch algorithm is computationally very demanding concerning time and space, and might be used to only relatively short sequence comparison \[18\].

Many proteins have functions described by shorter segments (protein domains), and hence sequence similarity can be defined by presence of these protein domains in homologous sequences. The Smith-Waterman algorithm, performs local sequence alignments, which searches for conserved regions instead of aligning the sequences entirely. The algorithm compares different length fragments and optimizes the matching score with respect to the scoring scheme being used. A local alignment is faster to calculate than a global alignment, but might report misleading homologs, when the query protein is multi-domain and shares only one domain with the compared sequence in the database. On the other hand, a global similarity algorithm may exclude possible homologs due to low similarity, if functional domains are short \[10\].

**Heuristic sequence alignment algorithms**

Heuristic alignments, such as FASTA \[19\] and BLAST \[20\], are approximations to Smith-Waterman algorithm. Heuristic approaches are much faster, easier to automate and can handle large amounts of data, but the increase in speed is usually comes with a prize of lower sensitivity and accuracy of prediction \[5\].

**BLAST**

BLAST was invented by Altschul et al. in 1990 \[20\] and since then has found many applications in different studies worldwide. It is the most widely used algorithm for sequence similarity search and functional characterization. BLAST algorithms are available in several versions, depending on the analysis type: BLASTn compares a nucleotide query sequence to the nucleotide database, BLASTp compares a protein query sequence to the protein database, BLASTx compares translated nucleotide query sequence to the protein database, tBLASTn compares protein query sequence to the translated nucleotide database, and tBLASTx compares translated nucleotide query sequence to a translated nu-
BLAST uses scoring method to evaluate the quality of pairwise sequence alignment, meaning that each position of the alignment is represented by a score, which is positive for a good match and is negative for a mismatch or gapped position. Scores for each pair can be obtained from the scoring matrix. DNA-DNA comparisons use straightforward scoring matrices, which gets a high score for base match and zero for base mismatch. In case of protein-protein comparison, more sophisticated scoring approach is used. There are 20 possible amino acids, which are grouped by properties, such as polarity, charge, and hydrophobicity; and overall 210 possible substitution pairs are available. Substitution matrix gives a measure of probability of a given amino acid to be substituted by another with respect to amino acid properties. Several substitution matrices were created to address this question. The first one, the measure of Percentage of Acceptable point Mutations (PAM matrix), was determined by Dayhoff in 1978 and reflects the measure of probability of one amino acid to be substituted by another in a given evolutionary distance. Higher score represents greater length of evolutionary time in PAM matrix. PAM-30 and PAM-70 are the most commonly used PAM matrices. Another type of substitution matrices is called BLOck SUbstitution Matrix (BLOSUM). BLOSUM matrix was suggested by Henikoff and Henikoff in 1992, and was derived by multiple local alignments of evolutionary divergent sequences. The blocks are built from conserved regions in the sequence (obtained with a similarity score over given threshold). BLOSUM-80 and BLOSUM-62 are the most used matrices. For instance, BLOSUM-62 was constructed from clusters of aligned proteins with identity score greater than 62. Generally PAM matrices with larger numbers would be more suitable for larger evolutionary distance, while in BLOSUM, matrices with higher scores would represent higher sequence similarity. The overall used measure of similarity between two sequences is called Expected value ($E$-value), which represents the probability of randomly occurring alignment.

BLAST is an heuristic approach, which provides rapid comparison of the query sequence to the database of known sequences and allows to retrieve available functional information. However BLAST doesn’t guarantee optimal alignment and looses sensitivity with the increase of speed.
1.2.3 Profile-based homology search

The assumption that functionally important regions are conserved over evolution, and that they can be detected in multiple sequences of different organisms, despite the overall low sequence similarity scores, led to the sequence-profile idea \[25\]. A variety of approaches, such as SHARP \[26\], MUSTER \[27\], HHpred \[28\], Metadomain \[29\] and Meta-MEME \[30\], were developed for adequate template-based sequence homology identification and structure prediction. Most of them include PSI-BLAST \[21\] and HMMER \[31\] algorithms. Profile-based searches are considerably more sensitive and accurate than simple pairwise search, however these methods can be computationally more demanding and slower.

**PSI-BLAST**

PSI-BLAST is a variation of BLAST algorithm, which looks for profiles - sets of evolutionary conserved sequence elements. It acquires a position-specific score matrix (PSSM) from multiple sequence alignment of high scoring sequences (above specified score) using BLASTp, and later, this PSSM is used to query database for new matches. The newly detected highly scoring sequences are used to update the profile \[32\].

**HMMER**

Hidden Markov models (HMMs) use stochastic processes that describe a probability distribution over potentially infinite number of possible sequences \[33\]. Profile HMMs can model divergent as well as conserved regions within multiple alignments, considering gaps, insertions and deletions. HMMs are applied to the problems of statistical modeling, database searching and multiple sequence alignment of protein families and protein domains.

HMMER is a widely used tool that uses profile-HMMs. It includes a set of programs sequence database and profile HMM search: *hmmscan* uses sequence as a query and searches it against the profile HMM database; *hmmsearch* takes profile HMM as query and uses it to search sequence database;
phmmer analogously to BLASTp takes a single sequence as query and used it to search sequence database; and jackhmmer analogously to PSI-BLAST takes sequence as a query and searches it against the sequence database. Profile HMM should be built from multiple sequence alignment or formatted using hmmbuild by HMMER software. Several different multiple alignment formats, such as CLUSTAL, SELEX, STOCKHOLM and aligned FASTA, are allowed. HMMER outputs two types of scores, bit-score and E-value; where bit-score is log-odds ratio score correlating the likelihood of profile HMM with the likelihood of the occurrence of independent, identically distributed random sequence model; and E-value is the number of randomly occurring hits, expected to reach equal or greater value of the bit-score.

1.3 InterPro

Advances in the sequencing technologies over the past fifteen years have resulted in rapidly growing genome datasets and the need to analyze them. A plethora of various analyses resulted in large number of databases, each with its own type of biological focus, signature prediction or search algorithms, and quality score schemes. In the year 2000, InterPro - an integrated tool for functional and structural classification, was introduced. InterPro provides a single resource of 13 protein signature collections, such as TIGRFAM, PIRSF, ProDom, PANTHER, SMART, PROSITE, HAMAP, Pfam, PRINTS, SUPERFAMILY, and Gene3D; and combines the signature recognition tools and quality check schemes from each of them into a single format output. Representative databases are integrated manually, and in principle a manual quality check of all signatures should lower the amount of false positives. Furthermore, InterPro provides mapping to Gene Ontology terms and relates InterPro entries to pathway and enzyme information containing resources, such as KEGG, PRIAM, Reactome, and UniPathway. GO mapping InterPro signature matches are determined using InterProScan software. InterProScan is implemented in Java programming language and includes a rapid pre-calculated match lookup service.
Pfam

Pfam is a large, widely used collection of domains, motifs, repeats, and protein families. Pfam contains two types of components: PfamA, high quality and manually curated entries; and PfamB, automatically generated models using Automatic Domain Decomposition Algorithm (ADDA) database \[51\]. Pfam-A profile HMMs are acquired from high quality multiple alignments; further, profile HMMs are searched against the UniProtKB sequence database; and family-specific sequence and domains gathering thresholds (GAs) are chosen \[40\]. The database counts 14,831 families in PfamA, and 544,866 families in PfamB, latest release (version 27.0).

TIGRFAM

TIGRFAM is a collection of full-length proteins and shorter regions at the level of superfamilies, subfamilies and equivalogs, where equivalogs are sets of homologous proteins conserved with respect to function. It is manually curated and described by Hidden Markov Models \[52\]. The TIGRFAM database counts 4,284 families in the latest release (version 13.0).

SUPERFAMILY

SUPERFAMILY is a collection of structural domains, described by HMMs. SUPERFAMILY employs Structural Classification of Proteins (SCOP) domains definitions at the superfamily level to determine structural annotations. It one by one models each sequence in the family, and later combines the result. The latest SCOP release (version 1.75) counts 3902 families and 1962 superfamilies \[42\].
Chapter 2

Microbial identification and characterization

Epidemic infectious diseases are one of the most serious mortality and morbidity causes worldwide. They are also responsible for significant economic loss around the world. Every year millions of people are infected by bacterial pathogens, most of which are transmitted through food and water \cite{53,54}. The *Vibrio cholerae* Haiti outbreak in 2010 is one of recent examples of outbreaks with a high infection rate, counting 526,524 suspected cases and 7025 death cases reported by Haitian government in the period of four month from the start \cite{55}. In light of this, rapid, accurate identification of microbial isolates is an essential task in modern epidemiology and clinical diagnostics.

2.1 Methods for microbial identification

The improvements in whole genome sequencing (WGS) techniques and bioinformatics led to the reduced cost of genome sequencing, therefore allowing the increase in the number of databases and development of new analytic tools for
microbial typing methods \cite{56}, such as Multi Locus Sequence Typing (MLST) \cite{57}. MLST is a typing method, which involves sequencing of 450-500 bp sequence fragments, of mostly six to eight housekeeping genes, that are nearly conserved in each genome. For each locus, unique sequence (allele) is given arbitrary number and, based on the combination of identified alleles (called allelic profile) the sequence type is determined \cite{58,59}. MLST established one of the first publically available typing marker databases, which led to the ability to easily share the sequence data among different research groups. In 2012, Jolley \textit{et al.} \cite{60} proposed the use of 53 genes, encoding ribosomal proteins, for ribosomal multilocus sequence typing (rMLST). This provided the possibility to in silico identify bacterial taxonomy down to the strain or subspecies level using WGS data.

Whole genome sequencing can allow several different sequence-based methods of taxonomy identification. Similarly, universally conserved genes or proteins, specific to particular taxonomic group can serve as novel targets for species and strain identification.

### 2.2 Epidemiological insight into microbial characterization

Microbial identification and characterization is also performed to support clinical diagnostics and infection control. Whole genome based analyses and comparative genomics are of raising interest in investigation of microbial outbreaks, especially when antibiotic-resistant pathogens, such as strains of \textit{Staphylococcus aureus}, \textit{Clostridium difficile}, \textit{Mycobacterium tuberculosis}, and \textit{Escherichia coli} species, are causing the infections. One of the recent outbreak examples is the outbreak of multi-drug resistant \textit{Escherichia coli} O104:H4 in Germany, in May of 2011. This strain caused bloody diarrhea and hemolytic uremic syndrome (HUS), with more than 3000 infection cases reported in Germany, and additional 100 in other European countries, USA and Canada. The total number of 46 cases resulted in death. Several groups were attempting to characterize and compare the multiple outbreak strains with the historical enterohaemorrhagic \textit{Escherichia coli} (EHEC) 0104:H4 isolate (2001, Germany) and enteroaggregative \textit{Escherichia coli} (EAEC) O104:H4 strain 55989.
(1990, Africa), during the ongoing outbreak, using whole genome based methods \cite{61, 62}. Results suggested that the outbreak is likely to be clonal and single-sourced. Strains from both German outbreaks had the similar EAEC genetic background, which doesn’t cause severe infections like HUS; and is only distant in relation to EHEC strains. However different from other typical EAEC strains, the 2001 and 2011 strains carry \textit{stx}_2-harboring prophage integrated in \textit{wrbA}, which serves also as integration site for \textit{stx}_2-phages in some EHEC O157:H7 outbreak strains. Stx-producing serotype O104:H4 are rarely extracted from patients with HUS. The plasmid content comparison showed that 2011 and 2001 strains contain a tellurite-resistance gene, which is absent in African strain. Furthermore, genes coding for aggregative adherence fimbri-aea type I (AAF/I) are also different between African and German outbreak strains \cite{61}.

This is one example of the whole genome characterization in early stages of outbreak; and in future, it may become a standard procedure, which will enable fast decisions about the treatment, source of origin, and prevention.
Part II

Projects
Chapter 3

BLAST-based comparative genomics

Comparative genomics usually starts with some sort of sequence similarity search, often performed with BLAST. This chapter includes two examples of BLAST-based comparative sequence analyses. Paper I shows the analysis between over 80 genomes of probiotic Bifidobacterium, Lactobacillus, Lactococcus, and Leuconostoc genomes, as well as a selection of Enterococcus and Streptococcus genomes, which are represented by both probiotic and pathogenic strains. Pairwise BLASTP genome comparisons were performed to define pan- and core-genomes within and between genera, as well as differences and similarities between probiotic and pathogenic strains.

Paper II demonstrates the use of the whole genome analysis of the two food-borne human-pathogen Listeria monocytogenes isolates with a purpose to identify genes or proteins that could contribute to persistence. Two sequenced strains were compared to three other publicly available strains of the same species. This study identified the genomic content that is different between the strains; however, clear conclusions could not be made about which genes are responsible for persistence.
3.1 Paper I. Comparative Genomics of *Bifidobacterium*,
*Lactobacillus* and Related Probiotic Genera
MINIREVIEWS

Comparative Genomics of *Bifidobacterium*, *Lactobacillus* and Related Probiotic Genera

Oksana Lukjancenko · David W. Ussery · Trudy M. Wassenaar

Received: 10 May 2011 / Accepted: 1 August 2011 / Published online: 27 October 2011
© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract Six bacterial genera containing species commonly used as probiotics for human consumption or starter cultures for food fermentation were compared and contrasted, based on publicly available complete genome sequences. The analysis included 19 *Bifidobacterium* genomes, 21 *Lactobacillus* genomes, 4 *Lactococcus* and 3 *Leuconostoc* genomes, as well as a selection of *Enterococcus* (11) and *Streptococcus* (23) genomes. The latter two genera included genomes from probiotic or commensal as well as pathogenic organisms to investigate if their non-pathogenic members shared more genes with the other probiotic genomes than their pathogenic members. The pan- and core genome of each genus was defined. Pairwise BLASTP genome comparison was performed within and between genera. It turned out that pathogenic *Streptococcus* and *Enterococcus* shared more gene families than did the non-pathogenic genomes. In silico multilocus sequence typing was carried out for all genomes per genus, and the variable gene content of genomes was compared within the genera. Informative BLAST Atlases were constructed to visualize genomic variation within genera. The clusters of orthologous groups (COG) classes of all genes in the pan- and core genome of each genus were compared. In addition, it was investigated whether pathogenic genomes contain different COG classes compared to the probiotic or fermentative organisms, again comparing their pan- and core genomes. The obtained results were compared with published data from the literature. This study illustrates how over 80 genomes can be broadly compared using simple bioinformatic tools, leading to both confirmation of known information as well as novel observations.

Introduction

The first bacterial genome sequences were published in 1995, and within 15 years, over a thousand fully sequenced bacterial genomes have become publicly available [16]. A number of these genome sequences are derived from bacteria used as probiotics or starter cultures in food fermentation, or both. Reid and co-workers [21] defined probiotics as “live microorganisms which when administered in adequate amounts confer a health benefit on the host”. A number of bacterial species from various genera are in use as probiotics, including members of *Lactobacillus*, *Lactococcus* and, less commonly, *Leuconostoc*. These Firmicutes are sometimes collectively described as lactic acid bacteria (LAB). Other commonly used probiotic species belong to *Bifidobacterium*, a genus within the phylum Actinobacteria. These genera exclusively contain species that are unlikely to cause disease while colonizing the intestine, and although some species (e.g. *Bifidobacterium dentium*) have been associated with dental disease, these are more commonly members of a normal oral flora. The distinction between normal gut flora (commensals) and
probiotic bacteria having a beneficial effect on their host’s health cannot always be made, for which reason we collectively describe them here as ‘non-pathogens’. Species belonging to LAB or Bifidobacterium are also frequently used in food fermentation, another application where the bacterial load of food is desirably increased. Besides LAB and Bifidobacterium, fermentation starter cultures can typically comprise of Streptococcus thermophilus, a non-pathogenic member of this genus that mostly contains pathogenic species. Some strains of Enterococcus are also in use as starter cultures or probiotics, whereby the used species also contain pathogenic strains. These two genera are therefore of interest, and their species that are used as starter cultures are included in our general description of ‘non-pathogens’. Other types of bacteria (particular strains of Escherichia coli, Pediococcus species and others) or yeasts used as starter cultures or probiotics are not treated here.

For all six genera of interest, multiple genome sequences are publicly available. In many cases, several genomes per species have been sequenced, so that the variation between and even within species can be assessed. One obvious question that could be addressed by comparison of these genomes is: what genes (if any) are common to all genomes of non-pathogens and distinct from genes found in (related) pathogens? Such a comparison requires including multiple species and genera of multiple bacterial phyla (in this case, the phylum of Firmicutes and Actinobacteria). As a general rule, genetic diversity increases with evolutionary distance, so that the genetic variation in such a collection of genomes will be enormous. One way of extracting information from such complex data is by grouping genes into functional groups or families, so that gene families rather than individual genes are compared. Such grouping is based on protein sequence similarity, as this approximately predicts conservation of gene function, ignoring the exceptions resulting from parallel evolution where function similarity does not coincide with sequence conservation. Slight differences in function, resulting from minor differences in sequences, are usually ignored in these groupings, so that fewer but broader groups can be achieved.

In this contribution, 2 approaches were used to compare over 80 genomes from 6 bacterial genera of interest. First, all protein-coding genes from these genomes were grouped into gene families based on sequence identity using a defined similarity cut-off, after which comparisons between and across genera could be performed. Genomes were then compared within their genus for both conserved and variable genes. Second, clusters of orthologous groups (COG) of genes were used to produce functional groups of genes. An attempt was made to identify differences in functional gene distribution between pathogenic and non-pathogenic members of the six genera of interest.

Materials and Methods

Selection of Genomes Used in This Study

Publicly available genomes of the six bacterial genera analyzed here were identified from the NCBI web pages. All completely sequenced genomes (as of July 2010) of 4 Lactococcus lactis strains, 3 Leuconostoc species and 21 Lactobacillus strains from 14 species were included. For Bifidobacterium, 11 completely sequenced and 8 incomplete genomes were selected; the latter were chosen when fewer than 70 contigs resulting in 19 genomes from 9 species. Since only 1 complete Enterococcus genome was available at the time of analysis, this genome was combined with 10 incomplete sequences, provided they were represented in fewer than 80 contigs, whereby animal isolates were excluded. This allowed inclusion of 2 strains obtained from normal gut flora to give 11 genomes from 4 species. For Streptococcus, all S. thermophilus genomes were included. All other species of this genus for which genome sequences were available are pathogens, and a selection of these was made of three genomes per species. These were chosen based on their strain characteristics to cover common but diverse serotypes. Animal isolates were excluded, although Streptococcus suis (a typical pig pathogen) was included as it has been responsible for a large human outbreak in China. This resulted in 23 genomes from 12 species. All genomes are listed in Table 1, which also provides characteristics such as their size, GC content and the strain description. The latter was extracted from the Genome Project pages at NCBI but checked in the corresponding genome publication when available. This resulted in a few small differences from descriptions listed on the Genome Project Description pages at NCBI. The derived proteomes (protein-coding sequences translated from the DNA sequence) were extracted from GenBank for completed sequences or produced with Prodigal [14] for incomplete sequences.

Definition of Gene Families and Pan- and Core Genome

The pan-genome of a collection of genomes represents all genes encountered in these genomes [27]. In order to define a pan-genome, the criteria to score a gene as ‘conserved’ or ‘novel’ were used as previously described [12]. Simply put, two genes are considered to belong to the same gene family and thus ‘conserved’ when their amino acid sequence is at least 50% identical over at least 50% of the length of the longest gene. All genes of a genome are thus grouped into gene families. Multiple genes per genome can belong to a single gene family, resulting in a lower number of gene families per genome than the reported number of genes. A gene not finding a match with the given criteria is put in its own gene family as a singleton.
An accumulative pan-genome was constructed according to Friis et al. [11], who built on work by Tettelin and co-workers [27]. A resulting pan-genome curve increases in size as more genomes are analyzed, and its shape is order-dependent, though the accumulative pan-genome is not influenced by the order of analysis. Similarly, a core genome is defined as all gene families conserved in all analyzed genomes, and this decreases in size as more genomes are analyzed.

Pairwise pan- and core genomes were calculated for all genome combinations as above, and for each combination, the obtained core genome was expressed as the fraction of the pan-genome. These percentages were visualized in a BLAST Matrix [11].

Core Genome Consensus Tree

Phylogenetic trees were constructed of all core genes that were conserved within the analyzed Firmicute genomes. Multiple alignments of all core sequences were performed with MUSCLE software [7]. PAUP was used to construct a set of core trees [10]. Later, these trees were compared and a best-fit consensus tree was constructed as described by Retief [22].

In Silico MLST Analysis

In silico multilocus sequence typing (MLST) analysis was performed with gene fragments extracted from the genome sequences. For Bifidobacterium, gene fragments from clpC, fusA, gyrB, IleS, purF, rplB and rpoB were extracted, according to the method proposed for Bifidobacterium bifidum, Bifidobacterium breve and Bifidobacterium longum [6]. For Enterococcus, the gene set of gdh, gyd, pstS, gki, arnE, xpt and yqlI, which is advised for use in Enterococcus faecalis (http://www.mlst.net), was compared with that designed for Enterococcus faecium, which is based on atpA, ddl, gdh, purK, gyd, pstS and adk. For Lactobacillus, de Las Rivas and co-workers [4] described an MLST gene set specified for Lactobacillus plantarum based on the target genes pgm, ddl, gyrB, purK1, gdh, mutS and tkt4. Two alternative combinations of genes have been proposed for Lactobacillus casei: fisZ, polA, mutL, metRS, trdD and pgm [1] or fusA, ileS, lepA, leuS, pyrG, recA and recG (http://www.pasteur.fr). A fourth gene set (gdh, gyrA, mapA, nox, pgmA and pta) has recently been described for Lactobacillus sanfranciscensis [20], but since this species is not represented in our dataset, this scheme was not used. For each genus, after concatenation of the gene fragments, a maximum likelihood phylogenetic tree was constructed.

Analysis of Variable Gene Content

The variable gene content of the analyzed genomes was compared using the method by Snipen and Ussery [24]. This method calculates Manhattan distances based on a matrix in which the presence or absence for each gene in each genome is scored with the binary score of 0 (absent) or 1 (present). Core genes and singletons are ignored. BLAST Atlases were produced according to Hallin and co-workers [12].

COG Analysis

COG is a database of proteins where each sequence is assigned to some group. All proteins within a group are believed to have a common ancestor and are likely to share a common function. The various groups are again clustered into some super-groups called functional groups [26]. In this analysis, each found protein was compared to the COG database using BLASTP to identify the functional groups to which they belong. An R-script was used to analyze the protein composition in pan- and core genomes, and the results were visualized in a pie chart. This was done using standard operating procedures [19].

Results

Comparison of Pan-Genomes

After the selection of genome sequences as described in the “Materials and Methods” section, 81 genome sequences were obtained from organisms listed in Table 1. These represented 43 different species and coded for 147,074 protein genes in total. Table 2 summarizes some average findings for each of the analyzed genera. Enterococcus has the largest average genome size and Leuconostoc is the smallest, a difference that is reflected in their average number of genes, since gene density is generally conserved in these bacteria. Bifidobacterium has a significantly higher CG content, which was one of the reasons to place this genus in the Actinobacteria [9]. The CG content varied most within the genus of Lactobacillus, with a CG content below 37.2% for Lactobacillus acidophilus, Lactobacillus crispatus, Lactobacillus gasseri, Lactobacillus helveticus, Lactobacillus johnsonii and Lactobacillus salivarius; genomes of the other members of this genus contain at least 38.9% CG. The average number of gene families (as defined in the “Materials and Methods” section) is also shown in Table 2. Since multiple genes per genome can belong to a single gene family, there are fewer gene families than genes per genome, but the difference is small for Bifidobacterium. This indicates that there is little gene redundancy in that genus. Lastly, the pan- and core genomes of these genera (based on the analyzed genomes) are quantified in Table 2. The plots resulting in these running totals are shown in Fig. 1, where the average
Table 1 Genomes selected for analysis

<table>
<thead>
<tr>
<th>GPID</th>
<th>Strain name</th>
<th>Size, bp or Mb</th>
<th>% CG</th>
<th>Contigs</th>
<th>Number of genes</th>
<th>Strain characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>Lactobacillus acidophilus NCFM</td>
<td>1,993,560</td>
<td>34.7</td>
<td>1</td>
<td>1,862</td>
<td>Commercial strain for yogurt, fluid milk production</td>
</tr>
<tr>
<td>404</td>
<td>Lactobacillus brevis ATCC 367</td>
<td>2,340,228</td>
<td>46.1</td>
<td>3</td>
<td>2,218</td>
<td>Starter culture for beer, sourdough, and silage</td>
</tr>
<tr>
<td>402</td>
<td>Lactobacillus casei ATCC 334</td>
<td>2,924,325</td>
<td>46.6</td>
<td>2</td>
<td>2,771</td>
<td>Starter culture for milk fermentation and flavour development of cheese</td>
</tr>
<tr>
<td>30359</td>
<td>Lactobacillus casei BL23</td>
<td>3,079,196</td>
<td>46.3</td>
<td>1</td>
<td>3,044</td>
<td>Probiotic strain</td>
</tr>
<tr>
<td>46813</td>
<td>Lactobacillus crispatus ST1</td>
<td>2,043,161</td>
<td>36.9</td>
<td>1</td>
<td>2,024</td>
<td>Normal oral/vaginal flora, chicken isolate</td>
</tr>
<tr>
<td>16871</td>
<td>Lactobacillus delbrueckii ATCC 11842</td>
<td>1,864,998</td>
<td>49.7</td>
<td>1</td>
<td>2,096</td>
<td>Yogurt</td>
</tr>
<tr>
<td>403</td>
<td>Lactobacillus delbrueckii bulgaricus ATCC BAA-365</td>
<td>1,856,951</td>
<td>49.7</td>
<td>1</td>
<td>1,721</td>
<td>Thermophilic starter culture for yogurt, Swiss and Italian-type cheeses</td>
</tr>
<tr>
<td>18979</td>
<td>Lactobacillus fermentum IFO 3956</td>
<td>2,098,685</td>
<td>51.5</td>
<td>1</td>
<td>2,044</td>
<td>Probiotic strain</td>
</tr>
<tr>
<td>84</td>
<td>Lactobacillus gasseri ATCC 3323</td>
<td>1,894,360</td>
<td>35.3</td>
<td>1</td>
<td>1,755</td>
<td>Human isolate, type strain</td>
</tr>
<tr>
<td>1781</td>
<td>Lactobacillus helveticus DPC 4571</td>
<td>2,080,931</td>
<td>37.1</td>
<td>1</td>
<td>1,610</td>
<td>Cheese culture</td>
</tr>
<tr>
<td>36575</td>
<td>Lactobacillus johnsonii FI9785</td>
<td>1,992,676</td>
<td>34.6</td>
<td>1</td>
<td>1,821</td>
<td>Probiotic strain</td>
</tr>
<tr>
<td>9638</td>
<td>Lactobacillus lactis lactis Il1403</td>
<td>2,365,589</td>
<td>35.3</td>
<td>1</td>
<td>2,266</td>
<td>Laboratory strain</td>
</tr>
<tr>
<td>403</td>
<td>Lactococcus lactis cremoris MG1363</td>
<td>2,529,478</td>
<td>35.7</td>
<td>1</td>
<td>2,516</td>
<td>Plasmid-cured NCDO712, lab strain</td>
</tr>
<tr>
<td>15766</td>
<td>Lactococcus lactis lactis DSM 20016</td>
<td>1,999,618</td>
<td>38.9</td>
<td>1</td>
<td>1,900</td>
<td>Type strain, human isolate</td>
</tr>
<tr>
<td>19011</td>
<td>Lactococcus lactis lactis DSM 1112</td>
<td>2,039,414</td>
<td>38.9</td>
<td>1</td>
<td>1,820</td>
<td>Human isolate</td>
</tr>
<tr>
<td>32195</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>2,075,763</td>
<td>37.7</td>
<td>2</td>
<td>2,005</td>
<td>Food fermentation, not specified</td>
</tr>
<tr>
<td>32197</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>1,884,661</td>
<td>43.1</td>
<td>1</td>
<td>1,885</td>
<td>Fermenting</td>
</tr>
<tr>
<td>13435</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>2,133,977</td>
<td>33.0</td>
<td>4</td>
<td>2,014</td>
<td>Probiotic strain</td>
</tr>
<tr>
<td>13280</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>2,598,348</td>
<td>35.8</td>
<td>6</td>
<td>2,504</td>
<td>Cheese production</td>
</tr>
<tr>
<td>18797</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>2,365,589</td>
<td>35.3</td>
<td>1</td>
<td>2,266</td>
<td>Laboratory strain</td>
</tr>
<tr>
<td>40637</td>
<td>Lactococcus lactis lactis ATCC 9293</td>
<td>2,635,654</td>
<td>34.9</td>
<td>1</td>
<td>2,575</td>
<td>Fermenting, non-dairy</td>
</tr>
<tr>
<td>16062</td>
<td>Leuconostoc citreum KM20</td>
<td>1,896,614</td>
<td>38.9</td>
<td>5</td>
<td>1,823</td>
<td>Kimchi (food, Korea)</td>
</tr>
<tr>
<td>20843</td>
<td>Leuconostoc kimchii IMSNU11154</td>
<td>2,101,787</td>
<td>37.0</td>
<td>1</td>
<td>2,130</td>
<td>Kimchi? not specified</td>
</tr>
<tr>
<td>315</td>
<td>Leuconostoc mesenteroides mesenteroides ATCC 8293</td>
<td>1,875,974</td>
<td>47.1</td>
<td>2</td>
<td>1,885</td>
<td>Human isolate</td>
</tr>
<tr>
<td>32949</td>
<td>Enterococcus faecalis T11</td>
<td>2,729,089</td>
<td>37.7</td>
<td>49</td>
<td>2,522</td>
<td>Urine isolate</td>
</tr>
<tr>
<td>32941</td>
<td>Enterococcus faecalis E1Sol</td>
<td>2,853,151</td>
<td>37.5</td>
<td>75</td>
<td>2,737</td>
<td>Faecal isolate, antibiotic-naïve, normal flora</td>
</tr>
<tr>
<td>20843</td>
<td>Enterococcus faecalis OG1RF</td>
<td>2,739,625</td>
<td>37.7</td>
<td>1</td>
<td>2,515</td>
<td>No info - lab strain?</td>
</tr>
<tr>
<td>32919</td>
<td>Enterococcus faecalis T3</td>
<td>2,821,089</td>
<td>37.6</td>
<td>40</td>
<td>2,603</td>
<td>Urine isolate</td>
</tr>
<tr>
<td>32927</td>
<td>Enterococcus gallinarum EG2</td>
<td>3,134,429</td>
<td>40.6</td>
<td>49</td>
<td>2,985</td>
<td>No info</td>
</tr>
<tr>
<td>32931</td>
<td>Enterococcus casseliflavus EC10</td>
<td>3,423,270</td>
<td>42.5</td>
<td>54</td>
<td>3,243</td>
<td>No info</td>
</tr>
<tr>
<td>32935</td>
<td>Enterococcus casseliflavus EC20</td>
<td>3,392,502</td>
<td>42.8</td>
<td>57</td>
<td>3,121</td>
<td>No info</td>
</tr>
<tr>
<td>46979</td>
<td>Enterococcus faecium PC4.1</td>
<td>2,811,160</td>
<td>37.9</td>
<td>78</td>
<td>2,705</td>
<td>Human microbiome, normal flora</td>
</tr>
<tr>
<td>32965</td>
<td>Enterococcus faecium Com12</td>
<td>2,685,402</td>
<td>38.1</td>
<td>67</td>
<td>2,573</td>
<td>No info</td>
</tr>
<tr>
<td>32967</td>
<td>Enterococcus faecium Com15</td>
<td>2,771,455</td>
<td>38.3</td>
<td>70</td>
<td>2,698</td>
<td>No info</td>
</tr>
<tr>
<td>330</td>
<td>Streptococcus agalactiae 2603V/R</td>
<td>2,160,267</td>
<td>35.6</td>
<td>1</td>
<td>2,124</td>
<td>Clinical isolate, common in adults</td>
</tr>
<tr>
<td>326</td>
<td>Streptococcus agalactiae A909</td>
<td>2,127,839</td>
<td>35.6</td>
<td>1</td>
<td>1,996</td>
<td>No info</td>
</tr>
<tr>
<td>334</td>
<td>Streptococcus agalactiae NEM316</td>
<td>2,211,485</td>
<td>35.6</td>
<td>1</td>
<td>2,134</td>
<td>Blood isolate</td>
</tr>
<tr>
<td>27849</td>
<td>Streptococcus dysgalactiae equisimilis GGS 124</td>
<td>2,106,340</td>
<td>39.6</td>
<td>1</td>
<td>2,100</td>
<td>No info</td>
</tr>
<tr>
<td>34729</td>
<td>Streptococcus galolyticus UCN34</td>
<td>2,350,911</td>
<td>37.6</td>
<td>1</td>
<td>2,261</td>
<td>Normally rumen flora, this is a clinical human isolate</td>
</tr>
<tr>
<td>GPID</td>
<td>Strain name</td>
<td>Size, bp or Mb</td>
<td>% CG</td>
<td>Contigs</td>
<td>Number of genes</td>
<td>Strain characteristics from endocarditis</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>----------------</td>
<td>------</td>
<td>---------</td>
<td>----------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>66</td>
<td>Streptococcus gordonii str. Challis CH1</td>
<td>2,196,662</td>
<td>40.5</td>
<td>1</td>
<td>2,051</td>
<td>Causes caries and periodontal diseases</td>
</tr>
<tr>
<td>20527</td>
<td>Streptococcus infantarius infantarius ATCC BAA-102</td>
<td>1,925,087</td>
<td>37.6</td>
<td>22</td>
<td>1,962</td>
<td>Human microbiome project, normal flora</td>
</tr>
<tr>
<td>16302</td>
<td>Streptococcus mitis B6</td>
<td>2,146,611</td>
<td>40.0</td>
<td>1</td>
<td>2,018</td>
<td>Clinical isolate</td>
</tr>
<tr>
<td>28997</td>
<td>Streptococcus mutans NN2025</td>
<td>2,013,587</td>
<td>36.8</td>
<td>1</td>
<td>1,895</td>
<td>Normally oral flora, can cause caries, endocarditis. Clinical isolate</td>
</tr>
<tr>
<td>333</td>
<td>Streptococcus mutans UA159</td>
<td>2,030,921</td>
<td>36.8</td>
<td>1</td>
<td>1,960</td>
<td>Oral flora, can cause caries, caries isolate</td>
</tr>
<tr>
<td>31233</td>
<td>Streptococcus mitis</td>
<td>2,121,351</td>
<td>39.5</td>
<td>1</td>
<td>2,135</td>
<td>Alternative name Spain 23FST81. Pandemic, high prevalence, invasive</td>
</tr>
<tr>
<td>29047</td>
<td>Streptococcus pneumoniae ATCC 70069</td>
<td>2,078,953</td>
<td>39.7</td>
<td>1</td>
<td>2,115</td>
<td>Resistant clinical isolate</td>
</tr>
<tr>
<td>277</td>
<td>Streptococcus pneumoniae</td>
<td>2,160,842</td>
<td>39.7</td>
<td>1</td>
<td>2,125</td>
<td>Virulent clinical isolate</td>
</tr>
<tr>
<td>269</td>
<td>Streptococcus pyogenes M1 GAS SF370</td>
<td>1,852,441</td>
<td>38.5</td>
<td>1</td>
<td>1,696</td>
<td>Group A</td>
</tr>
<tr>
<td>16364</td>
<td>Streptococcus pyogenes MGAS10270</td>
<td>1,928,252</td>
<td>38.4</td>
<td>1</td>
<td>1,987</td>
<td>Sequenced for comparative genome analysis</td>
</tr>
<tr>
<td>286</td>
<td>Streptococcus pyogenes MGAS8232</td>
<td>1,895,017</td>
<td>38.5</td>
<td>1</td>
<td>1,845</td>
<td>Serotype M18</td>
</tr>
<tr>
<td>13942</td>
<td>Streptococcus suis G54</td>
<td>2,388,435</td>
<td>43.4</td>
<td>1</td>
<td>2,270</td>
<td>Indigenous oral bacteria, causes dental decay, oral plaque isolate</td>
</tr>
<tr>
<td>17153</td>
<td>Streptococcus suis 05ZYH33</td>
<td>2,096,369</td>
<td>41.1</td>
<td>1</td>
<td>2,186</td>
<td>Causes disease in pigs and occasionally humans</td>
</tr>
<tr>
<td>32233</td>
<td>Streptococcus suis BM407</td>
<td>2,170,808</td>
<td>41.0</td>
<td>2</td>
<td>2,058</td>
<td>Human clinical isolate</td>
</tr>
<tr>
<td>18737</td>
<td>Streptococcus suis GZ1</td>
<td>2,038,034</td>
<td>41.4</td>
<td>1</td>
<td>1,978</td>
<td>Causes meningitis, arthritis, pneumonia in pigs human epidemic in China</td>
</tr>
<tr>
<td>13163</td>
<td>Streptococcus thermophilus CNRZ1066</td>
<td>1,796,226</td>
<td>39.1</td>
<td>1</td>
<td>1,915</td>
<td>Isolated from yogurt for industrial dairy fermentations</td>
</tr>
<tr>
<td>13773</td>
<td>Streptococcus thermophilus LMD-9</td>
<td>1,864,178</td>
<td>39.1</td>
<td>3</td>
<td>1,716</td>
<td>Used in the manufacture of fermented dairy foods</td>
</tr>
<tr>
<td>13162</td>
<td>Streptococcus thermophilus LMG 18511</td>
<td>1,796,846</td>
<td>39.1</td>
<td>1</td>
<td>1,889</td>
<td>Isolated from yogurt for industrial dairy fermentations</td>
</tr>
<tr>
<td>1621</td>
<td>Bifidobacterium adolescentis ATCC 15703</td>
<td>2,089,645</td>
<td>59.2</td>
<td>1</td>
<td>1,631</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>19423</td>
<td>Bifidobacterium animalis lactis AD01</td>
<td>1,933,695</td>
<td>60.5</td>
<td>1</td>
<td>1,528</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>42883</td>
<td>Bifidobacterium animalis lactis BB-12</td>
<td>1,942,198</td>
<td>60.5</td>
<td>1</td>
<td>1,642</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>32897</td>
<td>Bifidobacterium animalis lactis BI-04</td>
<td>1,938,709</td>
<td>60.5</td>
<td>1</td>
<td>1,567</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>32893</td>
<td>Bifidobacterium animalis lactis DSM 10140</td>
<td>1,938,483</td>
<td>60.5</td>
<td>1</td>
<td>1,566</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>32515</td>
<td>Bifidobacterium animalis lactis V9</td>
<td>1,944,050</td>
<td>60.4</td>
<td>1</td>
<td>1,572</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>28807</td>
<td>Bifidobacterium animalis lactis HN019</td>
<td>1,915,892</td>
<td>60.4</td>
<td>28</td>
<td>1,578</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>17583</td>
<td>Bifidobacterium animalis lactis Rd1</td>
<td>2,636,367</td>
<td>58.5</td>
<td>1</td>
<td>2,129</td>
<td>Normal oral and gut flora, can cause caries, caries isolate</td>
</tr>
<tr>
<td>20555</td>
<td>Bifidobacterium animalis lactis ATCC 27678</td>
<td>2,642,081</td>
<td>58.5</td>
<td>2</td>
<td>2,151</td>
<td>Human microbiome, faeces isolate</td>
</tr>
<tr>
<td>18773</td>
<td>Bifidobacterium animalis lactis LMG 18511</td>
<td>2,389,526</td>
<td>60.2</td>
<td>3</td>
<td>2,003</td>
<td>Normal gut flora, probiotic</td>
</tr>
<tr>
<td>328</td>
<td>Bifidobacterium animalis lactis NCC2705</td>
<td>2,260,266</td>
<td>60.1</td>
<td>2</td>
<td>1,729</td>
<td>Normal gut flora, probiotic</td>
</tr>
<tr>
<td>17189</td>
<td>Bifidobacterium animalis lactis ATCC 15697</td>
<td>2,832,748</td>
<td>59.9</td>
<td>1</td>
<td>2,416</td>
<td>Normal gut flora, probiotic</td>
</tr>
<tr>
<td>30065</td>
<td>Bifidobacterium animalis lactis CCUG 52486</td>
<td>2,453,376</td>
<td>60.2</td>
<td>55</td>
<td>2,085</td>
<td>Normal gut flora, human microbiome project</td>
</tr>
<tr>
<td>47579</td>
<td>Bifidobacterium animalis lactis JDM301</td>
<td>2,477,838</td>
<td>59.8</td>
<td>1</td>
<td>1,959</td>
<td>Normal gut flora, probiotic</td>
</tr>
<tr>
<td>29261</td>
<td>Bifidobacterium animalis lactis DSM 20098</td>
<td>2,007,108</td>
<td>59.4</td>
<td>17</td>
<td>1,586</td>
<td>Normal gut flora, type strain</td>
</tr>
<tr>
<td>30055</td>
<td>Bifidobacterium animalis lactis NCIMB 41171</td>
<td>2,186,140</td>
<td>62.8</td>
<td>33</td>
<td>1,810</td>
<td>Normal gut flora, probiotic</td>
</tr>
<tr>
<td>30749</td>
<td>Bifidobacterium animalis lactis NCC2705</td>
<td>2,058,429</td>
<td>56.1</td>
<td>31</td>
<td>1,720</td>
<td>Normal gut flora</td>
</tr>
<tr>
<td>30751</td>
<td>Bifidobacterium animalis lactis DSM 20098</td>
<td>2,019,802</td>
<td>57.5</td>
<td>27</td>
<td>1,580</td>
<td>Human microbiome project</td>
</tr>
<tr>
<td>30373</td>
<td>Bifidobacterium animalis lactis DSM 20438</td>
<td>2,304,808</td>
<td>56.3</td>
<td>36</td>
<td>1,870</td>
<td>Human microbiome project</td>
</tr>
</tbody>
</table>

a The official abbreviation 'subsp.' between species and subspecies name has been deleted throughout this contribution.

GPID genome project identification number (NCBI: see http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi), NA not available.
number of gene families present per genome is given as a green line. In all graphs, the pan-genome and core genome curves strongly diverge, indicative of a large variation in gene content between the analyzed genomes within each genus. The largest difference between the pan- and core genome, as a measure for the variance within the analyzed genera, is seen with Lactobacillus (21 genomes of 14 species) and Streptococcus (23 genomes of 12 species). The variance is larger in four genomes of Lc. lactis than in three different Leuconostoc species. Thus, intra-species variation in gene content of Lc. lactis exceeds inter-species variation of Leuconostoc, at least for these analyzed genomes.

The pan- and core genomes of pairwise genome comparisons were also determined to establish the percentage identity for each combination. This identity was expressed as the pairwise core genome divided by its pan-genome and was visualized by colour intensity in a BLAST Matrix. Figure 2 shows the BLAST Matrix for the Lactobacillus genomes. The strongest green, indicative of the highest fraction of genes found similar between two genomes, are reported for comparisons within a species, shown at the bottom of the figure. Some species also share a large fraction of genes between them. For instance, the two Lb. casei genomes share between 55.5% and 59.3% of their genes with those of the three Lactobacillus rhamnosus genomes (represented in the six darker green cells in the upper part of the matrix). An even higher similarity (62.2–62.8%) is found between Lb. gasseri and Lb. johnsonii. The highest similarity recorded is 93.3%, between two Lb. rhamnosus strains, and the lowest is 11.5%, between Lb. casei BL23 and Lactobacillus delbrueckii bulgaricus ATCCBAA-365.

A similar matrix is shown for Bifidobacterium in Fig. 3. In this case, the similarity between the six Bifidobacterium animalis genomes is obvious (visible as 15 strongly coloured cells at the bottom right). Two of these genomes reach a similarity of 95.5%. The lowest degree of similarity is seen between Bifidobacterium gallicum and B. longum infantis strain ATCC 15697 (28.5%).

When a BLAST Matrix was constructed with all genomes included in the analysis, the similarity between Bifidobacterium genomes and those of the other genera remained below 3%, illustrative of the difference of Bifidobacterium compared to the Firmicutes (results not shown). Thus, despite their sharing of an ecological niche, these bacteria share relatively few genes. A comparison of all Firmicute genomes is provided as Supplementary Fig. S1. As expected, the found percentage identity within any of these genera is much higher than that between genera. For instance, the three Leuconostoc genomes produced a similarity of 49.5–52.3% between them, but around 8% to 10% to genomes of other genera. The four Lc. lactis genomes gave slightly higher similarities of 16.1–18.4% to all other Firmicute genomes whilst sharing 59.5–66.1% between themselves. An Enterococcus and a Streptococcus genome typically share 10% to 15% of their genes, and two genomes of Enterococcus and Lactococcus 14% to 16%. Different Enterococcus species share around 30% of their genes, but multiple genomes within one species of this genus have around 70% of their genes being similar.

Comparison of Core Genomes and Conserved Genes

The pan-genomes of all six genera were combined to calculate the core genome shared by all genera. This resulted in only 63 core gene families out of a pan-genome of 37,053 gene families, using the criteria of gene similarity as described in the “Materials and Methods” section. These are listed in Supplementary Table S1. Exclusion of the distinct Bifidobacterium genus retained 243 core gene families for the Firmicute genomes that together produced a pan-genome of 30,615 gene families. Since these core genes are conserved in all Firmicute genomes analyzed here, phylogenetic trees could be generated and a consensus tree was generated, as shown in Fig. 4. The consensus core gene tree split all Lactobacillus genomes into three main clusters, though Lb. salivarius is excluded from these groups. The cluster shown at the top of the figure contains most Lactobacillus species.

### Table 2: Average findings per genus and their pan- and core genome

<table>
<thead>
<tr>
<th>Genus</th>
<th>Number of genomes included</th>
<th>Number of species</th>
<th>Average genome size (kbp)</th>
<th>Average % CG</th>
<th>Average number of genes (min–max values)</th>
<th>Average number of gene families (min–max values)</th>
<th>Pan-genome</th>
<th>Core genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactobacillus</td>
<td>21</td>
<td>14</td>
<td>2,369</td>
<td>42.4</td>
<td>2,235 (1,562–3,059)</td>
<td>2,071 (1,437–2,873)</td>
<td>13,069</td>
<td>363</td>
</tr>
<tr>
<td>Lactococcus</td>
<td>4</td>
<td>1</td>
<td>2,532</td>
<td>35.4</td>
<td>2,465 (2,266–2,504)</td>
<td>2,238 (2,118–2,341)</td>
<td>3,389</td>
<td>1,522</td>
</tr>
<tr>
<td>Leuconostoc</td>
<td>3</td>
<td>3</td>
<td>2,025</td>
<td>37.9</td>
<td>1,986 (1,820–2,130)</td>
<td>1,896 (1,724–2,050)</td>
<td>2,927</td>
<td>1,164</td>
</tr>
<tr>
<td>Enterococcus</td>
<td>11</td>
<td>4</td>
<td>3,041</td>
<td>36.6</td>
<td>3,078 (2,573–2,515)</td>
<td>2,707 (2,439–3,114)</td>
<td>7,519</td>
<td>1,092</td>
</tr>
<tr>
<td>Streptococcus</td>
<td>23</td>
<td>12</td>
<td>1,981</td>
<td>38.9</td>
<td>2,018 (1,696–2,270)</td>
<td>1,923 (1,643–2,180)</td>
<td>9,785</td>
<td>638</td>
</tr>
<tr>
<td>Bifidobacterium</td>
<td>19</td>
<td>9</td>
<td>2,209</td>
<td>59.5</td>
<td>1,796 (1,528–2,416)</td>
<td>1,746 (1,497–2,287)</td>
<td>6,980</td>
<td>724</td>
</tr>
</tbody>
</table>

*Number of gene families is given.
with lower CG content, though it also includes *L. delbrueckii*, whose CG content is quite a bit higher. This clustering, based on these core genes, corroborates the interstrain similarities already reported for their complete genomes, as shown in Fig. 2. The *Streptococcus* genus is separated into two large clusters in Fig. 4. Two clusters are also observed for the *Enterococcus* species, while *Lactococcus* is placed outside all other genera.

A more commonly used procedure is to compare only a small subset of core genes. In population biology, MLST of six or seven core gene fragments is frequently used to assess evolutionary distances between isolates within a species. MLST analysis is based on DNA sequences. We adapted this approach to perform in silico MLST for all isolates within a genus, as a measure for evolutionary distance of core genes, and used this for analysis of three genera. Unfortunately, despite the reputation of MLST as being generally applicable and despite a considerable number of gene families being conserved even between Firmicutes and Bifidobacteria...
(63 gene families), different MLST target gene sets have been proposed for various species, and most of these are not conserved between all species (Supplementary Table S1). In order to compare our findings with published data, we have used fragments of various genes depending on the genus, as suggested in the literature.

For in silico MLST analysis of *Bifidobacterium*, 7 gene fragments were extracted according to Deltoile and co-workers [6], as these happened to be conserved in all 19 *Bifidobacterium* genomes analyzed here. A phylogenetic tree of the extracted and concatenated MLST fragments is shown in Fig. 5. Although MLST was not designed for this purpose, the results show that this approach can reveal phylogenetic relationship of these core genes between species within a genus. All multiple isolates per species are correctly clustered, although subspecies are not correctly grouped (see the position of *B. longum longum* and *B. longum infantis*). Three major clusters can be recognized, separated in the figure by green lines. These findings are in accordance to the three groups within this genus recognized by Lee and O’Sullivan [17], based on an extensive 16S ribosomal RNA gene analysis.

The MLST website (http://www.mlst.net) lists two different gene sets to be used for Enterococci. Figure 5 (right side) shows the results obtained with each. Both trees produce little resolution within the species, especially when compared with the consensus tree based on 243 core genes in the previous figure.

For Lactobacilli, four MLST schemes are available: one for *L. plantarum* [4], two for *L. casei* ([1], http://www.pasteur.fr) and one for *L. sanfranciscensis* [20], which is not represented in our dataset. The first three MLST schemes were tested, which produced different trees (Supplementary Fig. S2). All three trees clustered multiple strains per species, but the branch positions of these species varied according to the gene set used. It cannot be stated which MLST tree is ‘correct’ as they all display the evolutionary relationship of the genes analyzed in question—but obviously, the phylogeny of core genes is not always conserved within a genome, as it is affected by recombination. This is also visible from the numbers of core genes producing consensus branches in Fig. 4. With this variation in mind, an MLST tree should be interpreted with caution, as it represents only a tiny fraction of the complete core genome of a strain.

Comparison of Variable Gene Content

The pan-genome of a species or genus comprises both conserved core and variable genes. The latter can also be

![Figure 2 BLAST Matrix for the Lactobacillus genomes. To the side, the total number of protein genes and gene families are listed for each genome. In the matrix cells, the shared protein genes are given as a percentage, based on the ratio of the core genome and pan-genome of each pair, as indicated.](image-url)
used to establish inter-genome relationships, although not by phylogeny. Instead, clustering of presence or absence of variable genes can be performed [24]. This method calculates Manhattan distances for genes variably present. Obviously, core genes and genes found present in only one genome were excluded from this analysis, as they cannot identify any correlation between genomes. Thus, only genes whose presence varies, found at least in two genomes but absent in at least one genome, are assessed. The resulting clustering is not a phylogenetic tree, since it is not based on phylogeny of individual genes. Instead, it shows which genomes share more of their variable genes than others.

Figure 6 shows the hierarchical clustering of the *Bifidobacterium* genomes based on their variable gene content. As can be seen, genomes of identical species cluster together and are separated from different species, but the subspecies of *B. longum* are not correctly separated. Since their variable gene content seems to be mixed, this suggests that these two subspecies share the same gene pool for horizontal gene transfer events. The similarity, in terms of variable gene content, between the two species *B. catenulatum* and *B. pseudocatenulatum* is not more than that between various *B. longum* subspecies. A deep division splits *B. animalis* combined with *B. gallicum* from the others, which correlates with the MLST tree shown in Fig. 5.

The analysis of variable gene content can simultaneously be performed with genomes of varying similarity, so that Fig. 6b combines all Firmicute genomes. The 21 *Lactobacillus* genomes are split into two major groups, which match a deep branch in the phylogenetic tree of 16S rRNA genes of this genus [2]. However, the clustering based on variable gene content produces a different picture to the consensus tree based on core genes (compare Figs. 4 and 6b). This probably reflects different evolutionary forces at play. Genes whose presence is variable may be located on mobile elements or may be more frequently subjected to DNA recombination than core genes. The three *Leuconostoc* genomes are placed within the *Lactobacillus* genus; apparently, these share a considerable number of variable genes.

The three major clusters within the *Streptococcus* genus visible in Fig. 6 largely match their taxonomic relationship as defined by 16S rRNA [8], although the distance between *S. thermophilus* and *Streptococcus infantarius*, which are both part of the ‘Salivarius group Streptococci’, is better captured by variable gene content than by 16S rRNA phylogeny. The discrepancy between this clustering and the consensus core gene tree is even more extensive for this genus.

The four *Lc. lactis* genomes are placed between *Streptococcus* and *Enterococcus*, which reminds of their
inclusion, prior to the 1980s, into the single genus *Streptococcus* [25]. Within the genus *Enterococcus*, the clustering in Fig. 6 separates each of the analyzed species and confirms that *Enterococcus casseliflavus* and *Enterococcus gallinarum* are more related to *E. faecium* than to *E. faecalis*.

Visualization of Conserved and Variable Gene Content

Conservation and variation in gene content between genomes can also be visualized by a BLAST Atlas [12], which contains information on gene location as well as on gene presence, at least for the reference genome on which a BLAST Atlas is based. Two different *Bifidobacterium* reference genomes were used in the two BLAST Atlases shown in Fig. 7 to which all other *Bifidobacterium* genomes were compared. Only genes present in the reference genome are captured in these atlases as these are used as query, for which the hits in the other genomes are recorded as colour in the BLAST lanes. The more strongly a protein gene is conserved, the more intense the colour is. Different colours are used to separate the different species, and these colours have been kept constant between the two panels, so that it is obvious that genes are mostly conserved within a species. The most inner BLAST lane included in Fig. 7 is that of the reference genome against itself. This shows the maximum colour that can be obtained for each location. Gaps in this “Blast-to-self” lane where BLAST hits are absent, for instance around 1,700 kb, are due to

---

**Figure 4** Consensus tree of 243 core genes conserved in all analyzed Firmicutes. The number of genes supporting the branches is shown in red. Values for fewer than 100 genes are not shown.

**Figure 5** In silico MLST of gene fragments extracted from the genomes of *Bifidobacterium* (a) and *Enterococcus* (b, c). b The genes selected for MLST of *E. faecium*; c the genes selected for *E. faecalis* were used.
Figure 6 Hierarchical clustering of the *Bifidobacterium* genomes (a) and the Firmicute genomes (b) based on their variable gene content. The scale at the bottom applies to both trees.
non-translated genes such as ribosomal RNA copies. In Fig. 7a, a large region around 350–400 kb appears to produce a gap of non-conserved genes in most *Bifidobacterium* genomes, with the exception of *B. longum* infantis CCUG 52486 and *B. longum* DJO10A. This represents a region with variable genes within the *B. longum* genomes (the red lanes in the atlas), which are completely absent in the other *Bifidobacterium* genomes. Other than that, there appears to be relatively little variation between the *B. longum* genomes. Strong conservation within the species is also observed for *B. animalis* when used as the reference, as shown in Fig. 7b. In that lower panel, the *B. animalis* lanes are far more darkly coloured than in the top panel, whereas the *B. longum* lanes are lighter in colour, illustrating that stronger homology is identified within a species than across species. Note that the large gap of the top atlas is no longer visible now, as the genes that were found in *B. longum* are absent in *B. animalis* and thus are no longer captured when the latter is used as a reference. Taken together, these data suggest that there is relatively strong conservation within a species of *Bifidobacterium*, an observation that has been made by others as well [30].

Figure 8 shows two BLAST Atlases of the *Lactobacillus* genomes. There appears to be considerably less conservation between species of this genus compared to *Bifidobacterium*. Even within the species of the two reference genomes of both panels, there are multiple gaps. This reflects the higher genetic diversity of the *Lactobacillus* genus compared to *Bifidobacterium*.

A BLAST Atlas of *Streptococcus* genomes with *S. thermophilus* LMD-9 as the reference is provided as Supplementary Fig. S3. Two non-pathogenic *E. faecalis* genomes were included as well, since these are normal human flora strains and could be considered to share a similar niche to *S. thermophilus*, at least when colonizing the human gut. There is quite a bit of variation in protein-coding genes between the three *S. thermophilus* genomes, and as expected, there is even fewer conservation in other species of *Streptococcus* or in the two *E. faecalis* genomes. Apparently, similarity in bacterial lifestyle is not necessarily represented by a significant homology in gene content.

**COG Comparison of Pan- and Core Genomes**

So far, conservation of genes was assessed and reported irrespective of their function, but that information is essential for a biological interpretation. The function of genes is not always known, but a large number of proteins have been assigned to a functional category of orthologous group, based on inference of sequence similarity to functionally characterized proteins. We have extracted the top-level COG groups for the genomes of interest and, in a first step, compared their core and pan-genomes genes. An example of such a statistical analysis for *Bifidobacterium* is shown in Fig. 9. At the bottom, the legend specifies the 3 top-level COG categories: ‘information storage and processing’, ‘cellular processes and signaling’ and ‘metabolism’, which are divided into 18 groups. The pie charts show what the fraction of the complete pan-genome genes of *Bifidobacterium* (left) or of the conserved core genes (right) belongs to each COG group. As expected, genes for which a function is not precise or not at all predicted build a significant fraction in the pan-genome, but these are mostly removed from the core genes, as their presence varies. More surprisingly, the three top categories are more or less similarly distributed in the two pie charts (thereby ignoring the contribution of the grey and black fractions), with a slight overrepresentation only of the information storage genes in the core genome compared to the pan-genome. Within these three broad categories, however, differences are visible when comparing the pan-genome or the core genome of these *Bifidobacterium* genomes. For instance, within ‘information storage and processing’, class J (translation, ribosomal structure and biogenesis) is enriched in the core genome, at the expense of K and L (transcription and replication, respectively). This means that the gene content related to these latter information storage processes is more variable and is hence captured in the pan-genome but less so in the core genome than the genes related to translation and ribosome biogenesis. Of interest is also the shift within the group ‘metabolism’ between classes E and G (for amino acid and carbohydrate transport/metabolism, respectively). The results indicate that the gene content for metabolism of amino acids is more conserved than that for carbohydrates, at least between these *Bifidobacterium* genomes. Lastly, enrichment in the core genome of class O, for post-translational modification and chaperones, is apparent within the group ‘cellular processes and signaling’.

The *Bifidobacterium* findings can be compared to those of *Lactobacillus*, shown at the top of Fig. 10. The distribution of the three top-level COG categories in the pan-genome of *Lactobacillus* is different to that of *Bifidobacterium*, with more information storage and fewer metabolism genes. This is more obvious from Table 3, which lists the relative fractions of these COG classes when the grey and black fractions are ignored. For the core genes of *Lactobacillus*, the relative increase (compared to its pan-genome) in the fraction of information, storage and processing genes, at the expense of metabolism genes, is far more pronounced than for *Bifidobacterium*. Within the information and storage group, the enrichment of class J genes in the core genome of *Lactobacillus* is also stronger than reported for *Bifidobacterium*.

Figure 10 also shows the plots for *Lactococcus* (middle) and *Leuconostoc* (bottom). Although these last two genera are represented by four and three genomes only, all pan-
genomes look surprisingly similar. However, when concentrating on the functionally annotated genes only (Table 3), some differences become apparent. The core genes of *Lactococcus* and *Leuconostoc* display a similar distribution of the three major COG classes as *Bifidobacterium* (which is taxonomically removed) that is different to the core genome of *Lactobacillus*, to which they are much closer related. Note that, in their pan-genomes, these three COG groups are similarly divided in *Bifidobacterium* and *Lactobacillus*. The shifts observed between pan-genome and core genome within a genus are contrasting between *Lactobacillus* and *Lactococcus*, whereas there is hardly a shift for *Leuconostoc*. From Fig. 10, it can be seen that, in the pan-genome of *Lactococcus*, class L genes make up a relatively large proportion. Within the metabolic gene classes, for *Lactobacillus*, a strong enrichment of nucleotide metabolism genes (class F) is observed in the core genes, whereas genes related to amino acid metabolism (class E) are more favoured in the core genome of *Lactobacillus*. A significant increase in the core genes of COG class O (post-translational modification and chaperones) is observed for all analyzed genera. This could be an indication of the importance for such genes in the natural habitat of these gut bacteria.

The COG distribution plots for the pan-genome genes and the core genes of *Enterococcus* and *Streptococcus* is provided as Supplementary Fig. S4; the percentages of the three functionally classified COG top levels are included in Table 3. In contrast to the above examples, these two genera contain both pathogenic and non-pathogenic isolates. As in the previous examples, the large fraction of genes with unknown function is minimized in the core genome, but for both genera. Metabolism genes are neither over- nor underrepresented in the core genome. As before, a strong conservation of genes of COG class J (translation, ribosomal structure and biogenesis) was observed. Carbohydrate transport and metabolism genes (class G) were more frequently found in the *Enterococcus* pan-genome than in the *Streptococcus* pan-genome, though this was less pronounced for their core genomes.

In an attempt to correlate findings with the presence or absence of pathogenicity, all genomes of pathogenic isolates (irrespective of their genus) were combined to collectively compare these with the non-pathogens (probiotic, fermentative and normal gut flora organisms) combined. The pathogenic group consisted of *Enterococcus* and *Streptococcus* genomes only, whilst the non-pathogenic group contained genomes of all genera analyzed. The COG analysis was then repeated for these two phenotypic collections, whereby the pan- and core genomes obviously were recalculated. The pathogenic collection had a pan-genome of 14,209 gene families and a core genome of 508. The pan-genome of the non-pathogenic collection was significantly larger (21,087), and this group produced a core genome of only 278 gene families. The results of the COG analysis are shown in Fig. 11. Surprisingly, the two pan-genome statistics look nearly identical, despite the obvious phenotypic differences between these two groups that both consist of diverse organisms, with a skewed genus distribution. However, the COG distribution between the two core genomes differs dramatically. The fraction of genes for which no homologue could be identified has (nearly) disappeared from the core genome of the non-pathogenic group, but a significant fraction of these genes was retained in the core genome of pathogens. The top level of metabolism genes has decreased in both core genomes, but more so in the group of the non-pathogens. Thus, the core genes of the non-pathogenic isolates are more frequently information storage genes and less likely metabolism genes than the core genes of pathogens (Table 4). Zooming in on shifts in single categories between pan- and core genomes, the enrichment of core genes belonging to class J, already observed for all single genus plots shown above, is even more extensive and far more extreme with the collection of non-pathogenic organisms. An enrichment for class O (post-translational modification and chaperones) within the top-level ‘metabolism’ is pronounced in the core genome of both groups, but the pathogens also show enrichment of class M genes (cell wall/membrane biogenesis) which is actually reduced in the core genome of non-pathogens.

**Discussion**

The comparative analysis presented here of 81 bacterial genomes, covering 6 genera and 43 different species, could be performed by grouping their genes into gene families and comparing core and pan-genomes of various subsets of genomes. The findings frequently confirmed taxonomic relationships but could not identify common signatures, in terms of gene content, for all non-pathogenic bacteria included in the analysis. This finding is surprising, as all these species occupy a similar niche. Conserved genes were compared by means of a consensus tree, while genes variably present were analyzed by cluster analysis. The latter indicated that *Leuconostoc* genomes share a considerable number of variable genes with *Lactobacillus*. Functional analysis of the proteins coded by the genes comprising a genus’ core genome
identified the relative strong conservation of information storage genes; this was observed for all genera analyzed. When all genomes were divided into a pathogenic and a non-pathogenic group, the pan-genome of both groups showed a surprisingly similar COG distribution; however, their core genome differed considerably. It was observed that, in the core genome of non-pathogenic genomes, genes for post-translational modification and chaperones were enriched.

**Figure 8** BLAST Atlas of *Lactobacillus* with *L. rhamnosus* strain Le705 (top) and *Lb. johnsonii* strain NCC533 (bottom) as the

---

**Figure 9** COG statistics for the genes found in the pan-genome (left) and core genome (right) of *Bifidobacterium* genomes. The key for the COG classes is explained below the pie charts. Percentages given in the pie chart are calculated by exclusion of classes R, S and X. Only values ≥5% are shown.
A simultaneous comparison of the pan- and core genomes of publicly available genomes of *Lactobacillus*, *Lactococcus*, *Leuconostoc*, *Enterococcus*, *Streptococcus* and *Bifidobacterium*, as was performed here, has not been published before, but similar analyses have been published for smaller selections of organisms. Canchaya and co-workers [2] performed comparative genomics of the then five available *Lactobacillus* genomes from different species and commented on the high variability within this genus. Schleifer and Ludwig [23] stated that “It is widely recognized that the taxonomy of this genus is unsatisfactory due to the highly heterogeneous nature of its members”. Indeed, data presented here illustrate the diversity within *Lactobacillus*. However, the heterogeneity of this genus is not larger than that of other bacteria. Using the same comparison criteria as applied here, the pan-genome of 53 *E. coli* genomes was found to comprise 13,000 gene families, even within this single species [18]. Similarly, an analysis of 27 genomes from 7 *Vibrio* species produced a pan-genome of nearly 15,000 gene families for this genus [31], and 38 genomes of 5 *Burkholderia* species contained as much as 26,000 gene families [28]. Thus, the diversity in gene content within the genus *Lactobacillus*, based on the genome sequences currently available, is not exceptional in the bacterial world.

Our analyses are mainly based on core genomes, an approach that others followed as well [2]. Those authors had defined a core genome for *Lactobacillus* whose size is similar to our findings. However, the fraction of identified orthologous genes in the pairwise comparisons performed by those authors range from 52.3% to 68.9%, which is much higher than our findings of between 12% and 42%, shown in the BLAST Matrix of Fig. 2. The difference may be due to the way these percentages were calculated. Whereas we express these as the fraction of gene families found in the reciprocal pan-genome of the pair of analyzed genomes, their calculations are different, and they do not state the cut-off used to recognize orthologous genes as such. In view of their limited reported range, we believe our way of expressing pairwise homology is more useful, as it gives a more sensitive measure. In a subsequent publication, comparative genomics was performed with a larger set of 12 *Lactobacillus* genomes [3]. Inclusion of 7 more genomes reduced their core genome to 141 genes which indicates they used more strict criteria of inclusion than the 50–50 rule we applied. Similar to our analysis, these authors compared the COG classes of the core genes they had identified, and their findings also reported the largest class represented to be genes involved in translation, followed by replication.

Comparative genomics of both *Lactobacillus* and *Bifidobacterium* was presented in a review [30], which mentioned the ability of *Bifidobacterium* to “synthesize at least 19 amino acids and (…) all of the enzymes that are needed for the biosynthesis of pyrimidine and purine nucleotides”. These authors further emphasized the importance of carbohydrate metabolism for *Bifidobacterium* with its capability to degrade complex sugars. Indeed, top-level metabolism genes form a major part of the *Bifidobacterium* core genome (Fig. 9) with class E (amino acid metabolism) as the largest fraction within that category. When we compare this core genome with that of *Lactobacillus* (Fig. 10), our analysis shows that class F genes (nucleotide metabolism) comprise the largest metabolism gene fraction in the *Lactobacillus* core genome.

Ventura and co-workers [30] used a known physiological characteristic (*Bifidobacterium* species are known for their prototrophy) and looked for evidence of this in the genomes. In contrast, we have done a neutral analysis of pan- and core genome COG class representation and then compared this between genera. Our approach reveals novel insights that would remain unnoticed when known phenotypes are taken as a start, for instance the conservation of COG class O genes, involved in post-translational modification and chaperones, in both of these genera.

The authors of a recent review on *Bifidobacterium* genomics [17] pointed out that most *Bifidobacterium*

---

**Table 3 Relative fractions of COG groups within the functionally annotated genes for the six genera**

<table>
<thead>
<tr>
<th>COG groups</th>
<th><em>Bifidobacterium</em></th>
<th><em>Lactobacillus</em></th>
<th><em>Lactococcus</em></th>
<th><em>Leuconostoc</em></th>
<th><em>Enterococcus</em></th>
<th><em>Streptococcus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pan (%)</td>
<td>Core (%)</td>
<td>Pan (%)</td>
<td>Core (%)</td>
<td>Pan (%)</td>
<td>Core (%)</td>
</tr>
<tr>
<td>Information storage</td>
<td>30.0</td>
<td>33.9</td>
<td>34.0</td>
<td>49.1 ↑↑↑</td>
<td>50.5</td>
<td>30.4 ↓↓↓</td>
</tr>
<tr>
<td>Cellular process, signalling</td>
<td>21.9</td>
<td>20.2</td>
<td>22.7</td>
<td>20.3</td>
<td>17.1</td>
<td>19.1</td>
</tr>
<tr>
<td>Metabolism</td>
<td>48.1</td>
<td>45.9</td>
<td>44.3</td>
<td>30.6 ↓↓↓</td>
<td>32.2</td>
<td>50.6 ↑↑↑</td>
</tr>
</tbody>
</table>

All percentages are expressed as the fraction of all COG classes C to V. The arrows indicate significant shifts between the pan-genome genes and core genes for a given genus. Percentages do not always add up to 100% due to rounding effects.
Figure 11  COG statistics for the genes found in the pan-genome (left) and core genome (right) of the collection of genomes from all included organisms, divided into non-pathogenic isolates (probiotic, fermentative and normal human gut flora) at the top and pathogenic isolates at the bottom.

Table 4  Relative fractions of COG groups within the functionally annotated genes for non-pathogens/pathogens. The arrows indicate how the reported percentages increase or decrease in the core genome compared to the pan genome.

<table>
<thead>
<tr>
<th>COG groups</th>
<th>Non-pathogens</th>
<th>Pathogens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pan (%)</td>
<td>Core (%)</td>
</tr>
<tr>
<td>Information storage</td>
<td>33.5</td>
<td>64.4 ↑↑</td>
</tr>
<tr>
<td>Cell. process, signalling</td>
<td>22.0</td>
<td>16.6 ↓</td>
</tr>
<tr>
<td>Metabolism</td>
<td>44.5</td>
<td>20.2 ↓↓</td>
</tr>
</tbody>
</table>
genomes have been sequenced from organisms that have a long history of culture outside their natural habitat, the gut, with the exception of B. longum DJO10A. There is good evidence that the genome of Bifidobacterium is subject to gene reduction to adapt to prolonged culture conditions. This could potentially bias our comparative analysis of Bifidobacterium genomes with that of the other probiotic organisms.

The term ‘lactic acid bacteria’ is commonly used to describe bacteria used as starter cultures and fermentation of foodstuffs. LAB can refer to species from the genera Lactobacillus, Lactococcus, Leuconostoc, Streptococcus, Enterococcus, Pediococcus or all of the Lactobacillales, and sometimes includes Bifidobacterium as well. However, there are good reasons why these bacteria have been placed into different genera and phyla. The analyses presented here support their current taxonomic positions and stress their differences in gene content. The term LAB incorrectly suggests all these organisms are somehow related; a view that is still being presented in the literature [15]. The use of the term LAB is a bit misleading, as the genetic content from these various genera differ significantly. Moreover, some of the genera within LAB comprise only non-pathogenic species (Leuconostoc, Bifidobacterium, Lactobacillus), whereas other genera are a mixture of pathogenic and non-pathogenic species and strains (Streptococcus, Enterococcus). It would be better to refrain from the term LAB as there is no common denominator, other than the production of lactic acid (which is not restricted to these organisms) to collectively describe all species and strains supposedly included in this diverse group of organisms.

An extensive comparative study of Enterococcus genomes could not be identified from the literature. Most studies concentrate on pathogenicity of E. faecalis. Vebo and co-workers [29] compared probiotic and (uro-)pathogenic E. faecalis genomes; however, those comparisons were not based on sequence data. The Enterococcus genomes we have included were mostly from pathogenic organisms (only two non-pathogenic E. faecalis strains whose sequences were nearing completion were publicly available at the time of analysis), which limits the strength of this analysis, as it cannot be used to compare and contrast multiple non-pathogenic with pathogenic Enterococcus genomes. The 11 genomes included represent only 4 species, giving a pan-genome of nearly 8,000 gene families. The first four species of Lactobacillus or Streptococcus genomes in the pan-genome plots of Fig. 1 produce smaller pan-genomes, which could suggest that the diversity of Enterococcus could be at least as extensive as that of Lactobacillus. The pairwise BLAST comparison within this genus resulted in homologues varying from 24% to 84%, again indicating extensive intra-genus diversity.

Streptococcus and Enterococcus are frequently considered as closely related, but the BLAST Matrix comparing all included genomes (Supplementary Fig. S1) does not support this. Instead, somewhat surprisingly, the observed homology between Leuconostoc and Streptococcus genomes is slightly higher than that between Streptococcus and Enterococcus. On the other hand, Lc. lactis was positioned in between these two genera in the tree based on variable gene content. A shared gene pool between these genera can be hypothesized. Based on the conserved core genes, however, Enterococcus is more related to Streptococcus, while Lactococcus is more distinct.

A small comparative study of Streptococcus genomes combined with MLST suggested that S. thermophilus is a relatively young clone, evolved by genome reduction which removed or inactivated Streptococcus virulence genes [13]. It is possible, however, that the reduced genomes observed are the result of prolonged use as starter cultures, as no fresh human isolates have been sequenced to date. In a short review, Delorme [5] states that “S. thermophilus is related to Lactococcus lactis…”. Indeed, from the all-against-all BLAST Matrix, a similarity between 17.3% and 20.2% is recorded between genomes of these two species, which is higher than that between S. thermophilus and any other non-streptococcal genome. However, Lc. lactis also shares 16.0% to 18.0% of reciprocal genes with S. suis, so these overlapping percentages of gene similarity are no indicator of similarity in (probiotic) phenotype. Within the Streptococcus genus, the stated similarity of S. thermophilus with Streptococcus sanguinis (the only member of the viridans group for which a genome sequence is available) is confirmed in our Matrix, but an even higher similarity is found with Streptococcus gordonii.

The COG analysis of the core genomes of separate genera identified both similarities and differences. The three top-level functional COG groups are relatively equally divided over the functionally annotated pan-genomes of all species, but their core genomes differ. Notably, Lactobacillus and Leuconostoc both have a smaller fraction of metabolism core genes than the other four genera and a larger information storage gene fraction. Information storage genes are essential, but redundancy allows so much variation between organisms that they are not all maintained in a core genome of diverse species. In the approach presented here, we first identified the core genomes of groups of bacteria and then sorted the genes in these core genomes for top-level COG categories. As a consequence, genes that were insufficiently conserved based on sequence similarity to be maintained in the core genome are removed despite their possible functional conservation. Using this approach, we found no correlation between the diversity within a genus (using the difference of their pan- and core genome as a measure)
and the fraction of their information/storage COG genes. This lack of correlation is illustrated by the core genome of *Bifidobacterium* (724, or 10% of its pan-genome) and *Leuconostoc* (1,164, or 40% of its pan-genome). These two core genomes contain 34% and 31% information/storage genes, respectively, despite a huge difference in the degree of variation in these two genera.

Of particular interest is the COG analysis where all genomes were divided into a pathogenic and a non-pathogenic group. Virulence genes are not a separate COG category, but from the comparison of the core genomes of the pathogenic group with that of the non-pathogenic group, we can hypothesize that genes belonging to COG categories M (cell wall/membrane biosynthesis) and O (post-translational modification, chaperones) would mostly contribute to virulence. Conversely, it could be assumed that genes highly overrepresented in the core genome of the non-pathogenic group (compared to the core genome of the pathogenic group) most likely contribute to their probiotic or fermentative lifestyle. We observe enrichment for genes belonging to COG class J (translation, ribosomal structure and biogenesis) and again O (post-translational modification and chaperones).

The finding that core genes of the non-pathogenic isolates are more frequently information storage genes and less likely metabolic genes than the core genes of pathogens is counter-intuitive. It is generally accepted that commensals and probiotic strains are most adequately equipped to live in the intestine, which would assume they share a large number of (conserved) metabolic genes to do so. Instead, the reduced metabolism gene fraction in their core genome suggests that there is a large variation within these genes, which reflects the diversity of the various commensals, fermentative and probiotic isolates. The vast enrichment for information/storage genes in the core genome of the non-pathogenic organisms is possibly a reflection of the relative poor conservation of all other functional classes in this group, an effect that appears to be less pronounced in the (ecologically more diverse) pathogenic group. The fact that *Bifidobacterium* are not present in the pathogenic group may have skewed these results slightly. A more accurate prediction for conserved genes with an important role in bacteria with a non-pathogenic lifestyle may become possible in the future, when more non-pathogenic *Enterococcus* genomes become available, which allows comparison of gene content within a genus or even species.

**Conclusions**

This study illustrates the value of comparative genomics of multiple genomes within and between related species and genera. The applied tools are relatively simple to analyze a vast number of genes, and the results can support or contradict existing hypotheses and taxonomic divisions, as well as generate novel hypotheses. We believe the data presented here can assist in understanding the commensal and probiotic relationship of bacteria with their human host. The work presented here demonstrates that the used analyses can be applied to large numbers of genomes, when searching for general mechanisms to predict trends even across genera. The presented analyses can be taken as a test case for comparison of multiple genomes from a largely variable dataset.

**Acknowledgements** The authors are grateful to all research groups that have submitted their genome sequences to public databases, without which this analysis would not have been possible. TMW acknowledges the support provided by the Safety and Environmental Assurance Centre at Unilever for part of this work. OL and DWU received support by the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

**References**

3.2 Paper II. Genome sequencing identifies two nearly unchanged strains of persistent Listeria monocytogenes isolated at two different fish processing plants sampled 6 years apart.
Genome Sequencing Identifies Two Nearly Unchanged Strains of Persistent *Listeria monocytogenes* Isolated at Two Different Fish Processing Plants Sampled 6 Years Apart

Anne Holch, Kristen Webb, Oksana Lukjancenko, David Ussery, Benjamin M. Rosenthal, Lone Gram

*Listeria monocytogenes* is a food-borne human-pathogenic bacterium that can cause infections with a high mortality rate. It has a remarkable ability to persist in food processing facilities. Here we report the genome sequences for two *L. monocytogenes* strains (N53-1 and La111) that were isolated 6 years apart from two different Danish fish processors. Both strains are of serotype 1/2a and belong to a highly persistent DNA subtype (random amplified polymorphic DNA [RAPD] type 9). We demonstrate using *in silico* analyses that both strains belong to the multilocus sequence typing (MLST) type ST121 that has been isolated as a persistent subtype in several European countries. The purpose of this study was to use genome analyses to identify genes or proteins that could contribute to persistence. In a genome comparison, the two persistent strains were extremely similar and collectively differed from the reference lineage II strain, EGD-e. Also, they differed markedly from a lineage I strain (F2365). On the proteome level, the two strains were almost identical, with a predicted protein homology of 99.94%, differing at only 2 proteins. No single-nucleotide polymorphism (SNP) differences were seen between the two strains; in contrast, N53-1 and La111 differed from the EGD-e reference strain by 3,942 and 3,471 SNPs, respectively. We included a persistent *L. monocytogenes* strain from the United States (F6854) in our comparisons. Compared to nonpersistent strains, all three persistent strains were distinguished by two genome deletions: one, of 2,472 bp, typically contains the gene for *inlf* and the other, of 3,017 bp, includes three genes potentially related to bacteriocin production and transport (*lmo2774*, *lmo2775*, and the 3′-terminal part of *lmo2776*). Further studies of highly persistent strains are required to determine if the absence of these genes promotes persistence. While the genome comparison did not point to a clear physiological explanation of the persistent phenotype, the remarkable similarity between the two strains indicates that subtypes with specific traits are selected for in the food processing environment and that particular genetic and physiological factors are responsible for the persistent phenotype.

Received 30 November 2012 Accepted 16 February 2013
Published ahead of print 22 February 2013

Address correspondence to Lone Gram, gram@bio.dtu.dk.
* Present address: Anne Holch, CMC Biologies A/S, Copenhagen, Denmark.
A.H. and K.W. contributed equally to the study.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/AEM.03715-12
Copyright © 2013, American Society for Microbiology. All Rights Reserved.
doi:10.1128/AEM.03715-12
invasive than human clinical strains (13, 16–18). Surprisingly, in a more complex biological model (using oral dosing of pregnant guinea pigs), the strains infected placentas and fetuses just as efficiently as the clinical strains (18). Hence, this particular subtype is of key interest since it is a recurrent contaminant and may be a risk, especially to pregnant women.

The genomes of several strains of *L. monocytogenes* have been sequenced in recent years (9, 19–22). At present, there are 34 *L. monocytogenes* genomes publicly available, of which 16 are finished and 18 are available as draft sequences. This rapid expansion in publicly available genome sequences is key to understanding the evolutionary history of *L. monocytogenes* and to elucidating virulence regulation. Our intent here was to harness genome-based analyses to better understand the basis of this organism’s persistence in particular food processing environments.

In this work, we initially addressed the discriminatory power of subtyping by comparing the genome sequences and predicted proteomes of two strains of *L. monocytogenes* isolated from different plants at different times but which share the same molecular subtype. These two strains were representative of the above-mentioned large group of strains that were isolated repeatedly from fish processing environments over many years and that were indistinguishable by molecular subtyping (8). Subsequently, we searched for features uniquely shared by these and another (previously sequenced) persistent strain in order to identify genes that may contribute to, or detract from, persistence in such environments.

### MATERIALS AND METHODS

**Listeria monocytogenes** strains. Two *L. monocytogenes* strains, representing a highly persistent molecular subtype, were sequenced for this study. Strain La111 was isolated from a package of cold-smoked salmon in 1996 (11), whereas strain N53-1 was isolated from a processing environment in 2002 (8). These isolates derived from different plants. Both strains were determined to be serotype 1/2a and lineage II strains. The strains were deemed identical based on random amplified polymorphic DNA (RAPD), pulsed-field gel electrophoresis (PFGE), and amplified fragment length polymorphism (AFLP) typing and similar to a large cluster of molecular subtypes that are often isolated from Danish fish smokehouses (8). The strains were isolated following a selective enrichment, streaking onto Oxford agar, and restreaking onto brain heart infusion (BHI) agar. Stock cultures were stored at −80°C in a medium containing 4% (wt/vol) glycerol, 2% (wt/vol) skim milk powder, and 3% (wt/vol) tryptone soya broth (TSB) (catalog number CM0129; Oxoid). Growth in the present study was performed with TSB at 37°C.

**DNA purification.** Genomic DNA was purified with a Fast DNA kit (catalog number 116540-400; MP Biomedicals), with modifications. Cells were harvested after growth for 24 h in TSB (catalog number CM0129; Oxoid), and the pellet was resuspended in 210 μl buffer 1 (0.58 M sucrose, 0.01 M Na-P, 10 μg/ml lysozyme). The suspension was heated for 1.5 h at 37°C, followed by washing. The pellet was resuspended in demineralized water, and the procedure for the Fast DNA kit was followed. RNA was removed by using Ambion RNase Cocktail (catalog number AM2286; Invitrogen).

**Genome sequencing.** *L. monocytogenes* N53-1 and La111 were sequenced by using second-generation methods on the Illumina Genome Analyzer II (GAI). Approximately 1 μg of total genomic DNA from each strain was used to generate a short-read library. Library preparation, DNA sequencing, and raw data processing via the Illumina Genome Analyzer Analysis Pipeline were carried out in accordance with the manufacturer’s protocols for single-end 36-bp reads (Illumina, San Diego, CA). The only exceptions involved the random fractionation of the genomic DNA via sonication (rather than nebulization) and the use of 5 μl (rather than 1 μl) of template for the final PCR amplification of the library. The GAI was employed for 36 cycles to generate the nucleotide data. Each strain was sequenced in one lane containing 2 pM template and in a second lane containing 3 pM template.

**Assembly of genomes.** Prior to assembly, sequences were filtered to remove those reads that contained one or more ambiguous base calls. The N53-1 and La111 sequences were assembled separately by using the de novo assembler Velvet version 1.1.04 (23), with parameters determined by Velvet Optimizer 2.1.7 (S. Gladman and T. Seeman). A high-resolution, ordered, and oriented restriction map (optical map) was generated for the N53-1 genome by using the OpGen system (OpGen Technologies, Madison, WI) and the Ncol endonuclease. This physical evidence was subsequently used to constrain genome assembly of N53-1 contigs using Map-solver software (OpGen) based on *in silico* digestion and comparison of restriction cut site patterns of each contig to the genome. The optical map of N53-1 was considered dispositive as evidence in placing contigs generated from the N53-1 isolate. We subsequently explored the applicability of the N53-1 physical evidence for its potential to assist in the assembly of La111, premised on the hypothesis that genomes so similar in sequence content would also share syntenic organization. A minimum score for the local alignment was set initially to 3 and then reduced to 2. Only unambiguous alignments were accepted. For both strains, contigs were concatenated in the order and orientation determined by the optical map alignment. Between each contig, the sequence 5′-NNNNNCAATTCGTCTAC TAATTAATTAATAAGTAATGGNNGNNN-3′ was inserted (24). This sequence was designed such that it introduces a stop codon in all six reading frames as well as a start codon in all reading frames, encouraging proper annotation of those genes residing near contig junctions (24).

**Genome annotation.** The predicted proteomes of all analyzed strains were extracted by using Prodigal software (25), which is able to recognize prokaryotic genes and identify translational initiation sites. tRNA-encoding sequences were located by using the tRNAscan-SE 1.21 server (26). Genome comparisons were made by using Mauve v 2.3.1 (27) and BLAST via the NCBI website.

**Genome sequences from online databases.** The genome of *L. monocytogenes* EGD-e (GenBank accession number NC_003210.1), which is a lineage II, serotype 1/2a strain, was downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/) and used as the reference strain (Table 1). Assembled genomes of *L. monocytogenes* F6854 and *L. monocytogenes* EGD-e were sequenced to generate a short-read library. Library preparation, DNA sequencing, and raw data processing via the Illumina Genome Analyzer Analysis Pipeline were carried out in accordance with the manufacturer’s protocols for single-end 36-bp reads (Illumina, San Diego, CA). The only exceptions involved the random fractionation of the genomic DNA via sonication (rather than nebulization) and the use of 5 μl (rather than 1 μl) of template for the final PCR amplification of the library. The GAI was employed for 36 cycles to generate the nucleotide data. Each strain was sequenced in one lane containing 2 pM template and in a second lane containing 3 pM template.

**TABLE 1. L. monocytogenes** strains used in the present study

<table>
<thead>
<tr>
<th>Strain</th>
<th>Serotype</th>
<th>Lineage</th>
<th>Source of isolation</th>
<th>Reference for isolation</th>
<th>Reference or source(s) for nucleotide database</th>
<th>Reference for genome sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>N53-1</td>
<td>1/2a</td>
<td>II</td>
<td>Smokehouse environment</td>
<td>8</td>
<td>This study</td>
<td>This study</td>
</tr>
<tr>
<td>La111</td>
<td>1/2a</td>
<td>II</td>
<td>Cold-smoked salmon</td>
<td>10</td>
<td>This study</td>
<td>This study</td>
</tr>
<tr>
<td>F6854</td>
<td>1/2a</td>
<td>II</td>
<td>Turkey franks</td>
<td>28</td>
<td>J. Craig Venter Institute, TraceDB</td>
<td>29</td>
</tr>
<tr>
<td>EGD-e</td>
<td>1/2a</td>
<td>II</td>
<td>Rabbit isolate, 1926</td>
<td>30</td>
<td>NCBI</td>
<td>30</td>
</tr>
<tr>
<td>F2365</td>
<td>4b</td>
<td>I</td>
<td>Mexican-style cheese</td>
<td>31</td>
<td>J. Craig Venter Institute</td>
<td>29</td>
</tr>
</tbody>
</table>

* Raw data.

* L. monocytogenes N53-1 and La111 have been sequenced in the present study, whereas DNA sequences from F6854, EGD-e, and F2365 were retrieved from online databases.
F2365 were downloaded from the J. Craig Venter Institute website (http://www.jcvi.org/). F6854 belongs to the same ribotype (DUP-1053A) as two other strains isolated 12 years later and linked to the same food processing facility (8) and is, hence, a highly persistent subtype. The raw sequence data of L. monocytogenes F6854 from TraceDB (ftp://ftp.ncbi.nlm.nih.gov/pub/TraceDB/) was included in the data set for the single-nucleotide polymorphism (SNP) analysis.

BLAST Ring Image Generator. Visual comparison of genome homology was done by using BRIG (BLAST Ring Image Generator) (32; http://sourceforge.net/projects/brig/). BRIG is capable of generating circular comparison images for prokaryote genomes and displays similarity between a reference genome in the center and other query sequences. EGD-e was used as the reference genome and was compared to the genomes of N53-1, La111, F6854, and F2365. As the similarity is calculated from the respective reference, regions that are absent from the reference genome but present in one or more of the query sequences will not be displayed. The BRIG method uses the software BLASTALL v 2.2.25+ for the searches. The comparisons were done with default settings.

BLAT and BLAT matrices. The similarity between N53-1 and La111, and the similarity to the other strains of L. monocytogenes, was also assessed by a pairwise genome comparison. A matrix showing the fraction of genome-specific genes was constructed. For each gene in one genome, a BLAST-Like Alignment (BLAT) was performed against the second genome. BLAT rapidly searches for relatively short k-mers and extends these to high-scoring pairs (HSPs) (33). A given gene was considered to be specific if there were no HSPs satisfying the 50/50 rule, meaning that no sequence in the queried genome was at least 50% identical to the gene over at least 50% of its length.

SNP analysis. For SNP detection, the raw data sequences from N53-1, La111, and F6854 were mapped to the reference strain EGD-e. N53-1 and La111 were mapped to both F6854 and F2365. Also, raw data sequences from N53-1 were mapped to the de novo-assembled La111 genome, and the raw data sequences of La111 were mapped to the de novo-assembled N53-1 genome. After mapping the raw data, open reading frames were identified, and the read mappings were analyzed for the presence of SNPs. All steps of the SNP analysis were conducted by using CLC Genomics Workbench v 4.8 (CLC, Aarhus, Denmark) with the default settings, except for the minimum variant frequency, which was set at 85%. A list of the identified SNPs was exported to an Excel spreadsheet. All SNPs coding for silent mutations were deleted, and further analysis was conducted with the remaining nonsynonymous SNPs.

In silico MLST analysis. Multilocus sequence typing (MLST) was used to analyze nucleotide variations in seven housekeeping genes (acet, bgkA, cat, dapE, dat, idh, and lhpA). An in silico PCR analysis was conducted on the N53-1 and La111 genomes by using CLC DNA Workbench v 6.5 with default settings. The obtained in silico PCR products were trimmed and uploaded to the L. monocytogenes MLST database (http://www.pasteur.fr/recherche/genopole/PPB2/mlst/Lmono.html) for determination of the sequence type (ST).

Nucleotide sequence accession numbers. Genome sequences have been submitted to the EMBL database at the EBI website and can be accessed under accession numbers HE999704 (strain La111) and HE999705 (strain N53-1).

RESULTS AND DISCUSSION

General genome features. The next-generation sequencing of L. monocytogenes N53-1 generated over 70.8 million reads, of which 69.3 million reads were retained after removing those containing ambiguous base calls within their sequence. The N53-1 reads assembled into 314 contigs (N50 [a statistic measuring assembly quality] = 100,675). For La111, over 57 million reads were generated, with 54.8 million reads subsequently analyzed after removing sequences containing ambiguous base calls. De novo assembly of the La111 short reads formed 279 contigs (N50 = 106,240).

By using MapSolver software, the in silico digestions of the de novo N53-1 assembled contigs were compared to the optical map. In total, 25 contigs were placed, representing 82.3% of the sequence data generated for N53-1 (assembled length excluding gaps divided by total length of all de novo-assembled contigs) (Table 2). Using BLAST, we found that of the remaining large contigs (>30 kb), two unplaced contigs aligned well to other published L. monocytogenes nuclear genomes, and one aligned to the plasmid sequence of L. monocytogenes 08-5578. Of the 279 de novo-assembled contigs of La111, 19 aligned to the optical map of N53-1 under the strict default parameters representing 78.5% of the genome (assembled length excluding gaps divided by total length of all de novo-assembled contigs) (Table 2). An alignment by using BLAST revealed that five of the six unmapped, large contigs ranging in size from 34.9 kb to 54.6 kb aligned closely with the other published L. monocytogenes 08-5578, 08-5923, and/or EGD-e, and one contig (37.7 kb) aligned to the plasmid sequence from L. monocytogenes 08-5578. A second BLAST alignment showed that four of the six large contigs showed a very high level of similarity (>99%) to the assembled N53-1 genome and, as such, were added to the La111 alignment based on this similarity. The final La111 assembly consisted of 34 contigs representing 84% of the sequence data generated.

Excluding the inserted gap sequences (24), the N53-1 genome assembly was 2,553,709 bp in length, while La111 totaled 2,534,555 bp, and both strains had a G+C content of 37.9% (Table 2). These genome sizes are similar to the sizes of other sequenced L. monocytogenes genomes, which have been estimated to be between 2.87 Mb (L. monocytogenes Finland 1988 [GenBank accession number CP002004.1]) and 3.02 Mb (L. monocytogenes Scott A [GenBank accession number AFG100000000.1]). Ninety-four and 86 tRNAs were predicted within the N53-1 and La111 genome sequences, respectively (Table 2). Using Prodigal for the protein BLAST matrix, N53-1 was predicted to have 3,323 proteins, and La111 was predicted to have 3,302 proteins (Table 2). The differences between the two strains likely derive from missing data in the La111 assembly. Differences in the number of predicted proteins and predicted tRNAs were observed when using
different programs. These differences are due to different algorithms and cutoff values used in the different programs.

**Comparative genomics.** N53-1 and La111 are very similar based on DNA subtyping (8), virulence gene sequencing (16), and phenotypic behavior (13, 16, 17). However, a whole-genome comparison of these two strains had not yet been attempted. Strains that are persistent might share genetic features that are not present in nonpersistent strains. This could include the presence or absence of entire genes, SNPs, or different patterns of gene expressions relative to presumably nonpersistent strains.

Conservation and variation in gene content between genomes were visualized by BRIG. The two newly sequenced genomes of N53-1 and La111 and the two downloaded genomes (F6854 and F2365) were included in the comparison, and EGD-e was used as a reference (Fig. 1). It should be noted that the F6854, N53-1, and La111 genomes are draft genomes and are not completely closed. Therefore, regions that are not included in the BRIG alignment most likely represent regions not sequenced in one or more genomes, deletions/insertions, or genome fragments replaced by a nonhomologous sequence.

A gap of 2,472 bp occurred in all three persistent strains (N53-1, La111, and F6854) relative to EGD-e and F2365, beginning at bp position 429629 and containing the \textit{inlF} gene in F2365 (Fig. 1). \textit{InlF} is a surface-anchored protein with unknown function; however, it plays a role in increased infection of L25 murine fibroblast cells (35) and is present in a large number of strains. Jia et al. (36) did not find any \textit{inlF}-specific PCR products in lineage I strains, and Tsai et al. (37) found \textit{inlF} in all tested lineage II strains and not in lineage I strains using gene sequencing. Doumith et al. (38) reported \textit{inlF} in a least two-thirds of both lineage I and lineage II strains using a DNA microarray. Further studies of strains from highly persistent subtypes are required to determine if the absence of \textit{inlF} promotes persistence.

A stretch of DNA of 3,017 bp was absent in N53-1, La111, and F6854 but present in EGD-e (at bp position 2857618) and F2365. The area covers \textit{imo2774}, \textit{imo2775}, and the 3'-terminal part of \textit{imo2776}. \textit{imo2774} encodes a homologue of a putative bacteriocin export ABC transporter, \textit{imo2775} a homologue of a bacteriocin-associated integral membrane protein, and \textit{imo2776} a homologue of lactococcin_972. The genes encoding these proteins are not well described, and no further information is available.

**FIG 1** Circular map of \textit{L. monocytogenes} N53-1, La111, F6854, and F2365 using EGD-e as a reference genome. The inner ring denotes the reference EGD-e genome with corresponding genetic coordinates. The next four rings denote the coding regions for the four queried strains, F6854 (blue), F2365 (purple), N53-1 (red), and La111 (green).
At bp position 2360713 in EGD-e, a large sequence of approximately 40,000 bp is not present in N53-1, La111, or F2365, whereas it is present in EGD-e and F6854. In F6854, the sequence has been identified as comK (major competence transcription factor). A prophage was previously shown to be inserted into comK in F6854 at this position (9, 39). Oriş et al. (9) used whole-genome sequence comparison to analyze four strains from the same processing plant: a food and outbreak pair from 1988 and a food and outbreak pair from 2000. These four strains differed by only 11 SNPs in the backbone sequence (excluding comK and the Thr-4 prophage) by an interstrain comparison. In all four sequenced strains (9), comK contained a prophage insertion of approximately 40,000 bp. In spite of the near uniformity of the backbone sequences, the prophage insert contained 1,274 SNPs that differentiated the pair from 1988 from the pair from 2000.

Recently, it was found that the presence of a prophage in comK could be a marker for rapid niche-specific adaptation, biofilm formation, and persistence (39); however, the two processing-persistent strains used in the present study may lack an intact prophage insertion in comK (gap of around 40 kb in N53-1, La111, and F2365) (Fig. 1). We searched the La111 and N53-1 draft genomes for intact prophages using software described previously by Bohlin et al. (40) and found none. However, as our genome assemblies contain gaps representing regions where assembly of sequence data was not achieved, it is difficult to determine whether the full-length 42-kbp prophage is inserted into the comK gene within these two Listeria strains. We explored the possibility that the prophage is not present as one contiguous piece in our assemblies. Using nucleotide BLAST, portions (approximately 0.9 kbp) of the 28.5-kb comK prophage sequence from F6854 aligned well to the La111 and N53-1 assembled contigs. The most significant alignments occurred in the same area of the scaffold, and some of the alignments ended because of a gap in the sequences. Using MAQ (Mapping and Assembly with Qualities) (http://maq.sourceforge.net/), we found significant alignment of the raw sequence data from both strains across approximately 50% of the comK prophage reference sequence. Hence, there is strong evidence that at least a portion of a prophage is present in the La111 and N53-1 draft genomes. However, we are unsure as to whether the prophage, in its entirety, persists. This may be attributed to limitations in the assembly of repetitive regions and/or the inability to map reads that differ by more than 2 bases (a parameter of MAQ). Alternatively, the results may represent a relic of a previous phage insertion and subsequent deletion event. If the two Listeria strains do contain a prophage in comK, it could potentially be involved in the persistence mechanism (39).

At bp position 473841 in EGD-e, there is a gap of 7,500 bp in N53-1 and La111, whereas the gap size in F6854 and F2365 is 8,625 bp. The genes present in this region in EGD-e (lmo0444, lmo0445, lmo0446 [pva], lmo0447 [gadD1], and lmo0448 [gadT1]), designated stress survival islet 1 (SSI-1), are responsible for growth at low pH and at high salt concentrations and the ability to survive and grow in model food systems (41). The size of the gap is larger in F6854 and F2365, as the islet in those strains contains only one gene (LMOF2365_0481 homologue), whereas the islet in N53-1 and La111 contains genes homologous to lmo0446 and lmo0445. A more detailed description of SSI-1 is presented below.

Comparative proteomics. The gene content of strains was compared in a BLAT matrix (Fig. 2). It displays the frequency of genes found in the “row” genome that are not also found in the “column” genome, as a proportion of the total number of genes in the row genome. Strains N53-1 and La111 are extremely similar, with only 2 (0.06%) of the predicted proteins in N53-1 not present in La111. In contrast, 144 and 143 (3%) of the predicted proteins in EGD-e were not present in N53-1 and La111, respectively. The genomes of both N53-1 and La111 are not fully sequenced, which could explain the missing predicted proteins in these two strains compared to EGD-e.

Of two predicted proteins present in N53-1 but absent in La111, one with unknown function has NACHT and WD repeat domain-containing protein 1. The WD40 domain is found in a number of eukaryotic proteins that cover a wide variety of functions, including adaptor/regulatory modules of signal transduction, pre-mRNA processing, and cytoskeleton assembly (http://www.ncbi.nlm.nih.gov/protein/308736994). An uncharacterized protein, YdeI, is the only predicted protein present in N53-1 and absent from EGD-e, while the glutamate synthase (NADPH) large chain and glycine betaine/carnitine/choline transport ATP-binding protein OpuCA are present in La111 but absent from EGD-e. As none of these proteins are present in both N53-1 and La111, none can independently suffice as a cause for persistence each of these strains. When turning to the 5% of predicted proteins that are unique to EGD-e, they cover a broad range of protein functions (see Table S1 in the supplemental material).

Single-nucleotide analysis. Although some SNPs may derive from sequencing errors, true SNPs may be either silent or nonsynonymous, in which case they may change the function of the transcribed protein or result in a truncated protein. An example of the latter possibility entailed the lialisterial surface protein InlA, where a nucleotide substitution from C to T results in a stop codon and production of a truncated protein (16, 42, 43). In the present study, SNP analyses assessing only nonsynonymous changes were carried out after mapping raw reads of the queried strain against a reference strain (Fig. 3). First, the three persistent strains (N53-1, La111, and F6854) were mapped against EGD-e, and between 3,471 and 5,037 SNPs were detected (Fig. 3A). Among these, 1,980 SNPs were shared between the three strains.

The numbers of SNPs detected between F6854 and our two newly sequenced strains were 3,829 and 3,819 for N53-1 and La111, respectively (Fig. 3B). All three strains belong to serotype 4b, compared to EGD-e, a serotype 4b strain, identified 5,848 and 5,840 SNPs, respectively (Fig. 3C).

In contrast, testing of N53-1 and La111 against each other identified no SNPs when using N53-1 as the reference; using La111 as the reference suggested only 18 SNPs, substantiating an extraordinarily close relationship between the two strains in spite of the 6-year interval separating their dates of isolation. The complete lack of SNPs between strains N53-1 and La111, despite being isolated 6 years apart from two different factories, may indicate that this genome type is especially well adapted to persisting in this environment.

In silico MLST analysis. Seven in silico PCR products of between 458 and 702 bp were obtained from N53-1 and La111. After trimming, the sequences were uploaded to the L. monocytogenes MLST database, and both were identified as belonging to ST121: abcZ-7, bgla-6, cat-8, dapE-8, dat-6, ldh-37, and lihkA-1. F6854 belongs to ST11, which corresponds to abcZ-7, bgla-6, cat-10, dapE-6, dat-1, ldh-2, and lihkA-1, and EGD-e belongs to ST35 (abcZ-6, bgla-5, cat-6, dapE-20, dat-1, ldh-4, and lihkA-1) (44). A
recent study by Hein et al. (45) described ST121 strains isolated in Austria and Belgium from different ecological niches, including food, food processing facilities, and human cases, over several years. Two of the strains were isolated from the same dairy plant over a course of at least 3 years. 

$L.\ mono\text{cytogenes}$ ST121 strains have also been reported in France, Italy, and Spain (34, 46, 47). By PCR, Hein et al. (45) showed that the ST121 strain had a 2.2-kbp fragment (in N53-1 and La111), whereas the majority of serotype 1/2a strains had a 9.7-kbp fragment. The 9.7-kbp fragment is described as a five-gene stress survival islet (SSI-1) and contributes to growth under suboptimal conditions (41). A BLAST search of the 2.2-kbp fragment showed 95% identity with the two genes $\text{lin0464}$ and $\text{lin0465}$ from $\text{Listeria innocua}$ CLIP 11262 (GenBank accession number AL596165.1). Hein et al. (45) speculated that the two $\text{L. innocua}$ genes $\text{lin0464}$ and $\text{lin0465}$ both contribute to fitness of the ST121 strains in the environment. Furthermore, the ST121 strains also had the same premature stop codon in $\text{inlA}$, leading to a truncated InlA, as in our two processing-persistent strains. Altogether, we can conclude that the ST121 strains described in a variety of studies are identical to our two processing-persistent strains, whose genomes we have now sequenced. Evidence that this group of strains persists in the processing environment is mounting, and the basis for this attribute warrants investigation.

**Conclusions.** Several studies have reported the ability of particular molecular subtypes of $\text{L. monocytogenes}$ to persist in food processing plants (7–9), where they constitute a recurrent source of product contamination. In the Danish fish processing industry, strains belonging to one particular subtype of $\text{L. monocytogenes}$ have been isolated over several years in different processing plants. Strains of this subtype were isolated from four out of eight different processing plants and were the persistent and dominant type in three plants over a period of 6 years (8, 11). These data indicate that certain subtypes of $\text{L. monocytogenes}$ may be specifically adapted to processing plant environments and are able to persist over long periods of time. However, our data do not allow us to conclude on the underlying ecology and evolution. Thus, we cannot say if a particular subtype at random enters the processing environments and, due to low growth rates, remains unchanged for years or if the conditions in the environment select for particular mutational changes over time. The use of genome sequencing of strains isolated repeatedly from a plant over longer

![BLAT matrix of pairwise genome analysis between $\text{L. monocytogenes}$ strains N53-1, La111, EGD-e, F2365, and F6854.](image)

Each row represents the specific genome of one genome compared to another, while the diagonal shows comparison to itself. In the matrix cells, the numbers of nonshared protein-coding genes are given both as a number and as a percentage, based on the ratio of the specific genome and total number of predicted genes of the query genome, as indicated. The cells in the matrix are colored darker as the fraction of similarity decreases.

**FIG 2** BLAT matrix of pairwise genome analysis between $\text{L. monocytogenes}$ strains N53-1, La111, EGD-e, F2365, and F6854. Each row represents the specific genome of one genome compared to another, while the diagonal shows comparison to itself. In the matrix cells, the numbers of nonshared protein-coding genes are given both as a number and as a percentage, based on the ratio of the specific genome and total number of predicted genes of the query genome, as indicated. The cells in the matrix are colored darker as the fraction of similarity decreases.
periods of time could potentially unravel this. Such approaches have recently been used to analyze the changes in persistent food processing environments and that particular genetic and physiological factors are responsible for the persistent phenotype.

ACKNOWLEDGMENTS

Anne Holch was supported by the Danish Research Council for Technology and Production Sciences (project 274-08-042). Kristen Webb was supported by the ARS Research Associate Program.

We thank Alicia Beavers, Tony Capuco, Chris Clover, Detiger Dunams, Monica Santin-Duran, Garrett Gleeson, Steve Schroeder, and Tad Sonstegard for support toward the completion of the project. We thank Jon Bohlin for help with prophage analyses.

REFERENCES


Chapter 4

HMM-based comparative genomics

Profile-based methods provide an alternative way for sequence similarity search and whole genome comparison. Such algorithms as PSI-BLAST and HMMER, suggest statistically significant similarity between homologous sequences and are generally more sensitive than simple pairwise homology search. This chapter introduces PanFunPro - a new approach for pan-genome analysis (Paper III), and includes three examples of its application.

One of the main questions in comparative genomics is the number of universally conserved genes, which can be found in all prokaryotic genomes. In 2010, a study by Lagesen et al. showed that there is not even a single protein conserved in the set of 1000 prokaryotic genomes, using BLAST-based comparison. Paper IV demonstrates the comparison of 2110 bacterial and archaeal genomes using PanFunPro approach, with the purpose to re-examine the core set of proteins found within analysed set of genomes. The results suggest a minimal genome of perhaps about 100 conserved functional domains and provides the functional annotation of the conserved proteins.

Paper V illustrates the analysis of chromosome-specific families in Vibrio
genomes. Whole genome comparison included chromosome-specific genome estimation within and a mixture of complete and draft genome sequences. Resulting specific proteins families were searched for available Gene Ontology information in order to access functional categories and possible processes that differ between two chromosomes.
4.1 Paper III. (Manuscript). PanFunPro: Pan-genome analysis based on Functional Profiles
PanFunPro: PAN-genome Analysis Based on FUNctional PROfiles

Oksana Lukjancenko∗1, Martin Christen Frølund Thomsen1, Mette Voldby Larsen1 and David Wayne Ussery1,2

1Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
2Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Abstract
PanFunPro is a tool for pan-genome analysis that integrates functional domains from three HMM collections, and uses this information to group homologous proteins into families based on functional domain content. We use PanFunPro to compare a set of Lactobacillus and Streptococcus genomes. The example demonstrates that this method can provide analysis of differences and similarities in protein content within user-defined sets of genomes. PanFunPro can find various applications in comparative genomic study, starting with the basic comparison of newly sequenced isolates to already existing strains, estimation of shared and specific genomic content; and furthermore, it can be potentially used in determination of target sequences for in silico bacterial identification, and epidemiological studies.

Introduction

Whole genome sequencing continues to become faster and less expensive with time; currently there are more than 2000 complete microbial genomes that are publically accessible, and the number of sequences is still growing exponentially. Availability of numerous strains from the same species led to the development of new analyses, such as the bacterial species pan-genome (1). Pan-genomic studies aim to determine differences in protein content between organisms and characterize the complete genomic repertoire of certain taxonomic group. Therefore, comparative genomics is the first fundamental step in pan-genome analysis.

Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor through a speciation event, or a duplication event. As a result, comparative genomics usually starts with a sequence similarity search using standard approaches, such as local alignment search (BLAST (2), FASTA (3)); orthology detection and clustering (CD-HIT (4), OrthoMCL (5), Inparanoid (6)); or search tools based on Hidden Markov Models (HMM) (7). The comparison of homologous sequences and analysis of their phylogenetic relationships has important implications in understanding evolutionary processes and provide very useful information regarding the structure and function of proteins (8).

Here we present a tool for pan-genome analysis. It is stand-alone tool providing several functionalities – homology detection and genome annotation by three HMM-collections, pan-/core genome

∗Corresponding author, e-mail: oksana@cbs.dtu.dk
calculation within a set of proteomes, pairwise pan-/core-genome analysis, specific genome estimation for different sets of genomes as well as pairwise analysis of specific proteomes, basic statistics for the output proteins from the pan-/core-/specific-genome calculation, and finally analysis of available Gene Ontology (GO) information for the output proteins from the pan-/core-/specific-genome calculation.

**Design and Implementation**

**Approach overview**

There are four basic steps in the PanFunPro approach, as shown in Figure 1: (1) genome selection; (2) functional domain collection; (3) construction of functional profiles and and protein grouping; (4) and finally, analysis of the pan, core and accessory genomes.

Figure 1: Schematic of PanFunPro approach. Method includes four basic steps: (1) genome selection; (2) functional domain collection; (3) construction of functional profiles and and protein grouping; (4) and finally, analysis of the pan, core and accessory genomes. Blue colour explains profile construction steps, while green colour indicates possible types of analysis.
(1) Genome selection
The PanFunPro programme first imports a list of genomes, selected for analysis. Each genome is represented by a FASTA file of amino acid sequence for all the encoded proteins. In the case of DNA sequences with no annotated genes, prediction of open-reading frames (ORFs) from the DNA sequence of the genome is carried out using Prodigal software (9).

(2) Acquiring the functional domains
To form a set of functional profiles for each genome, all proteins are scanned against three collections of HMMs: PfamA (10), TIGRFAM (11), and Superfamily (12) using InterProScan software (13).

(3) Construction of functional profiles and proteins grouping
Briefly, the functional profile or architecture is a combination of non-overlapping functional domains (HMMs) found in a particular protein. Only HMM hits with an E-value below 0.001 are considered significant and are used to create functional architectures. Furthermore, domains of only one database at a time are considered, meaning that if the protein has any matches in PfamA database, the hits in TIGRFAM and Superfamily databases are not considered. However, if the scan against the PfamA database does not result in any hit, analogously TIGRFAM and Superfamily databases are checked. HMM collections are searched in the following order: PfamA, TIGRFAM, and then Superfamily. For each protein the functional profile name is created based on alphabetically sorted non-repeating accession numbers of all non-overlapping domains found in the protein sequence. Multiple proteins can belong to a single protein family if they share the same functional architecture, resulting in a lower number of families per genome than the reported number of proteins. Sequences with no significant matches to any searched HMM-database are collected from each of analysed genomes and clustered using the CD-HIT tool (4). Clustering is implemented with a five amino acid window search, allowing two proteins to be in the same protein family if similarity between sequences is at least 60%. Resulting clusters are considered to be protein families, where the profile name is prefixed with ‘CL’ (stands for clustering) and followed by cluster identification number. Later, HMM-based and clustering-based protein families for each genome are joined together to form a whole genome profile collection.

(4) Analysis
Analysis part includes description of possible ways of result acquisition and visualization.

Core- and pan-genome calculation
The pan-genome is defined as the complete collection of all proteins found in a set of genomes (1); in our case, this is represented by the collection of all unique functional profiles found in those genomes. Starting with the first genome, as more genomes are added, an accumulative pan-genome is constructed and the resulting pan-genome number increases with the addition of more genomes. Similarly, the core-genome is the collection of conserved proteins (functional profiles) that are conserved across the analysed genomes, and the size of the core genome decreases as more genomes are added. Conservation data are stored as table and can be visualized in an accumulative pan-/core-genome plot. Additionally, lists of profiles, comprising pan- and core-genomes, can be assesses as a...
Pairwise comparison between genome is visualized as a triangle-shaped ‘matrix’, showing the number of protein families that are shared between two proteomes, both as percentage and absolute number; as well as the total amount of protein families found in both genomes. When a strain is compared to itself, the fraction of protein families with more than one member is provided. The blue colour gradient indicates homology between different genomes, and the red triangles at the bottom of the figure represent homology within a genome (e.g., duplicate proteins).

### Accessory genome analysis

Differences between proteomes can be assessed by identification of accessory profiles. The accessory genome includes proteins that are present in several, but not all analysed genomes; or are specific to particular genome or group of genomes. A protein is considered to be ‘specific’ if the functional profile is present in the query set of genomes and is absent in subject set of organisms. Estimation of accessory or specific genomes requires two sets of organisms and can follow four assumptions: (1) proteins, present in core-genome of first set of genomes, and absent in the core-genome of the second set of genomes; (2) proteins, present in pan-genome of first set of genomes, and absent in the core-genome of the second set of genomes; (3) proteins, present in core-genome of first set of genomes, and absent in the pan-genome of the second set of genomes; (4) and proteins, present in pan-genome of first set of genomes, and absent in the pan-genome of second set of genomes. Options (1) and (2) introduce specific-core-genome, while options (3) and (4) – specific-pan-genome. Given that the first and the second sets of genomes are the same, application of options (3) and (4) will yield in accessory genome of input set of genomes.

Pairwise analysis of specific content can be visualized as a square-shaped matrix, where each row represents the specific genome of one organism compared to another, while the diagonal shows the comparison to itself. In the matrix cells, the amount of non-shared sequences is provided as a ratio of specific genome to a total number of proteins in the query strain. When compared to itself result is 0. The colour code indicates the level of similarity.

### Basic statistics and Gene Ontology analysis

For a given collection of genomes, the set of core, pan, and accessory proteins is calculated, and the share of PfamA-, TIGRFAM-, Superfamily-, and CD-HIT-based profiles, as well as protein length distribution are visualized using R ggplot2 package and can be assessed as a table.

In addition, available GO (14) information can be extracted. Interproscan tool provides possible GO identification numbers (GO ID) for each domain in the profile. Consequent GO IDs for each of the profiles are searched for GO term description and grouped by more common functional category using map2slim tool, part of GO::Parser module. Results are visualized using R package ggplot2.

### Results

**The case study**

The PanFunPro approach was tested on genomes of Lactobacillus and *Streptococcus* genera, previously used in comparative genomics study by Lukjancenko et al. (15), further mentioned as BLAST-
Based study. All *Lactobacillus* genomes were probiotic, whereas *Streptococcus* strains contained both pathogenic and probiotic species.

Here we focus on the types of results PanFunPro (further mentioned as PanFunPro-based analysis) can generate: a pan-/core-genome plot; a pairwise pan-/core-genome matrix; a pairwise specific-genome matrix; distribution of database source by which protein was annotated; and finally, distribution of predicted GO terms among profiles.

**Pan- and core-genome overview**

Accumulative pan- and core-genome were calculated for both example genera and are shown in Figure 2. *Lactobacillus* genus resulted in a total of 467 core and 7009 pan gene families (Figure 2A). Most of the shared architectures consisted of PfamA domains and for 73% of them GO terms were available (Figure S1.A), whereas only 37% of pan-genome gene families were HMM-based profiles and barely half of them had Gene Ontology information available (Figure S1.B). Analysis of GO IDs distribution among the 3 general functional groups: biological process, molecular function, and cellular component, resulted in 239, 176 and 26 GOs, respectively, in the core-genome; and 470, 418 and 60 GOs, respectively, in the pan-genome.

Similar analysis, done for genomes of the *Streptococcus* genus, yielded in 576 shared functional profiles and a total amount of 6263 architectures found within the genus (Figure 2B). Similarly to the *Lactobacillus* results, core-genome profiles consisted of PfamA domains and 72% of them contained pathway information (Figure S2.A), whereas only 23% pan-genome profiles were based on HMM-domains and for more than half of them pathway information was accessible (Figure S2.B). Analysis of GO IDs distribution among the 3 general functional groups: biological process, molecular function, and cellular component, resulted in 269, 211 and 36 GOs, respectively, in the core-genome; and 492, 434 and 56 GOs, respectively, in the pan-genome.

Pairwise pan- and core comparison of strains within the *Lactobacillus* genus showed that pairs of genomes from different species share 30-60% of the protein families (profiles), while 70-90% are shared within the same species (Figure 3). Homology estimation within single proteomes revealed that approximately 20% of protein families in each genome had more than 1 member. Comparison of core- and pan-genome analyses, performed by BLAST-based and PanFunPro-based approaches, found that typically HMM-based grouping of homologous sequences is more sensitive, and result in significantly reduced number of pan-genome families, 7,009 compared to 13,069 for *Lactobacillus* genus, and 6,263 compared to 9,785 in *Streptococcus* genus. Furthermore, the amount of shared profiles increased for *Lactobacillus* genus (363 to 467); however the core of *Streptococcus* genus did not follow the expansion tendency, and yielded in 576 compared to 638 profiles.

**Specific genome overview**

*Streptococcus* genomes were used as an example of accessory genome analysis. The genus contains twelve species for which complete sequenced genomes are available; *S. thermophilus* is used in making yoghurt, and considered probiotic, while other strains are pathogenic. Single representatives of each pathogenic species and all probiotic genomes were selected for specific genome analysis. Proteomes were compared in pairs to estimate the fraction of specific profiles, which is present in one genome and absent in another. The resulting overview is visualized in Figure 4. On average each proteome contained 30-40% specific profiles compared to other species and 6-20% within the non-pathogenic species.
Figure 2: Pan- and core-genome plot. A. Analysis performed on *Lactobacillus* genomes. B. Analysis performed on *Streptococcus* genomes.
Figure 3: Pairwise pan- and core-genome comparison of strains within the *Lactobacillus* genus.
Further, proteomes from pathogenic genomes were compared to non-pathogenic proteomes. Profiles, conserved in each pathogenic strain and absent in probiotic *Streptococcus* genomes, were considered to form specific core profiles. Specific-core-genome estimation resulted in 23 functional architectures formed from PfamA domains (Figure 5A), 14 of them contained Gene Ontology information. Each protein could serve multiple functions, though more than one GO ID was available. The classification of proteins into three common gene ontology groups, as well as less broad term groups, are shown in Figure 5B. Specific core protein families were involved in metabolic processes, transport, signal transduction, and various binding and enzyme activity. Similar analysis of specific pan-genome for pathogenic *Streptococcus* strains yielded in 4,603 profiles, 31% of which were based on HMM-domains and 703 contained pathway information (Figure S3). An overview of the GO functional groups reveals a broader collection of processes that proteins of pathogenic strains are involved in; however, they are not shared among all the *Streptococcus* pathogens and are most likely to be species-specific. The BLAST-based analysis included pathogenic strains from other genera, and thus cannot be comparable.

**Performance**

The PanFunPro method was designed to integrate the information of functional domains from three HMM-based databases and group proteins into families according to the domain content within the protein, and then to further analyze differences and similarities within defined groups of genomes based on functional architectures and visualize them. The approach includes a complex construction and assignment of functional profiles step. Therefore, we have measured the time required to collect functional domain information and perform profile formation for a set of 21 *Lactobacillus* genomes (15). The test was performed both on MacBookPro, 2.4 GHz Intel Core i5, 8GB 1067 MHz DDR3; and on a Cluster with x86_64 architecture using 1 processor per genome and the default InterProScan settings. As illustrated in Table 1, single genome annotation by the PanFunPro approach takes about...
Figure 5: Protein architecture and available GO functional categories distribution within specific core-genomes of pathogenic *Streptococcus* strains. A. Specific core-genome profile distribution. B. Specific core-genome GO functional categories distribution.
25 and 14 min, on a laptop and cluster, respectively. To prepare profiles for the whole genus of 21 genomes, scanning one genome at a time, took more than 8h on MacBookPro and approximately 5h on the cluster. However if we allow scanning of genomes to run simultaneously on the cluster, the pan-genome calculation takes less than an hour.

Table 1: PanFunPro profile construction performance.

<table>
<thead>
<tr>
<th></th>
<th>MacBookPro</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 genome (1 genome per scan)</td>
<td>25 min 52 sec</td>
<td>14 min 8 sec</td>
</tr>
<tr>
<td>21 genome (1 genome per scan)</td>
<td>8h 52 min 10 sec</td>
<td>5h 2 min 43 sec</td>
</tr>
<tr>
<td>21 genome (21 genome per scan)</td>
<td>NA</td>
<td>21 min 33 sec</td>
</tr>
</tbody>
</table>

Availability and Future Directions

The source code for PanFunPro is developed in the Perl programming language for UNIX systems, and requires access to the following programs: BioPerl, GO Parser, HMMER packages, R program, Interproscan, Oracle/Sun Java 1.6, CD-HIT clustering tool. Software and instructions are available via http://www.cbs.dtu.dk/~oksana/PhD_Thesis/PanFunPro/

PanFunPro has been also implemented as a web server (http://cge.cbs.dtu.dk/services/PanFunPro/). The user can select a set of genomes from the provided database, including 1982 Bacterial and 128 Archaeal strains; or can upload genome sequence and compare it to the genomes listed in the database (optional). The input file can be uploaded either in Genbank/FASTA format, or can already contain predicted proteins. Web server provides 6 analysis possibilities: core-, pan-, specific-genomes, pan-/core-plot, pan-/core-matrix, and specific-matrix. Results of analysis can be downloaded as a table and postscript file. For core-, pan-, and specific-gene families basic statistics and Gene Ontology information can additionally be predicted as described above. More detailed instructions and output examples are provided on the server web page.

In the future we plan to update the approach with the analysis features and data visualisation possibilities. Moreover, a web-interface will provide the possibility to compare known genomes to multiple user-submitted isolates.

Acknowledgement

Authors received supported by the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.
References


4.2 Paper IV. (Manuscript). Life’s Set of Core Genes, Revisited
Life's Set of Core Genes, Revisited

Oksana Lukjancenko and David Wayne Ussery

1Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
2Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Abstract
There is a core set of functions required for all of life, and in principle one would expect a corresponding set of core proteins to be conserved across genomes. Unfortunately, as more genomes have been sequenced, the set of core genes has continually dropped, from 256 proteins, based on 2 genomes, to 31 proteins, based on about 200 genomes, to zero proteins conserved in a thousand bacterial genomes. We have developed a novel method - PanFunPro, and used this to re-examine the core set of proteins found in 2110 genomes. PanFunPro is based on models of functional domains, present in more than 85% of the proteins for most genomes. We find a stable set of 39 profiles and more than a hundred domains that are conserved across more than 99% of the genomes. The majority of these proteins are involved in protein synthesis, including many ribosomal proteins. We find nearly 100% conservation of amino-acyl tRNA synthetases, and strong conservation of the 36 large and 21 small ribosomal proteins across all genomes. Further, we find protein families responsible for the basic functions for life (replication, regulation, metabolism) to be conserved across all organisms.

Introduction
Comparison of the first two sequenced genomes, Mycoplasma genitalium (1) and Haemophilus influenzae (2), found 256 genes as a first estimate of the mini-
Table 1: History of minimal genome analysis.

<table>
<thead>
<tr>
<th>Number of genomes analysed</th>
<th>Number of conserved genes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>256</td>
<td>(3)</td>
</tr>
<tr>
<td>21</td>
<td>81</td>
<td>(6)</td>
</tr>
<tr>
<td>45</td>
<td>23</td>
<td>(7)</td>
</tr>
<tr>
<td>66</td>
<td>32</td>
<td>(8)</td>
</tr>
<tr>
<td>100</td>
<td>63</td>
<td>(6)</td>
</tr>
<tr>
<td>34</td>
<td>80</td>
<td>(9)</td>
</tr>
<tr>
<td>147</td>
<td>34</td>
<td>(10)</td>
</tr>
<tr>
<td>27</td>
<td>71</td>
<td>(11)</td>
</tr>
<tr>
<td>191</td>
<td>31</td>
<td>(12)</td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>(13)</td>
</tr>
</tbody>
</table>

Genes shared by distantly related organisms are likely to be essential and collection of these genes would reflect a minimal genome (4; 5). However, as more complete genome sequences have been published – the core genome has steadily declined, as seen in Table 1. Currently over 2600 complete prokaryotic genome sequences are available in addition to more than 10,000 draft genomes.

Prokaryotic pan-genome analysis provides insight to the genomic variation within groups of related microorganisms and can identify gene families strongly conserved within phylogenetic groups, although even within a large group, such as Proteobacteria, the number of conserved genes drops to zero (14).

A number of computational approaches are available for pan-genome analysis and finding possible essential genes. Many of these approaches use fast algorithms for pair-wise analysis and rely on the assumption that close relatives, sharing, and overall sequence identity above a certain threshold can be grouped into proteins families. For example, BLAST performs by identification of closely matching words, which are subsequently joined to build final alignment. However, there is no certainty that evolutionary processes and functions will be accurately represented by significant sequence similarity (15; 16). Another category of algorithms for sequence comparison is achieved by looking at the basic functional units that form proteins – protein domains (16). Protein domains are defined as sequential and structural motifs that are found independently in
different proteins, in different combinations (17). A variety of computational
approaches have been developed to identify protein domains and predict protein
function. Perhaps the most widely used are those that search Hidden Markov
model (HMMs) collections, such as PfamA (18), Superfamily (19) or TIGRFAM
(20), and combine proteins with the same domain architecture. Functional pre-
diction using probabilistic models can improve protein annotation and provide
a better understanding of organism complexity and evolutionary processes.
Here, we performed analysis using PanFuPro approach (21), which uses combi-
inations of functional domains (functional profiles) to group genes into protein
families, with the purpose of estimating the pan-genome of the fairly large set
of more than two thousand genomes, and to identify the minimal genome set of
core genes conserved across all of the genomes.

Results and Discussion

We have combined the sequence information of 2110 prokaryotic genomes, 1982
Bacterial and 128 Archaeal (all of the ‘complete’ prokaryotic genomes from
NCBI available in September 2012, see Table S1). Proteomes of each genome
were scanned against three HMM collections and the fraction of genes covered
by each of these databases is shown in Figure 1.

On average, more than 80% of proteins encoded by a genome have at least
one significant match in the PfamA database; 0.4% and 2.1% of the remaining
genomes could be covered by TIGRFAM and Superfamily databases, respectively.
However, no HMM domains were detected for approximately 15% of the pro-
teins for most genomes. In Figure 1, there are seven genomes (belonging to M.
haemofelis, M. hemocanis, M. wenvonii, Candidatus M. haemoninutum, and M.
leprae) that are clear outliers, with more than half of the proteins not having
any match to the HMMs. This might be due to several causes such as gene
prediction errors (Fig. S1), absence or inability to detect functional domains in
the sequence (Fig. S2), or genome decay (that is, the presence of large numbers
of pseudogenes, as in the known case of M. leprae). At any rate, the proteins in
these seven genomes have significantly fewer matches to any of the databases,
compared to the other 2103 genomes.
The proteome of each genome was grouped into a set of protein families based
on the presence or absence of a functional profile (combination of functional
domains) within the set of proteins. The collection of 2110 single-genome func-
tional profiles was combined into the complete pan-genome (Table S2). The
Figure 1: Distribution of genes covered by HMM-based databases. Each of 2110 proteomes was searched against PfamA, TIGRFAM, and Superfamily databases in corresponding order. The figure illustrates the fraction of genes covered by each database with respect to the order in which the groups of proteins were scanned. The last column represents the fraction of genes, which did not result in any significant hit in any considered HMM collection.

pan-genome contained a total number of 737,692 distinct protein families, where 10,858 different HMM-based domains served as structural units to compose 40,920 functional families, and more than 720,000 families resulted from clustering of dispensable genes with no matches in PfamA, TIGRFAM or Superfamily. Approximately one fourth of the HMM-based profiles appeared to be a single domain and almost one half of the domains were seen in only one type of profiles, while most of the domains tended to combine with other protein units to form different combinations (Figure S3). On average, each profile consisted of 9 domains (median = 2), and respectively, each domain was involved in 2
different combinations (median = 2).
The pan-genome contains all gene families, and it consists of a small set of highly conserved genes (the core genome) and a large set of accessory proteins, which are present in some but not all genomes or are unique to a certain strain (22).

We find 19 functional profiles that are strictly conserved amongst all 2110 genomes (17 of these are ribosomal proteins); there are 60 different individual functional domains conserved across all genomes. As shown in Table 2, allowing the absence of a functional profile or domain in even a single profile increases the number, and in 99% of the genomes (that is, missing in fewer than 21 genomes) we find 39 profiles (Table S3) and a total of 102 domains conserved (Table S4). We compared core functions estimated by this study to the commonly used set of 31 universally conserved genes (UCGs), previously suggested by Ciccarelli et al. (12). Comparison showed that 14 UCGs are present in the core of 100% of genomes; and 20 are shared by 99% of genomes. On the other hand, if we consider the core genome of single functional domains, 27 UCGs showed 100% conservation, three were missing in one genome, and one protein was absent in more than 1% of the genomes. However, the UCGs are missing several important proteins, including, for example, translation initiation and elongation factors, as well as some of the 54 well-conserved ribosomal proteins found in nearly all bacterial genomes. Computational analysis tends to underestimate the minimal gene set by considering only those genes that have remained similar enough during the course of evolution or that share strict domain architecture (23).

To investigate the major functional roles of core genes and characterize biological processes shared within the microbial organisms, we compared functional domains to the Clusters of Orthologous Groups (COG) database (24), as shown in Figure 2. Nearly all of the conserved domains (101 out of 102) have well-defined biological functions. Three COG functional groups (J, K, and L) involved in genetic information storage and processing are most abundant within the set of core domains. These categories contained, on average, 85% of domains and 91% of profiles in each threshold group. Additionally, proteins involved in metabolism (C, E and F) and cellular processes and signaling (D, O and U) are included in the core-genome, although these functions are less enriched.

The strong bias in core genes involved in protein synthesis is consistent with previous work (12). An alternative approach, based on conserved protein structural folds in a set of 420 genomes (25), found a minimal core of 70 folds, of which 40 were in the metabolism COG functional group and only 5 folds belonged to
Table 2: Distribution of COG functional categories in the Pfam domains, found in 100% and 99% of all prokaryotic genomes. The core genes were compared to the COG database and the functional category for each gene was defined.

<table>
<thead>
<tr>
<th>General Functional Group</th>
<th>COG Functional Group</th>
<th>100% conservation Domains</th>
<th>99% conservation Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Profiles</td>
<td>Domains</td>
</tr>
<tr>
<td>Information storage and processing</td>
<td>(J)</td>
<td>18</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>(K)</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(L)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Metabolism</td>
<td>(C)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(E)</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(L)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cellular Process and Signalling</td>
<td>(D)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(O)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(U)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Poorly characterized</td>
<td>(R)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination and repair
C: Energy production and conversion
E: Amino acid transport and metabolism
F: Nucleotide transport and metabolism
D: Cell cycle control, cell division, chromosome partitioning
O: Posttranslational modification, protein turnover, chaperones
U: Intracellular trafficking, secretion, and vesicular transport
R: General function prediction
the "J" translation COG group. However, these results were based on a set of genomes where parasitic organisms were excluded. We find that exclusion of the small parasitic genomes has little effect on the distribution and size of the core (Table 3). Exclusion of genomes encoding less than a thousand proteins yielded only three new architectures and thirteen domains.

The number of conserved genes varies with the number of analysed genomes and taxonomic diversity (14). Previously, in the study by Segata et al., sizes of taxa-specific core-genomes were identified. Genomes belonging to large phyla, such as Proteobacteria, Firmicutes or Actinobacteria, have one or zero genes in common. In this study we find that even in large and diverse taxonomic groups, the core genome size should consist of at least 49 genes. An overview for each prokaryotic phylum is summarized in Table 4; as more sequenced genomes become available for genomes of the same bacterial species, it is possible to extend this list to taxa-specific gene families for genera, species, and strains or serovars.

Genes involved in fundamental functions are more universally conserved and are expected to have a lower number of duplications than non-essential genes. This occurs because an essential gene's function is crucial for the organism to survive and is less likely to be compensated by its paralog (26; 27). In contrast, membrane proteins, such as transporters or participants of metabolic and cellular signaling processes can count more than 100 members within the same family. We extracted the list of the top five most abundant gene families, as shown in Figure 3. Perhaps not surprisingly, enzymes of these families participate in the following processes: transmembrane transport (PF00005 and PF07690), peptide transport (PF00528), oxidation-reduction (PF00106), and transcriptional regulation (PF00126_PF03466). Even though members within these five families encode the same function, protein sequences are more distant (Table S5). On average, protein sequences of the same family are 14% to 26% identical.

Transporters are important for all molecular processes within a living organism: metabolism, cellular communication, reproduction and biosynthesis. They allow all essential nutrients to enter the cell and its compartments; catalyze export and uptake of macromolecules (proteins, complex carbohydrates, lipids and DNA) and signaling molecules; promote the generation of ion electrochemical gradients; and prevent toxic effects of drugs and toxins by catalyzing their active efflux (28). Helix-turn-helix are DNA-binding motifs which are associated with regulation of transcription and short-chain dehydrogenases catalyse NAD(P)(H)-dependent oxidation/reduction reactions.
Table 3: Profile and domain core genomes within two sets of genomes, excluding small genomes.

<table>
<thead>
<tr>
<th>General Functional Group</th>
<th>COG</th>
<th>Genomes with 700+ genes</th>
<th>Genomes with 1000+ genes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100% conservation</td>
<td>99% conservation</td>
</tr>
<tr>
<td></td>
<td>Profiles</td>
<td>Domains</td>
<td>Profiles</td>
</tr>
<tr>
<td>Information storage and processing</td>
<td>(J)</td>
<td>19</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>(K)</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(L)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Metabolism</td>
<td>(C)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(G)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(E)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(F)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(H)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(I)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cellular processes and signalling</td>
<td>(D)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(M)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(O)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(U)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Poorly characterized</td>
<td>(R)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>(R)</td>
<td>22</td>
<td>69</td>
</tr>
</tbody>
</table>

J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination and repair
G: Carbohydrate transport and metabolism
F: Nucleotide transport and metabolism
H: Coenzyme transport and metabolism
I: Lipid transport and metabolism
M: Cell wall/membrane biogenesis
C: Energy production and conversion
E: Amino acid transport and metabolism
D: Cell cycle control, cell division, chromosome partitioning
O: Posttranslational modification, protein turnover, chaperones
U: Intracellular trafficking, secretion, and vesicular transport
R: General function prediction
Table 4: Estimation of core genome sizes for different phyla.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>Number of genomes</th>
<th>Number of shared genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acidobacteria</td>
<td>7</td>
<td>879</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>225</td>
<td>132</td>
</tr>
<tr>
<td>Aquificae</td>
<td>11</td>
<td>525</td>
</tr>
<tr>
<td>Bacteroidetes</td>
<td>78</td>
<td>172</td>
</tr>
<tr>
<td>Caldiserica</td>
<td>1</td>
<td>1202</td>
</tr>
<tr>
<td>Chlamydiae</td>
<td>73</td>
<td>422</td>
</tr>
<tr>
<td>Chlorobi</td>
<td>11</td>
<td>737</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>17</td>
<td>352</td>
</tr>
<tr>
<td>Chrysiogenetes</td>
<td>1</td>
<td>1904</td>
</tr>
<tr>
<td>Crenarchaeota</td>
<td>43</td>
<td>265</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>44</td>
<td>487</td>
</tr>
<tr>
<td>Deferrribacteres</td>
<td>4</td>
<td>867</td>
</tr>
<tr>
<td>Deinococcus-Thermus</td>
<td>17</td>
<td>614</td>
</tr>
<tr>
<td>Dictyoglomi</td>
<td>2</td>
<td>1154</td>
</tr>
<tr>
<td>Elusimicrobia</td>
<td>2</td>
<td>466</td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>81</td>
<td>2389</td>
</tr>
<tr>
<td>Fibrobacteres</td>
<td>2</td>
<td>2156</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>448</td>
<td>90</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>5</td>
<td>453</td>
</tr>
<tr>
<td>Gemmatimonadetes</td>
<td>1</td>
<td>2516</td>
</tr>
<tr>
<td>Ignavibacteria</td>
<td>2</td>
<td>1191</td>
</tr>
<tr>
<td>Korarchaeota</td>
<td>1</td>
<td>1229</td>
</tr>
<tr>
<td>Nitrospirae</td>
<td>3</td>
<td>681</td>
</tr>
<tr>
<td>Planctomycetes</td>
<td>6</td>
<td>661</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>886</td>
<td>49</td>
</tr>
<tr>
<td>Spirochaetes</td>
<td>47</td>
<td>190</td>
</tr>
<tr>
<td>Synergistetes</td>
<td>5</td>
<td>551</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>62</td>
<td>86</td>
</tr>
<tr>
<td>Thaumarchaeota</td>
<td>2</td>
<td>789</td>
</tr>
<tr>
<td>Thermodesulfobacteria</td>
<td>2</td>
<td>939</td>
</tr>
<tr>
<td>Thermotogae</td>
<td>15</td>
<td>510</td>
</tr>
<tr>
<td>Verrucomicrobia</td>
<td>4</td>
<td>562</td>
</tr>
</tbody>
</table>
Closer look into conservation of functional domains, representing protein families, which are involved in transcriptional regulation, free energy production, transmembrane transport, and genetic information processing, are shown in Figures S4-S9. Figure S4 shows the conservation of enzymes involved in the flow of genetic information. In order for protein synthesis to occur, several essential enzymes are necessary for three basic steps: Replication, Transcription, and Translation, coloured green, red and blue in Figure S4, respectively. Note that in general the DNA polymerization process (green) is highly conserved across essentially all genomes, and the translation (protein synthesis) is also quite well conserved, but the transcriptional enzymes (red) appear to be less well conserved. Genomes were scored for the presence of at least one functional domain per step (column in the figure). Each group is further divided into factors involved in the polymerization process - Initiation, Elongation, or Termination.

Figure 2: Top 5 most abundant protein families. The plot shows distribution of the number of genes per each family within a set of 2110 genomes.
of each step; the names of factors are highlighted in light green, yellow, or red, respectively. Figure S5 examines the conservation of helix-turn-helix regulator families across the 2100 genomes – only a few (AraC, GntR, RpiR, and BirA) are found in more than 2000 genomes. But still, this list of different helix-turn-helix families distributed across the genomes is impressive. As before, the calculation was done with the assumption that if at least one functional domain representing the particular helix-turn-helix transcriptional regulator/regulator family is present in the genome – then this function is conserved in the genome. The ribosomal proteins and amino-acyl tRNA synthetases are quite well conserved across all genomes, as can be seen in Figures S6 and S7. In contrast, Figure S8 shows that the proteins involved in membrane transport are not nearly as well conserved, although a few families are conserved in nearly all genomes. Finally, conservation of enzymes involved in glycolysis and the TCA were examined, based on conserved functional domains, as shown in Figure S9. In general, profiles for enzymes involved in glycolysis and the TCA cycle can be found most of the genomes (1800 genomes on average, or about 85% of the genomes). Closer look into conservation of protein families, representing transcriptional regulation, free energy production, and transmembrane transport, within this study, found lower conservation levels than for proteins involved in genetic information processing.

Conclusions

In conclusion, an analysis of more than two thousands prokaryotic genomes suggests that the minimal genome contains approximately one hundred functional domains found in more than 99% of the genomes. We find 19 functional profiles and more than 60 single functional domains strictly conserved in 100% of the genomes. This study also confirms that proteins shared amongst all the genomes are largely involved in processes responsible for genetic information processing and some metabolic pathways. In addition, enzymes involved in different metabolic pathways, communication gain more flexibility due to the ability of microbial organisms to adapt to different environmental conditions and to obtain less-essential genes from the environmental niches in which they live. The core genome size varies with phylogenetic diversity and depends on the number of analysed genomes. Nevertheless, the number of genes shared by the taxonomic group should be more than zero or one. We find that the set of
approximately one hundred core domains encodes functions that could allow organisms to reproduce, respond to the environment, and metabolize food.

Acknowledgement

The authors are grateful to all research groups that have submitted their genome sequences to public databases, without which this analysis would not have been possible. Authors received supported by the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

References


REFERENCES


4.3 Paper V. (Manuscript). Chromosome-specific families in Vibrio genomes
Chromosome-specific families in *Vibrio* genomes

Oksana Lukjancenko\(^1\) and David Wayne Ussery\(^{1,3}\)

\(^1\)Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
\(^3\)Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Abstract

We have compared chromosome-specific genes in a set of 18 finished *Vibrio* genomes, and also compared more than 250 draft sequences. These genomes represent a total of 9 known species and 2 unknown species. Within the finished chromosomes, we find a core set of 1269 gene families for chromosome I, and a core of 252 gene families for chromosome II. Many of these core genes are also found in the draft genome sequences (although of course which chromosome they are located on is unknown.) Of these chromosome specific core gene families, 1169 and 153 are uniquely found in chromosome I and II, respectively. We found gene ontology (GO) terms for the gene families, and compared the different sets for each chromosome. A total of 363 different ‘Molecular Function’ GO categories were found for chromosome I specific gene families, and these include several broad activities: pyridoxine 5’ phosphate synthetase, glucosylceramidase, heme transport, DNA ligase, amino acid binding, and ribosomal components; in contrast, chromosome II specific gene families have only 66 Molecular Function GO terms, and include many membrane-associated activities, such as ion channels, transmembrane transporters, and electron transport chain proteins. Thus, it appears that there are distinct sets of functions that are unique to each chromosome.

Introduction

Strains of the *Vibrio* genus belong to *Gammaproteobacteria*, are abundant and highly variable. These bacteria have the ability to form biofilm on biotic and abiotic surfaces, and are ubiquitous in marine and estuarine environments, at notably high densities in fish, corals, shrimps, plankton, and mammals (1; 2; 3). Currently the *Vibrio* genus consists of more than 60 different species, although complete genome sequences are available for only 10 species. Several species are known to be pathogens in human, fish, and marine invertebrates, and are well studied. *V. cholerae* can act as the causative agent of the severe and sometimes lethal disease cholera, and is probably the most sequenced and clinically important member of *Vibrio* species (4; 5). *V. vulnificus* causes septicemia in wound infections; however, despite its high fatality rate, human infections of *V. vulnificus* are rare (6; 7). *V. parahaemolyticus* and *V. furnissii* infections may lead to gastroenteritis in human via consumption of raw seafood (8; 9). Strains of *V. anguillarum* species are life-threatening to many economically important fish, including Atlantic salmon, seabass, cod, and rainbow trout (10). *V. fischeri* participates in beneficial symbioses with many marine organisms, especially squids (11). *V. harveyi* causes luminous vibriosis, which infects prawns, oysters, and lobsters (12). Finally, *V. splendidus* is known as extensive bivalve
All known *Vibrios* have two chromosomes. Chromosome I is usually larger, with relatively constant size, and possess essential functions; whereas chromosome II is smaller, varies in size, and shows diversity in the encoded genes. The existence of two chromosomes in all *Vibrio* genomes, and variance of chromosome II, has been an insight to many investigations worldwide and brought up multiple discussions about the purpose and origin of smaller chromosome. One of such speculations proposed that chromosome II originated as as megaplasmid, although later Heidelberg et al. have suggested that it may play important role in the organism and could help optimize the fast replication rate (2; 14; 15; 16).

The aim of this study is to compare more than 300 strains of *Vibrio* genus, both complete and available draft genomes, and to focus on distribution of functional genes and available Gene Ontology information between two chromosomes. Furthermore this study could be extended to other *Vibrios* analysis, using information about whether a gene belongs to chromosome I, or chromosome II.

**Material & Methods**

**Selection and Characteristics of Bacterial strains**

Publically available *Vibrio* strains were selected for this study and obtained from NCBI (July 2012). Initial set included 368 genomes, 18 of them were complete and 350 were retrieved as Illumina raw reads from NCBI Sequence Read Archive (SRA). Of these, 188 genomes were sequenced using a HiSeq 2000 sequencer and the remaining 162 with an Illumina Genome Analyzer II.

Open-reading frame (ORFs) predictions were carried out by gene-finding tool Prodigal (17). 16S ribosomal RNA sequences were extracted for both complete and draft *Vibrio* genome using RNAmmer (18). For each of assembled genome, the number of fragments (contiguous pieces), genes, and the mean gene length were calculated; strains with an average gene length below 700 bp were excluded from the further analysis. The resulting set consisted of 18 complete genomes and 284 draft sequences. The distribution of these characteristics for each genome is shown in Figure 1.

**Proteome comparison**

Proteome comparison was performed by PanFunPro tool (19). Briefly, genes of each proteome were annotated as described by Lukjancenko *et al.* and grouped into gene families. Results of pan- and core-genome analysis for chromosomes I and II were both visualized as an accumulative pan-/core-plot and pairwise comparison matrix.

The distribution of unique functional profiles between the large and small chromosomes was examined, following by brief investigation of available GO functional categories, specific for each of the chromosomes.

One representative proteome for each species was chosen from the pool of complete genomes and interspecies analysis of specific-genomes was performed between each pair of species. Results were visualized as a specific-matrix.

**Results & Discussion**

The bacterial dataset consisted of 302 genomes, representing 9 known and 2 unknown *Vibrio* species. A list of the species and numbers of representing genomes are shown in Table 1. Only
18 of the strains were completely finished, and for those independent proteomes for both chromosome I and II could be extracted. However most of the genomes (284) were assembled and present in multiple contigs, with no available information of which gene belongs to which chromosome. Thus it was decided to build analysis around 2 sets: finished genomes (set_18) and the whole dataset, including the WGS draft genomes (set_302).

Table 1: List of species analysed in this study. For each species the number of available genomes and sequence status are provided. Species are listed alphabetically.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of genomes</th>
<th>Sequence status</th>
</tr>
</thead>
<tbody>
<tr>
<td>V. alginolyticus</td>
<td>1</td>
<td>Draft</td>
</tr>
<tr>
<td>V. anguillarum</td>
<td>1</td>
<td>Complete</td>
</tr>
<tr>
<td>V. cholerae</td>
<td>279</td>
<td>Complete, Draft</td>
</tr>
<tr>
<td>V. furnissii</td>
<td>1</td>
<td>Complete</td>
</tr>
<tr>
<td>V. fischeri</td>
<td>1</td>
<td>Draft</td>
</tr>
<tr>
<td>V. harveyi</td>
<td>1</td>
<td>Complete</td>
</tr>
<tr>
<td>V. parahaemolyticus</td>
<td>1</td>
<td>Complete</td>
</tr>
<tr>
<td>V. splendidus</td>
<td>12</td>
<td>Complete, Draft</td>
</tr>
<tr>
<td>V. vulnificus</td>
<td>3</td>
<td>Complete</td>
</tr>
<tr>
<td>Vibrio sp. EJY3</td>
<td>1</td>
<td>Complete</td>
</tr>
<tr>
<td>Vibrio sp. Ex25</td>
<td>1</td>
<td>Complete</td>
</tr>
</tbody>
</table>

The calculated basic features for each analyzed genome is shown in Figure 1, including the number of contiguous pieces, predicted genes, average gene lengths, and predicted 16S rRNAs. A large fraction of the assembled genomes contained between 150 and 190 contigs, with a group of outlier strains, showing more than 200 pieces per genome. An obvious correlation can be seen between number of contigs and amount of predicted rRNAs and genes, following by smaller average gene length in assembled genomes with higher number of contigous sequences.

**Vibrio Chromosome I and Chromosome II comparison**

Vibrio chromosome one is larger and more stable, carrying essential genes, whereas chromosome two is smaller, more variable in size and believed to contain more specific functions. Pairwise comparison of pan- and core-genome was performed on set_18 for both chromosomes and visualized in Figure 2. It can be seen that chromosome I and chromosome II share between 10% and 15% of gene families, while similarity within smaller chromosome ranges between 25% and 96%, and between 55% and 95% in larger chromosome. Since there are multiple genome sequences for several different strains available for the *V. cholerae* species, a high similarity within chromosomes can be found with confidence, although on average only 10% are shared between chromosomes I and II.

Analysis of the total pan- and core-genome of complete strains resulted in 1269 conserved protein families shared within chromosome I, and 252 core families in chromosome II; only 104 functional profiles are shared between two chromosomes. When the draft genomes were included, the core-genomes of chromosomes I and II dropped to 673 and 140 protein families, following by decrease of shared functional profiles to a total number of 96. The pan- and core-genome summary results are shown in Table 2 and conserved profiles and their functions in Table S1.

A closer look to the distribution of functions within core-genome of two chromosomes showed that all of the shared genes are annotated by PfamA database (Figure S1) and most of them are involved in biological processes or molecular function (Figure 3). The presence of proteins
involved in essential metabolic and regulatory processes in the shared genomic pool of both chromosomes validates the claim that the smaller chromosome is not a plasmid, but is fundamental for growth and biological activity.

Figure 1: Predicted genome characteristics. **A.** Distribution of number of contiguous pieces; **B.** Distribution of protein number per genome; **C.** Distribution of average gene length per genome; **D.** Number of predicted 16S rRNA sequences.

Are there genes that are conserved in each of two chromosomes and absent in another? For this purpose, we extracted genes, which would be in the core of chromosome I and are absent in the core of chromosome II (Figure 4); and vise versa, present in the core of smaller chromosome, and absent in the core of larger chromosome (Figure 5). A total number of 639 GO IDs could be extracted for chromosome I core-specific profiles (1169 profiles). Of these 438 were involved in biological process, 53 in cellular component functions and 363 carried molecular functions. Equivalent analysis of chromosome II core-specific profiles yielded in total 109 Go IDs (of 153 profiles), and the distribution among three main groups is as follows: 57 in biological process, 10 in cellular component, and 66 in molecular function. It is not surprising that core of larger chromosome carries more genes, essential to sustain life and reproduce; and the specific core for the smaller chromosome contains proteins involved in metabolic processes, enzyme and membrane associated activity. Addition of 284 draft genomes slightly reduced the number of specific genes and specific pathway groups in chromosome I, remaining 265 GO terms involved in biological process, 39 in cellular component functions, and 197 - molecular function (Figure S2). Whereas, chromosome II contained 15, 4, and 14 GO Terms in each of the following groups: biological process, cellular component, and molecular function, respectively (Figure S3).
Figure 2: Pairwise pan- and core-genome comparison. Comparison was performed for set 18 genomes. Blue and green square boxes separate chromosome I and II, respectively. Red-coloured box in the middle of the figure indicates inter-chromosomal comparison of *V. cholerae* species, when black-coloured triangles highlight similarities within the same chromosome of the species.
Table 2: Pan- and core-genome calculation.

<table>
<thead>
<tr>
<th></th>
<th>set_18</th>
<th>set_302</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Core-genome</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome I</td>
<td>1269</td>
<td>673</td>
</tr>
<tr>
<td>Chromosome II</td>
<td>252</td>
<td>140</td>
</tr>
<tr>
<td>Both chromosomes</td>
<td>104</td>
<td>96</td>
</tr>
<tr>
<td><strong>Pan-genome</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome I</td>
<td>5498</td>
<td>NA</td>
</tr>
<tr>
<td>Chromosome II</td>
<td>3742</td>
<td>NA</td>
</tr>
<tr>
<td>Both chromosomes</td>
<td>7825</td>
<td>17363</td>
</tr>
</tbody>
</table>

NA - proteomes of assembled genomes cannot be separated

**Interspecies comparison**

The genus *Vibrio* genus comprises a diverse group of bacteria, which can be symbiotic or pathogenic to mammals and organisms of marine environments. Species-specific genomes can contain proteins responsible for pathogenicity or crucial for surviving in a given environment.

![Figure 3: GO term analysis in genes shared by chromosome I and II. Distribution is shared both as percentage on the axis and absolute number above the bar. Absolute number shows the amount of GO IDs that were connected to the pathway. Colour code is as follows: red is biological process, green is cellular component, and blue is molecular function.](image-url)
Figure 4: GO term analysis in genes shared within chromosome I and missing in the core of chromosome II. Distribution is shared both as percentage on the axis and absolute number above the bar. Absolute number shows the amount of GO IDs that were connected to the pathway. Colour code is as follows: red is biological process, green is cellular component, and blue is molecular function.

Figure 5: GO term analysis in genes shared within chromosome II and missing in the core of chromosome I. Distribution is shared both as percentage on the axis and absolute number above the bar. Absolute number shows the amount of GO IDs that were connected to the pathway. Colour code is as follows: red is biological process, green is cellular component, and blue is molecular function.
To show the level of specificity between species of the same chromosome, for 9 strains representing 7 known and 2 unknown species, pairwise comparison of specific-genome was performed. Within larger chromosome, fraction of unique proteome varies from 18% to 33% (Figure 6), whereas genomes of chromosome two differ in larger portion of proteins, ranging from 18% to 64% (Figure 7).

**Figure 6:** Pairwise interspecies-specific genome comparison for chromosome I. Analysis included single representative of 7 known and 2 unknown species. Resulting percentage shows the ratio between the amount of species-specific families and size of total proteome. On average each species contained between 18% and 33% specific protein families. Colour intensity indicates the level of specificity.

**Figure 7:** Pairwise interspecies-specific genome comparison for chromosome II. Analysis included single representative of 7 known and 2 unknown species. Resulting percentage shows the ratio between the amount of species-specific families and size of total proteome. On average each species contained between 17% and 68% specific protein families. Colour intensity indicates the level of specificity.
**Vibrio cholerae** spp. are known pathogens in human and were chosen as an example of investigation of which proteins specific-genome contains and what processes species-specific genes are be involved in. Representative strains of *V. cholereae* species were compared to other strains, as shown in Figure 7. Chromosome I and II contained similar amount of specific profiles, 190 and 192, respectively. Most of them were CD-HIT clustering-based, however 81 and 47 were annotated by PfamA and TIGRFAM collections. A complete list of profiles and corresponding functions are listed in Table S2.

### Proteomes of *V. cholerae* draft genomes

*V. cholerae* is one of the most important, highly documented, and mostly sequenced species of *Vibrios*. Our dataset included 279 cholera-causing strains, 8 completely sequenced and 271 draft genomes. Information of proteome separation into chromosome I and II was not available. Core-genome analysis of 279 *V. cholerae* strains yielded in 776, 250, and 182 gene families, in large, small, and both of chromosomes, respectively. Further we extracted all the proteins, which were not found in pan-genome of both chromosomes within set_18 genomes. Distribution of total number of 8325 functional profiles is as follows: 2333, 341 and 73 families assigned to PfamA, Superfamily, and TIGRFAM databases, respectively (Figure 8). Analysis shows, that 271 newly sequenced *V. cholerae* strains bring at least 2000 possible profile combinations to the pool of previously known functions, which represent more than 70 different GO functional categories (Figure 9). This extracted proteome might as well contain genes belonging to plasmids.

In conclusion, multiple analysis of similarities and differences between *Vibrio* species, showed that Vibrios are variable between species and chromosomes. Proteomes of larger chromosome are more similar, and carry important functions to sustain life.

### Disclosure/Conflict-of-Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

### Acknowledgements

The authors are grateful to all research groups that have submitted their genome sequences to public databases, without which this analysis would not have been possible. Authors received supported by the Center for Genomic Epidemiology at the Technical University of Denmark; part of this work was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.
Figure 8: Annotation and length distribution of genes within specific-proteomes within draft genomes of *V. cholerae*. **A.** Distribution of profiles by assignment source: PfamA, Superfamily, TIGRFAM, and CD-HIT clustering. **B.** Gene length distribution by each profile type.

Figure 9: GO term analysis in specific to *V. cholerae* draft genomes. Distribution is shared both as percentage on the axis and absolute number above the bar. Absolute number shows the amount of GO IDs that were connected to the pathway. Colour code is as follows: red is biological process, green is cellular component, and blue is molecular function.
References


Chapter 5

Microbial Identification Using Whole Genome Sequences

Identification of microbial genomes is usually carried out using traditional typing methods, such as 16S rRNA and MLST. This chapter provides insight into alternative microbial identification methods, which employ whole genome sequence comparison. Paper VI demonstrates the use of species-specific genes for high-density microarray design, which is used to evaluate the genomic content of unsequenced bacterial genomes within Enterobacteriaceae family.

Paper VII provides an insight into the use of comparative genomics for non-traditional epidemiological typing. A number of genes, shared between 79 Salmonella genomes were extracted; and genomic variation within these core gene was used to infer phylogenetic relationships between closely related genomes. Results were compared to traditional typing methods, such as 16S rRNA and MLST.

Taxonomy prediction can also be performed using protein functional content as a target. TaxonomyFinder is a new in silico approach, which uses HMM-profile combinations to infer microbial identification of unknown isolates. Performance of the method is shown in Paper VIII. TaxonomyFinder
was compared to other genomic typing methods, and the performance was evaluated on two different sets of genomes: Draft genome sequences and SRA genomes, which were publicly available in Genbank database.
5.1 Paper VI. Design of an Enterobacteriaceae Pan-genome Microarray Chip
Design of an *Enterobacteriaceae* Pan-Genome Microarray Chip

Oksana Lukjancenko and David W. Ussery

Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, 2800 Kongens Lyngby, Denmark

**Abstract.** Microarrays are a common method for evaluating genomic content of bacterial species and comparing unsequenced bacterial genomes. This technology allows for quick scans of characteristic genes and chromosomal regions, and to search for indications of horizontal transfer. A high-density microarray chip has been designed, using 116 *Enterobacteriaceae* genome sequences, taking into account the enteric pan-genome. Probes for the microarray were checked *in silico* and performance of the chip, based on experimental strains from four different genera, demonstrate a relatively high ability to distinguish those strains on genus, species, and pathotype/serovar levels. Additionally, the microarray performed well when investigating which genes were found in a given strain of interest. The *Enterobacteriaceae* pan-genome microarray, based on 116 genomes, provides a valuable tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

**Keywords:** *Enterobacteriaceae*, Pan-genome, DNA microarray analysis, gene, *Escherichia coli*.

1 Introduction

The risk of dying from disease caused by a bacterial infection is greater than that associated with any other type of disease, including cancer or heart attacks [1, 2]. Epidemic infectious diseases are the most serious causes of mortality and morbidity worldwide, more than all other diseases combined. Infections contribute to significant economic loss in most parts of the world, including first world countries that have high income and developed surveillance and control systems [3, 4]. Every year thousands of people are infected by bacterial pathogens, most of which are transmitted through food [5]. The outcome from food-borne human infections can range from mild self-limiting diarrhea to severe illness that requires hospitalization. In rare cases, food-borne illnesses are even fatal [5, 6]. Enteric bacteria, particularly *Salmonella enterica* subsp. *enterica*, are among the leading food-borne pathogens [6, 7]. In light of this, the detailed and rapid investigation of enteric pathogens is essential in modern epidemiology and clinical diagnostics.

*Enterobacteriaceae* are pervasive. They are widespread in the environment, existing in water, soil, food, and plants, as well as in the normal intestinal flora of many animals and humans [8-12]. Pathogens within this group have developed a diversity...
of strategies to overcome protective host barriers in order to invade the host, resist innate immune response, multiply in specific and normally sterile body sites, and damage cells in order to establish and maintain a successful infection [13, 14]. Genera within Enterobacteriaceae family are of interest, as well, because of problems from food spoilage and for that reason are of considerable economic importance [15].

Bacterial genomes vary in size, even among the strains of the same species. Bacterial species can be characterized by its pan-genome. As defined by Tettelin et al., the microbial pan-genome is a complete collection of various genes located within populations at a particular taxonomic level, commonly within a species. The pan-genome concept can of course be expanded to higher levels, such as genus or even a bacterial family. The pan-genome includes a core-genome, which is a minor fraction of the entire gene pool that is shared between all the given strains. Furthermore, there is a much larger, dispensable portion of bacterial genes, that are missing in one or more strains. Also there are some genes that appear to be unique to each strain [16, 17]. Strain-specific genes can, even among a particular species, make up a notably large portion of the pan-genome [18].

Many methods have been developed for characterizing genetic variation. Use of DNA microarrays is becoming a standard procedure for evaluating genotyping – that is, looking at the genetic content of a bacterial species. The price for microarrays used for genotyping was historically expensive, but now is becoming competitive with the cost of other commonly used typing methods, such as previously widely used multi-locus sequence typing (MLST). Moreover, it is becoming increasingly popular, quick, and cost-effective to define the presence and absence of each of the assigned genes in the pan-genome of a species. Thus, microarrays, imprinted with all the genes from species’ pan-genome can be used to compare and characterize the genomic content of unknown bacterial isolates and to achieve accurate typing information, that can be useful in epidemiological investigations and clinical diagnostics [1, 19]. For instance, array comparative genomic hybridization (aCGH) is frequently used in human cancer studies to genotype cell lines by determination of gene loss and copy number variations [20] or to detect single nucleotide polymorphisms at target loci [21]. Additionally, microarrays have been widely used in human screenings for the determination and genotyping of bacterial species. Microarrays have changed considerably since they were first introduced. Early microarrays for the E. coli genome consisted of long fragments of chromosomal DNA (~1000 to 2000 base-pairs), attached to a microscope slide. Later, Affymetrix made an array covering the entire E. coli K-12 genome using a set of 10 to 15 probes (synthetic 25mers) for each gene [22], followed shortly by an array which contained 4 E. coli genomes [23, 24]. Custom-designed NimbleGen chips have been made including 7 and then 32 E. coli genomes [25, 26].

This study describes the design and use of a high-density oligonucleotide microarray covering the pan-genome of 116 genomes within the Enterobacteriaceae family. Probes are designed to distinguish among organisms at the level of genera, species, and even single strains. Moreover, probes for determination of particular gene families, comprising Enterobacteriaceae pan-genome, are defined. The performance of this microarray is evaluated both in silico and experimentally. Its utility is illustrated for the hybridization of genomic DNA in order to compare uncharacterized isolates which have not been sequenced with the 116 known, sequenced strains. A microarray chip approximating the complete pan-genome of Enterobacteriaceae
provides optimal sensitivity to characterize isolates. Gene family microarray analysis is useful for medical and environmental diagnoses and will provide an alternative to costly genome libraries, as well as to the sequencing of environmental samples.

2 Materials and Methods

2.1 Bacterial Strains

In this study, one hundred and twelve complete Enterobacteriaceae genome sequences and four in progress, which were publically available in GenBank database at the time of analysis (February, 2010), were used for custom microarray design. An overview of the used strains is shown in Table 1 and the complete collection of the strains is described in supplementary Table S1.

Table 1. Enterobacteriaceae genera used in the design of the microarray chip

<table>
<thead>
<tr>
<th>Genus</th>
<th>Number of strains</th>
<th>Genus</th>
<th>Number of strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buchnera</td>
<td>6</td>
<td>Photorhabdus</td>
<td>2</td>
</tr>
<tr>
<td>Citrobacter</td>
<td>3</td>
<td>Salmonella</td>
<td>18</td>
</tr>
<tr>
<td>Cronobacter</td>
<td>2</td>
<td>Serratia</td>
<td>1</td>
</tr>
<tr>
<td>Dickeya</td>
<td>3</td>
<td>Shigella</td>
<td>8</td>
</tr>
<tr>
<td>Edwardsiella</td>
<td>2</td>
<td>Sodalis</td>
<td>1</td>
</tr>
<tr>
<td>Enterobacter</td>
<td>2</td>
<td>Wigglesworthia</td>
<td>1</td>
</tr>
<tr>
<td>Escherichia</td>
<td>35</td>
<td>Xenorhabdus</td>
<td>1</td>
</tr>
<tr>
<td>Klebsiella</td>
<td>4</td>
<td>Yersinia</td>
<td>14</td>
</tr>
<tr>
<td>Pectobacterium</td>
<td>3</td>
<td>Erwinia</td>
<td>4</td>
</tr>
<tr>
<td>Proteus</td>
<td>3</td>
<td>Candidatus*</td>
<td>3</td>
</tr>
</tbody>
</table>

* Candidatus is not a genus; however some strains were included as they were classified as Enterobacteriaceae at the time of study.

Twelve bacterial strains included in experimental evaluation of the chip are listed in Table 3 (Results section).

2.2 Pan-Genomics

The pan-genome was estimated, as described by Snipen et al [27]. Briefly, all protein sequences were compared by BLASTP [28]. Two proteins were attributed to a single gene family if they satisfied the 50/50 rule, meaning that when they could produce a pairwise BLASTP alignment covering at least 50% amino of the length of the longest protein with at least 50% of amino acid identity. Each genome was compared successively: for each n additional genome, that genome was compared to any combinations of n-1 genomes and the number of identical ‘core genes’ and ‘genome specific genes’ (specific for genome n) were counted for each n. All cumulative BLASTP hits found in the whole set of genomes were plotted as a running total and were considered as pan-genome, which increases as more genomes are added. The number of gene families with at least one representative in every genome was plotted for the core-genome.

1 Available at http://www.cbs.dtu.dk/~dave/Supplementary_TableS1.pdf
2.3 The Custom-Microarray Design

The custom probe set for the microarrays was designed around 78 different groups of genomes (the list of groups is presented in the Results section, Table 2) including a collection of generic probes for the entire enteric core (97 genes), as well as for the probes that differentiate each genus within *Enterobacteriaceae*. The custom probe set was followed by more specialized probe sets for species-specific classification within *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera and further probe groups were specific for strain and pathotype for *Escherichia coli* genus. Additionally, sets of probes for all the gene families, comprising pan-genome, were included. The custom microarrays, manufactured by NimbleGen, were based on the NimbleGen 12-plex platform.

2.4 Constructing Target Gene Sets

The genome sequences in this study (Table S1) were searched for genes using the Prodigal gene-finding approach [29] in order to standardize gene finding. All protein-coding sequences were aligned all-against-all using BLASTP [28], and similarity was decided according to 50/50 rule. Proteins that satisfy this rule were assigned to one protein family. ‘Group specific gene families’ (as described above) were found using batch Perl script, which outputs a list of gene families that are either common to or complementary to the genomes included in pan- and core-genome plots (depending on whether unique or core genes are extracted). Representative sequences from each gene network were selected by choosing the organism from which the genes should be extracted. Unique genes were considered to be those that appeared to be conserved only among the strains belonging to a particular group.

2.5 Probe Selection for Target Genes

Probes for target genes were selected using the OligoWiz program, previously described by Wernersson *et al.* [30][31]. At each position along all the input sequence, the suitability of placing a probe was evaluated according to several criteria: melting temperature ($\Delta T_m$), cross-hybridization, folding (self-annealing), position (within the transcript), and ‘low-complexity’ (absence of subsequences that occur very commonly in the genome/transcriptome). The weighting scores for these criteria are as follow: cross-hybridization, 39%; $\Delta T_m$, 26%; folding, 13%; position, 13%; and low-complexity, 9%. No probes were accepted unless an overall score of at least 0.3 was obtained, and all probes were required to have a length in the range of 42 bp to 50 bp. OligoWiz was originally designed for single genome use, and thus, the program was modified in order to make the mechanisms screening for cross-hybridization less strict as described by Vejborg *et al.* [32]. A new modified scheme included a log-transformation in the underlying calculations. The net effect is insignificant near the upper boundary of the score, but next to the lower boundary it increases the discriminatory power of the tool.

\[
BLAST\ max\ score = 1 - \sum_{i=1}^{n} \log(1 + \sum_{m=1}^{m} \frac{hm_i}{100})
\]
2.6 Probe Evaluation in silico

Probes were aligned against a database consisting of all possible gene sequences in the total data set using BLASTN. The affinity of each probe for every gene was determined and expressed as the number of identical base pairs and by the E-value. Sequences for which the E-value was lower than 0 were extracted using a batch Perl script. Probes that matched strains not expected to belong to particular group were excluded from the further analysis. If more than ten probes per gene remained available after filtering, only not-overlapping ones were used for subsequent analysis. This resulted in the reduction of candidate probes from 106,657 to 53,644. Consequently, the number of probes targeting each gene ranged from 3 to 14 with a median coverage of about 7 probes per gene.

2.7 DNA Preparation and Hybridization

All the experimental isolates were kindly provided by the laboratory of Frank Møller Aarestrup (DTU Food, The Technical University of Denmark). All test strains were grown overnight on blood agar and genomic DNA was isolated as described in the protocol for the Easy-DNA kit from Invitrogen [33]. The method used is briefly described here: the lysis of the cells was performed by the addition of solution A and subsequent incubation at 65°C. Proteins and lipids were precipitated and extracted by the addition of solution B and chloroform. The solution was then centrifuged to separate the solution into two phases. The DNA was in the upper, clear aqueous phase, the proteins and lipids were in the solid interface, and the chloroform formed the lower phase. The DNA was then removed, precipitated with ethanol, and re-suspended in TE buffer.

The genomic DNA was labeled with cy3 dye and hybridized to NimbleGen custom arrays according to Arrays User’s Guide for CGH analysis as provided by the manufacturer of the arrays (Roche NimbleGen, Madison, Wisconsin, USA).

2.8 Analysis Methods

In the initial step, the raw data from multiple microarrays was extracted using NimbleScan software, developed by Roche NimbleGen, and combined as a single input. Data analysis was performed in R (a statistical software program), using the ‘oligo’ package for analyzing oligonucleotide arrays at the probe level. The package was obtained from Bioconductor [34]. The probes were mapped to each gene group, including position, according to the design. Chip analysis workflow then continued as follows:

1. Performance of probe-level normalization using robust multi-array average (RMA) algorithm. RMA method had a three-step procedure consisting of background correction, normalization, and summarization to obtain gene-level relative intensity measures from probe-level intensities [35].
2. Estimation of gene ‘on/off’ status based on the summarized gene relative intensities and the median of these intensities for each of the 78 groups.

Supporting microarray chip design information is publicly available\(^2\).

\(^2\) http://www.cbs.dtu.dk/~dave/Microarray_Chip_Design_Lukjancenko_2010.ndf
3 Results

3.1 Pan-Genome and Core-Genome Estimation

For each of the considered bacterial strains listed in Table S1 (Supplementary data), the genome sequence was downloaded from NCBI/GenBank. Genes were predicted by Prodigal [29], and translated into proteins. This resulted in a dataset of 887,184 entries with considerable redundancy due to the presence of the same gene in multiple genomes. To reduce the homology, proteins were grouped into the gene families. Proteins were considered conserved (belonging to the same gene group) if they showed at least 50% amino acid identity in a BLASTP alignment covering at least 50% of the length of the longest protein. The combined pan-genome of 116 genomes within *Enterobacteriaceae* was estimated and appeared to contain 44,838 gene families. The core-genome, that is, the number of conserved genes present in all 116 genomes, was estimated to be comprised of 97 conserved gene families.

3.2 Probe and Microarray Design

In the presented *Enterobacteriaceae* pan-genome microarray design strategy, the probe set was designed around 78 different groups of genomes. The microarray was made up of a collection of probes for each genus within *Enterobacteriaceae*, being species-specific for *Klebsiella*, *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* genera; strain and pathotype specific for *Escherichia coli* genus; core genes; and all protein families, comprising pan-genome. Using the data from the pan- and core-genome estimation step, the number of ‘group-specific’ genes and probes was determined and are shown in Table 2. Genes were considered to be ‘group-unique’ if they were found only within genomes, belonging to a particular group, and were absent in all of the rest genomes among a set of 116 genomes.

The final result was a set of 52,356 *Enterobacteriaceae* target sequences, representing genes of both specific groups and pan-genome gene families. The oligos were then selected using OligoWiz [31] based on several criteria, including their specificity, self-annealing, presence of low-complexity sequences, and their lengths adjusted so as to standardize the hybridization strength. Probes were filtered in order to avoid complimentarity with unwanted targets. In the end a set of 130,540 non-overlapping probes with an average length of 49 bp were obtained. The average number of probes per target gene was about 7, although the actual number for any given target depended on the length of the sequence, since shorter sequences have space for fewer non-overlapping probes. For set of probes that represent gene families an average of 3 probes per family was used.

3.3 Validation of the Custom Arrays

The chip design was evaluated by analyzing and comparing hybridization data from twelve control strains, shown in Table 3. Microarray data can have noise, coming from multiple variations which can occur during the array manufacturing process, the preparation of the biological sample for the hybridization, the hybridization of the samples to the array itself, and the quantification of the spot intensities [35]. To remove such variation, which obviously will affect the measured gene intensity levels,
Table 2. Number of ‘group specific’ gene families and probes before and after in silico validation

<table>
<thead>
<tr>
<th>Probe group</th>
<th>Number of genes before validation</th>
<th>Number of probes before validation</th>
<th>Number of genes after validation</th>
<th>Number of probes after validation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Buchnera</em> genus</td>
<td>14</td>
<td>200</td>
<td>14</td>
<td>123</td>
</tr>
<tr>
<td><em>Candidatus</em> strains</td>
<td>41</td>
<td>584</td>
<td>41</td>
<td>373</td>
</tr>
<tr>
<td><em>Citrobacter</em> genus</td>
<td>20</td>
<td>171</td>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td><em>Cronobacter</em> genus</td>
<td>271</td>
<td>3224</td>
<td>270</td>
<td>2002</td>
</tr>
<tr>
<td><em>Dickeya</em> genus</td>
<td>155</td>
<td>2129</td>
<td>155</td>
<td>1398</td>
</tr>
<tr>
<td><em>Edwardsiella</em> genus</td>
<td>318</td>
<td>3803</td>
<td>317</td>
<td>2447</td>
</tr>
<tr>
<td><em>Enterobacter</em> genus</td>
<td>40</td>
<td>511</td>
<td>40</td>
<td>318</td>
</tr>
<tr>
<td><em>Erwina</em> genus</td>
<td>217</td>
<td>2919</td>
<td>217</td>
<td>1840</td>
</tr>
<tr>
<td><em>Escherichia</em> genus</td>
<td>1</td>
<td>15</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td><em>Escherichia</em> coli 042</td>
<td>106</td>
<td>1047</td>
<td>79</td>
<td>450</td>
</tr>
<tr>
<td><em>Escherichia</em> coli 536</td>
<td>142</td>
<td>1207</td>
<td>95</td>
<td>436</td>
</tr>
<tr>
<td><em>Escherichia</em> coli 55989</td>
<td>72</td>
<td>646</td>
<td>45</td>
<td>272</td>
</tr>
<tr>
<td><em>Escherichia</em> coli APEC</td>
<td>116</td>
<td>1287</td>
<td>14</td>
<td>83</td>
</tr>
<tr>
<td><em>Escherichia</em> coli APEC O1</td>
<td>116</td>
<td>1287</td>
<td>14</td>
<td>83</td>
</tr>
<tr>
<td><em>Escherichia</em> coli Avirulent</td>
<td>69</td>
<td>508</td>
<td>39</td>
<td>241</td>
</tr>
<tr>
<td><em>Escherichia</em> coli B phylogroup</td>
<td>14</td>
<td>175</td>
<td>14</td>
<td>100</td>
</tr>
<tr>
<td><em>Escherichia</em> coli CFT073</td>
<td>292</td>
<td>2251</td>
<td>115</td>
<td>393</td>
</tr>
<tr>
<td><em>Escherichia</em> coli E24377A</td>
<td>249</td>
<td>1700</td>
<td>90</td>
<td>511</td>
</tr>
<tr>
<td><em>Escherichia</em> coli EAEC</td>
<td>72</td>
<td>646</td>
<td>45</td>
<td>272</td>
</tr>
<tr>
<td><em>Escherichia</em> coli ED1a</td>
<td>159</td>
<td>1545</td>
<td>146</td>
<td>823</td>
</tr>
<tr>
<td><em>Escherichia</em> coli EHEC</td>
<td>21</td>
<td>173</td>
<td>13</td>
<td>27</td>
</tr>
<tr>
<td><em>Escherichia</em> coli EPEC</td>
<td>142</td>
<td>1685</td>
<td>126</td>
<td>893</td>
</tr>
<tr>
<td><em>Escherichia</em> coli ETEC</td>
<td>249</td>
<td>1700</td>
<td>90</td>
<td>511</td>
</tr>
<tr>
<td><em>Escherichia</em> coli ExPEC</td>
<td>52</td>
<td>392</td>
<td>17</td>
<td>131</td>
</tr>
<tr>
<td><em>Escherichia</em> coli HS</td>
<td>90</td>
<td>642</td>
<td>44</td>
<td>313</td>
</tr>
<tr>
<td><em>Escherichia</em> coli IAI1</td>
<td>67</td>
<td>499</td>
<td>39</td>
<td>238</td>
</tr>
<tr>
<td><em>Escherichia</em> coli IAI39</td>
<td>77</td>
<td>609</td>
<td>48</td>
<td>262</td>
</tr>
<tr>
<td><em>Escherichia</em> coli K-12</td>
<td>11</td>
<td>159</td>
<td>11</td>
<td>113</td>
</tr>
<tr>
<td><em>Escherichia</em> coli O103:H2</td>
<td>65</td>
<td>693</td>
<td>50</td>
<td>377</td>
</tr>
<tr>
<td><em>Escherichia</em> coli O111:H-</td>
<td>148</td>
<td>1536</td>
<td>54</td>
<td>250</td>
</tr>
<tr>
<td><em>Escherichia</em> coli O127:H6</td>
<td>142</td>
<td>1685</td>
<td>126</td>
<td>893</td>
</tr>
<tr>
<td><em>Escherichia</em> coli O157:H7</td>
<td>68</td>
<td>709</td>
<td>52</td>
<td>379</td>
</tr>
<tr>
<td><em>Escherichia</em> coli O26:H11</td>
<td>74</td>
<td>690</td>
<td>48</td>
<td>280</td>
</tr>
<tr>
<td><em>Escherichia</em> coli S88</td>
<td>52</td>
<td>392</td>
<td>17</td>
<td>131</td>
</tr>
<tr>
<td><em>Escherichia</em> coli SE11</td>
<td>178</td>
<td>1692</td>
<td>70</td>
<td>360</td>
</tr>
<tr>
<td><em>Escherichia</em> coli SE15</td>
<td>58</td>
<td>609</td>
<td>49</td>
<td>328</td>
</tr>
<tr>
<td><em>Escherichia</em> coli SMS-3-5</td>
<td>145</td>
<td>1064</td>
<td>106</td>
<td>501</td>
</tr>
<tr>
<td><em>Escherichia</em> coli UMN026</td>
<td>113</td>
<td>1026</td>
<td>85</td>
<td>505</td>
</tr>
<tr>
<td><em>Escherichia</em> coli UPEC</td>
<td>121</td>
<td>983</td>
<td>49</td>
<td>179</td>
</tr>
<tr>
<td><em>Escherichia</em> coli UTI89</td>
<td>85</td>
<td>754</td>
<td>35</td>
<td>192</td>
</tr>
<tr>
<td><em>Escherichia/Shigella</em> genera</td>
<td>15</td>
<td>184</td>
<td>15</td>
<td>113</td>
</tr>
<tr>
<td><em>Klebsiella</em> genus</td>
<td>242</td>
<td>3296</td>
<td>242</td>
<td>2090</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em> 342</td>
<td>11</td>
<td>93</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em> MGH 78578</td>
<td>21</td>
<td>237</td>
<td>14</td>
<td>49</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em> NTUH-K2044</td>
<td>339</td>
<td>2636</td>
<td>233</td>
<td>863</td>
</tr>
</tbody>
</table>
normalization was performed. A set of twelve arrays (one 12plex array) used in the experiment was printed at the same time, so background noise effects were expected to be reasonably similar across all arrays. Only one out of the twelve the results were not as anticipated. The single exception being for the *Salmonella enterica* serovar Choleraesuis isolate, which shows variation. Thus it was decided to exclude hybridization data of this isolate from further analysis. RMA normalization, performed for microarray data of the remaining eleven samples, made the distribution of probe intensities for each array in a set of arrays nearly the same.

In the workflow of further microarray data analysis, the evaluation of which genus, species, pathotype/serovar or strain, the experimental isolate is most likely to be similar to. For each of the seventy-eight gene sets, the median of signal intensities were calculated. The analysis was performed based on both distribution of probe log intensities and the signal median. The examples are shown in Figures 1-3, which visualize
the resulting plots for single representative of three chosen genera *Escherichia*, *Salmonella* and *Yersinia*. Those were *Escherichia coli* ECOR20, *Salmonella enterica* serovar Dublin and *Yersinia frederiksi*, respectively. Table 3 overviews the results for all the eleven isolates, used in the study.

Both box-and-whisker and bar plots for *Escherichia coli* ECOR20, represented in Fig. 1, show high signal intensity among the genes comprising core and *Escherichia* and-*Shigella* groups. Additionally, results show high similarity to several pathogenic *E. coli* strains, such as *Escherichia coli* CFT073, and strains of O111:H-, UPEC and EHEC pathotypes. Apart from being highly expressed among the genes belonging to *Escherichia* genus, microarray data show relatively high signal level to *Shigella* genus.
Fig. 2. Distribution of signal intensity and signal median for Salmonella enterica serovar Dublin strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

strains, thus, resulting in another proof of Escherichia and Shigella genera strains being very similar.

Fig. 2 visualizes the comparison of data for Salmonella enterica serovar Dublin isolate. Genes have high intensity values within strains belonging to Salmonella genus and core group. The highest similarity is shown to be Dublin serovar; however, DNA sequences appeared to hybridize with the high strength to Newport, Choleraesuis and Paratyphi A serovar representing probes as well.

In the case of the chosen representative for Yersinia genus, Yersinia frederiksi, results, shown in Fig. 3, are not that positive, since any obvious high intensity signal cannot be seen. This might occur as a consequence of impropriate isolation of genomic DNA, low concentration of labeled DNA, which was obviously not enough for proper hybridization to target genes, or cross-hybridization effect.
Fig. 3. Distribution of signal intensity and signal median for Yersinia frederiksii strain among the set of seventy-eight groups, mentioned previously in Table 2. a. Box-and-whisker plot, showing signal intensity distribution. b. Bar plot, showing expression signal median distribution. X-axis elements are sorted by genus, based on the order showed in Table 2. Colour code is based on the genera, where 12-colour palette represents 20 genera.

Isolates, results for which are presented in Table 3, show different chip performances. Several of them can be easily proved to belong to a particular genus, specific species and be most likely similar to a particular genus, species or serovar/serotype. However, some samples, likewise Yersinia frederiksii, do not show obvious results. This can consider the presence of uncertainties included in genomic DNA purification and sample preparation for the hybridization.
Table 3. Overview of experimental validation results

<table>
<thead>
<tr>
<th>Isolate / Distinguishing level</th>
<th>Genera</th>
<th>Species</th>
<th>Pathotype/Serovar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli ECOR20</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Salmonella enterica serovar Dublin D6</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Salmonella enterica serovar Paratyphi B var Java b</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Salmonella enterica serovar Isangi 2005-60-2087-1</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Salmonella enterica Typhimurium HN-GSS-2007-016</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Salmonella enterica serovar Choleraesuis 2870/08</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Shigella sonnei phase 12006-077</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shigella flexneri 4 2006-054</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Shigella boydii 9S</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yersinia enterocolitica O3 98-30624-5</td>
<td>-</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Yersinia ruckerii NCTC 10476</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yersinia frederiksi P963</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

‘+’ is a positive result, ‘-’ is a negative result and absence of any mark means no analysis with this purpose was made or results are not analysed.

4 Discussion and Perspective

The design of a microarray chip covering 116 bacterial genomes has proven to be a considerable challenge. Multiple aspects had to be examined, such as the number of possible sequences to be included in the database, various criteria to select the unique set of genes to particular groups of genomes, and to design probes for them. The greatest difficulty was to optimize these criteria and to filter out the false positive representative sequences for each sequence of interest. Some genera within Enterobacteriaceae, such as Escherichia and Shigella, are quite similar, thus it was difficult to find genus-specific genes. For example, the Escherichia genus appeared to have only a single gene family conserved among all the strains belonging to this genus, and being absent in the other enterics. Thus it was an obvious decision to design probes for Escherichia-and-Shigella genera-specific genes.

Along with choosing representative sequence for each of unique gene family, a problem of selecting the right organism to extract representative sequences for core-genome set became evident. In this study, core-genome genes were extracted from type species of the type genus Escherichia coli K-12 MG1655 strain. The unique sets of genes were selected on protein level, that is, similarity/dissimilarity was based on alignment using BLASTP, and gene family members were considered based on the 50/50 rule, described above. Thus this might be an explanation of why some probes did not show high intensity levels at the DNA level as was predicted.

Selecting the probes is indeed a challenging aspect. On the one hand, probes should cover all versions of the same gene, however, at the same time they should be able to distinguish between different genera, species, pathotypes/serovars, and strains. Furthermore, the array should allow various numbers of probes per gene in order to acquire the sufficient coverage of genes. Longer sequences require higher numbers of probes, whereas design of the same number of probes for short genes would result in low quality probes [36]. Therefore, the challenge is to find the best possible solution, with least time, money, and personal energy consumption.
Several improvements and suggestions could be considered for the design of an *Enterobacteriaceae* pan-genome microarray chip. To obtain more sufficient unique gene finding, searches should be done on DNA level with an appropriate cut-off value. Alignment using the BLASTN algorithm would be able to efficiently identify homologous nucleotide sequences based on similarity and would be helpful in avoiding non-specific probes.

Furthermore, for the validation of the chip step, sample preparations, such as genomic DNA isolation, labeling, and preparation to hybridize an array should be done according to protocols. Purity of DNA should be checked before the DNA labeling step to avoid small quantities of labeled DNA, which hybridizes to wrong sequences and fails to recognize the expected target sequence.

5 Conclusion

In this study, an *Enterobacteriaceae* pan-genome microarray chip was developed based on 116 genomes within this bacterial family. The typical genome size (with the exception of the reduced endosymbiont genomes of *Buchnera*, *Wigglesworthia* and *Sodalis* genera) contained between 3500 and 5500 genes. This made it possible to find at least 10 genus-, species- and pathotype/serovar-genes among all the analysed genomes. This resulted in 53644 unique probes, which were expected to hybridize to particular target sequence. High-density pan-genome microarrays can be very useful in both characterizing DNA content and monitoring expression levels for thousands of genes simultaneously. The comparison of two or more arrays can display the distinct patterns of gene expression or signal intensity level that are useful in the definition of unknown strains or genes included in these genomes. Using some experimental tests the ability of the microarray to determine bacterial strains within *Escherichia* spp., *Shigella* spp., *Salmonella* spp. and *Yersinia* spp. was demonstrated. Most of the results showed discriminative power, although some samples did not show a clear connection to the bacterial strain they are most likely to be similar to. This could be due to low quality DNA from the experiment.

It can be concluded that a *Enterobacteriaceae* pan-genome microarray, based on 116 genomes provides a perfect tool for determination of the genetic makeup of unknown strains within this bacterial family and can introduce insights into phylogenetic relationships.

Acknowledgments. This work is supported by grants from the Danish Center for Scientific Computing and the Danish Research Council. The authors would like to thank Colleen Ussery for help in editing the manuscript.

References


5.2 Paper VII. Genomic variation in Salmonella enterica core genes for epidemiological typing
Genomic variation in *Salmonella enterica* core genes for epidemiological typing

Pimlapas Leekitcharoenphon1,2, Oksana Lukjancenko2, Carsten Friis1, Frank M Aarestrup1 and David W Ussery2*

**Abstract**

**Background:** Technological advances in high throughput genome sequencing are making whole genome sequencing (WGS) available as a routine tool for bacterial typing. Standardized procedures for identification of relevant genes and of variation are needed to enable comparison between studies and over time. The core genes – the genes that are conserved in all (or most) members of a genus or species – are potentially good candidates for investigating genomic variation in phylogeny and epidemiology.

**Results:** We identify a set of 2,882 core genes clusters based on 73 publicly available *Salmonella enterica* genomes and evaluate their value as typing targets, comparing whole genome typing and traditional methods such as 16S and MLST. A consensus tree based on variation of core genes gives much better resolution than 16S and MLST; the pan-genome family tree is similar to the consensus tree, but with higher confidence. The core genes can be divided into two categories: a few highly variable genes and a larger set of conserved core genes, with low variance. For the most variable core genes, the variance in amino acid sequences is higher than for the corresponding nucleotide sequences, suggesting that there is a positive selection towards mutations leading to amino acid changes.

**Conclusions:** Genomic variation within the core genome is useful for investigating molecular evolution and providing candidate genes for bacterial genome typing. Identification of genes with different degrees of variation is important especially in trend analysis.

**Background**

With the increasing number of available bacterial genome sequences, when these genomes are compared, the genetic variation within bacterial species is greater than previously predicted [1,2]. Rapid and reliable sub-typing of bacterial pathogens is important for identification of outbreaks and monitoring of trends in order to establish population structure and to study the evolution among bacterial genomes especially within and between the outbreak strains. Today, the most widely used typing methods for bacterial genomes include multilocus sequence typing (MLST), pulsed field gel electrophoresis (PFGE), sequencing of 16S rRNA genes, and multilocus variable-number of tandem-repeat analysis (MLVA).

PFGE and MLVA have major benefits, but are time consuming and the results are difficult to standardize [3]. Other typing methods which rely on one or a few ubiquitous genes, such as the 16S rRNA gene or a set of housekeeping genes in MLST, are capable of classification at the species level and sometimes also at the subspecies level, but the biological information in a narrow selection of genes will rarely be sufficient to clearly distinguish between closely related strains such as several isolates of the same serotype [4-6]. Thus, more of the genome content should be considered rather than just one or a few genes [4].

The price and time for whole genome sequencing will soon be in the same range as the traditional typing methods mentioned above. Genome sequencing can be a powerful method in epidemiological and evolutionary investigations [7-9]. Although, to date, this has only been used in more limited epidemiological investigations where isolates suspected to be part of the same outbreak have been compared to a reference genome. In the
future, it is likely that WGS will become a routine tool for identification and characterization of bacterial isolates, as hinted at in the first ‘real-time’ sequencing of the *E. coli* O104 outbreak in Germany in the summer of 2011 [10] and the *Vibrio cholerae* outbreak in Haiti in October 2010 [11]. This requires standard procedures for identifying variation and for analyzing similarities and differences.

Conserved genes are present across bacterial genomes of the same species (or genus). A fraction of these genes—those conserved in all (or most) of the genomes of a given bacterial taxonomic group—is called the ‘core-genome’ of that group. The core-genome can be identified either within a genus or species [3] and can be used to identify the variable genes in a given genome [12]. In addition, the conserved genes in general appear to evolve more slowly, and can be used for determining relationships among bacterial isolates [13].

Currently there are more than a hundred bacterial species for which sufficient genomic data are available to estimate the species core-genome (that is, there are at least three genomes sequenced from the same species) [14]. Among these, *Salmonella enterica* is a good candidate species for conserved gene identification because the genomes are quite similar [15]. Moreover, *S. enterica* is one of the most important food-borne pathogens and is responsible for global outbreaks [16] which makes international standard typing procedures of major importance in order to allow for global comparisons [17]. The *Salmonella* genus has only two species with sequenced genomes: *Salmonella bongori* and *Salmonella enterica*. In turn, *S. enterica* is divided into 6 sub-species: *enterica*, *salaarna*, *arizonae*, *diarizone*, *houstenae* and *inda*. Presently, *S. enterica* is classified into more than 2,500 serotypes [18].

In order to investigate an outbreak caused by *Salmonella*, characterization of *Salmonella* isolates from genome data is a crucial step. *Salmonella* genomes are highly similar, particularly within subspecies *enterica*, where little variance exists in the genomes [15]. This high similarity presents a challenge for typing and classification.

In their pioneering work Tettelin et al. [1] defined the core genes of a species by being those genes found present in (nearly) all known members of the species. Since then others have studied core and pan genomes at the genus level or even at the kingdom level [19], but for our purposes the original definition at the species level is suitable. In this work we identify the core genes within *S. enterica* genomes and determine variation between the different available genomes, both in terms of sequence and presence/absence of non-core genes; in the latter case using a method originally published by Snipen & Ussery [20]. We evaluate the value of different approaches for classification of isolates in epidemiological settings and compare our findings to currently used sequencing methods, both in long term trend analysis and outbreak investigations.

**Results and discussion**

The 73 *Salmonella* genomes used in this study are summarized in Additional file 1: Table S1. The set comprises 21 completed genomes and 52 nearly completed genomes. Of these, 35 genomes are closely-related *S. Montevideo* strains pertaining to an outbreak of salmonellosis from Italian-style spiced meat [21]. All genomes were retrieved from GenBank [22] except *S. Typhimurium* str. DT104, which was received from the Sanger Institute’s bacterial genome database. All *Salmonella* genomes are from sub-species *enterica* with the exception of the single *S. enterica* subsp. *Arizonae*.

**Evaluation of traditional bacterial sequence-based typing**

The ribosomal genes are essential for the survival of all cells, and their structure cannot change much because of their involvement in protein synthesis [23]. Thus, 16S rRNA genes are highly conserved among isolates belonging to the same bacterial species [4]. Exceptions may be *N. meningitidis* [24] and *Mycoplasma* [25]. However, due to limited variation within a given species, the 16S sequencing is often not useful for epidemiological studies, where the classification of highly similar strains is needed. Jacobsen et al. shows a phylogenetic tree based on 16S rRNA genes, extracted from 26 *Salmonella enterica* genomes, using RnAmmer [15,26]. As expected, there is not sufficient resolution to distinguish among the *Salmonella* subspecies *enterica*.

Genes such as *rpoB* or *sodA* have been suggested as substitutes for 16S rRNA and have shown improved efficacy in species identification [27], although it remains unlikely that a single gene can always reflect the subtle differences between genomes of the same species.

The limitations of using a single gene may be improved by the simultaneous analysis of multiple genes. Multi Locus Sequence Typing (MLST) has found wide applications, especially in phylogenetic studies and is most commonly based on seven housekeeping genes - each bacterial species having its own set. For *Salmonella* these are: *aroC*, *dnaN*, *hemD*, *hisD*, *purE*, *sucA* and *thrA* http://www.mlst.net. A MLST tree, based on an in silico analysis of the 73 available *Salmonella enterica* genomes in Genbank, is shown in Figure 1. Strains of the same serovar generally cluster into distinct groups, although exceptions exist; for example the *S. Weltevreden* str. HI_N05-537 is mixed with *S. Montevideo*. Furthermore, recent work on 61 sequenced *E. coli* genomes [4], found that the 16S rRNA tree cannot resolve well within the genus level and also that MLST cannot differentiate pathogenic strains from non-pathogenic strains. Still, MLST has proven useful for long-term analysis of population structures, but often fails
Figure 1 In silico MLST tree. Seven housekeeping genes were extracted from Salmonella genomes. Concatenated sequences were aligned by MUSCLE. The phylogenetic trees were generated by MEGA5 using bootstrap maximum likelihood method. Each color represents a different serogroup (O antigen). The confidence value is the bootstrap value calculated by sampling with replacement from the multiple sequence alignment.
to detect differences between closely related strains [28]. Indeed, improved MLST schemes that include more than 7 genes have been suggested [4].

For *Salmonella*, sequencing specific short repeats and virulence genes have recently been suggested as an alternative and improved method for typing of *S. Enteritidis* [29]. The usefulness of this approach in epidemiological studies and typing is currently unknown, although the choice of repeats must be tailored for the specific bacterial species studies.

### Identification of core genes

Determining gene conservation across multiple genomes is not overly difficult, but certain choices must be made which will affect the final outcome. Using a previously published method [20,30,31] which employs single-linkage clustering on top of BLASTp alignments, sets of pan- and core-genomes were estimated, based on all 73 *Salmonella* genomes. The progression of the pan- and core-genomes is shown in Figure 2A. The number of novel gene clusters in the pan-genome gradually increases when more genomes are considered, while the number of conserved gene clusters constituting the core genome decreases slightly. When all *Salmonella* genomes have been considered, there are 10,581 pan gene clusters and 2,882 core gene clusters (Additional file 2) in species *enterica*. In the step going from *S. Typhimurium* to *S. Typhi*, the number of core genes drops suddenly, most likely because the *S. Typhi* genome has undergone considerable pseudogene formation resulting in gene loss [32]. The number of core genes drops again when adding a genome of the subspecies *arizonae* which is associated with cold-blooded animals. This technique has previously been applied successfully in finding core genomes for Proteobacteria genera *Burkholderia* [33], *Escherichia coli* [4], *Vibrio* [34] and *Campylobacter jejuni* [30], as well as Bacteroides [35] and Lactic acid bacteria [36].

### Genomic variation within the core genes

The core genes as calculated above were used for constructing a gene variation plot by performing all-against-all BLAST alignments between 2,882 core gene clusters and all 73 *Salmonella enterica* genomes. The resulting average identities within each core gene cluster is displayed in Figure 2B. From this figure, the average percent identity was very high (> 98%) in most of the core genes, but dropped sharply for around 5% of the core genes. From this plot, the identified core genes can be divided into two categories: a small group of highly variable genes and the majority of genes which show little variation.

For the highly variable core genes, the variation in amino acid sequences (Figure 2B, green dots) was higher than for the nucleotide sequences (Figure 2B, red dots), whereas the opposite was the case for the more conserved core genes. This indicates that for core genes with low variation there is a selection against mutations leading to amino acid changes, whereas for the highly variable genes, positive selection for amino acid changes seems to be the case. In order to confirm these hypothesis, the approximation of dN/dS has been performed by dividing the number of non-synonymous changes per non-synonymous sites with the number of synonymous changes per synonymous sites [37] using *S. Typhimurium* str. LT2 as a reference genome. The median dN/dS ratio for conserved and highly variable core genes are 1.0 and 1.25 respectively. Therefore, the amino acid changes in highly variable core genes might be due to an increase in positive selection at some sites. Nonetheless, the importance of this needs to be confirmed by additional analysis, although one could imagine, for example, a selective pressure to vary the surface proteins to avoid immune response.

The seven genes used for MLST are marked in the Figure 2B, and are scattered throughout the highly conserved part of the core genes (Figure 2B, black dots) and, as expected, little variation exists in these genes. Including core genes from both the highly conserved and variable regions might be beneficial in evolution studies. On the one hand, the more slowly evolving genes are useful in distinguishing between divergent and convergent evolution, while faster evolving genes can help in strain identification.

### Functional analysis of conserved genes

In order to determine the functional profile of core genes, the core gene clusters were aligned against UniProt [30]. Functional profiles were determined based on Gene Ontology (GO) terms and visualized in Figure 3. Though the difference is generally small, some terms common in conserved core genes tend to be less frequent in highly variable core genes; for example, electron carrier activity, structural molecule activity and metallochaperone activity. These functions are essential for living cells and are therefore enriched in conserved core genes. On the other hand, highly variable core genes encode many proteins that are associated with the extracellular region. In general, genes located outside the cell are known to be more variable [38].

### Consensus tree based on core gene clusters

Figure 4 shows a phylogenetic tree generated from the sequence of all 2,882 *Salmonella* core gene clusters. The tree generally divides the serotypes up well, but the bootstrap value in several branches is very low. This uncertainty could be due to the large number of core gene trees being analyzed individually; the low bootstrap values near the root reflect a lack of consensus at the higher levels. In contrast, the low bootstrap values found in *S. Montevideo* strains likely reflect uncertainty due to the high similarity of gene sequence of the clonal
Figure 2 Pan-core-genome plot and variation plot. (A) Pan- and core-genome plot of 73 *Salmonella enterica*. The plot shows an increase of the pan-genome (blue line) and a decrease of the core-genome (red line) as more genomes are added. The last points show the total number of gene clusters in the pan-genome and the core-genome. (B) Variation plot. This plot shows the variation within core gene clusters in amino acid levels (green dots) and nucleotide levels (red dots). Black dots show the distribution of housekeeping genes in the core genes. The Y- and X-axes represent average percent identity and numerical core gene cluster name respectively.
outbreak. All S. Montevideo strains sequenced were from a single outbreak [21] and as expected this analysis confirmed the almost complete identity of these isolates.

A previous study described that there are 69 genes unique to Salmonella [39]. Instead of using all core genes, we generated a consensus tree based on these 69 Salmonella-specific genes (Additional file 3: Figure S1). We also constructed an additional four consensus trees based on sets of 69 core genes randomly picked from different areas in the variation plot (Figure 2B): from a mixture of high, medium and low variable core genes (Additional file 4: Figure S2), from medium variable core genes (Additional file 5: Figure S3), from highly variable core genes (Additional file 6: Figure S4) and from the area where the curve decreases in the variation plot (Additional file 7: Figure S5). The appearance of these 5 consensus trees was similar to the tree from Figure 4, with two exceptions: the trees based on the 69 specific genes (Additional file 3: Figure S1) and the highly variable core genes (Additional file 6: Figure S4). In the former, S. arizonae, which is not part of the subspecies enterica, was still mixed in with other enterica, while for the latter, S. Agona str. SL483 clustered away from the other subspecies enterica. Thus, based on these results, it appears that using only Salmonella unique genes or highly variable genes does not provide phylogenetically useful information and should probably not be used for future WGS studies. Comparisons using more genomes in more species can further test this.
Pan-genome tree

In principle, genome similarity is not only measurable by shared genes, but also by the absence of genes. Figure 5 is another tree, based on gene presence/absence across all the *Salmonella* genomes [20]. This tree bears a striking resemblance to the consensus tree based on core genes (Figure 4), although the bootstrap values are higher in many of the branches, especially near the root. Of all methods investigated in this study, the pan-genome tree presents itself as the best solution for a tree that can resolve strain differences in a biologically meaningful way, even if it would be expected to correlate more with phenotype than phylogeny. It is, however, important to note that creating pan-genome trees requires higher quality sequencing data and assemblies than what are typically obtained using short reads from second-generation sequencing methodologies. Even so, we have found that pan-genome trees with good correspondence to known bacterial types can be constructed from Solexa data (100 bp reads), if care is taken to ensure good assembly and gene finding (data not shown).

The power to discriminate between variants differs between the methods used. The phylogenetic analysis for the MLST tree is based on the identified informative sites among the seven housekeeping genes, for the pan-genome tree on presence and absence of genes and for the consensus tree based on the informative sites of core gene clusters from alignments of all core genes. The number of informative sites for *in silico* MLST tree, pan-genome tree and consensus tree based on core gene clusters were...
877 bp (10,008 total base-pairs in the seven genes), 7,699 genes (10,581 total genes) and 880,832 bp (2,868,821 bp in all core genes), respectively. The pan genome and core gene analysis were based on much more variation than the MLST analysis and have a much stronger power to discriminate closely related strains.

Conclusions

Bacterial typing should provide meaningful information for both epidemiological and evolutionary studies. For epidemiology, the ability to differentiate unrelated isolates (discriminatory power) and the ability to cluster related isolates are crucial. 16S rRNA and the MLST genes rarely provide separation between closely related strains. The performance of the pan-genome tree, however, is valid for epidemiological investigation in both discriminatory and clustering abilities. One caveat is that this method depends on good quality genomic data.

Comparative genomics can determine the conserved genes (core-genome) among bacterial genomes at either
genus or species level. Genomic variation within the core-genome can then be used to reveal highly variable genes (fast evolving genes) and conserved genes (slow evolving genes). These core genes are useful for investigating molecular evolution and remain useful as candidate genes for bacterial genome typing—even if they cannot be expected to differentiate highly similar isolates from e.g. outbreak cases, such is not always desirable. Even in cases where a deeper distinction of isolates is of interest, e.g. in mapping outbreaks, core genes might still be useful as a reference fragment for SNPs calling instead of using whole genome analysis. However, in term of computational costs, the consensus tree based on core genes requires more computational time than the other methods.

In the near future, global real-time surveillance of Salmonella and other pathogens giving simultaneous information on population structure and evolution, as well as outbreak detection, may well be possible.

Methods
Salmonella genome data and gene annotation
From public genome databases (NCBI and Sanger Institute’s bacterial genome databases), 83 Salmonella enterica genomes available at the time (April, 2011) were downloaded. These genomes consisted of 21 completed genomes and 62 draft genomes. Due to the large number of contigs in some genomes, only 73 genomes were selected for this study (Additional file 1: Table 1). The gene finder Prodigal was used on DNA sequences of all genomes to eliminate biases in annotation quality and to standardize the genes found in all genomes [15]. Gene clusters were then inferred according to [15,20,30]

In silico MLST trees
The in silico MLST tree was constructed from seven housekeeping genes: aroC, dnaN, hemD, hisD, purE, sucA and thrA http://www.mlst.net. These genes were extracted from Salmonella genomes and concatenated. The concatenated sequences were aligned using MUSCLE [40]. Phylogenetic trees were generated by MEGA5 using the maximum likelihood method [41]. The confidence value is, in this case, the same as the bootstrap value, calculated by sampling with replacement from the multiple sequence alignments [42]. Thus, the in silico MLST differs from traditional MLST in that complete genes are used and not just the MLST alleles. However, since the alleles typically cover the majority of the genes, the difference is small.

Consensus trees
All core gene clusters from 73 Salmonella genomes were used for generating a consensus tree. Multiple alignments for each core gene cluster from all strains were performed using MUSCLE [40]. A phylogenetic tree for each core gene was generated using PAUP [43]. The Phylip package was used to construct the consensus tree from all the trees [44]. The bootstrap values are shown in the consensus tree.

GO annotation
The core gene clusters were compared in an all-against-all BLAST with protein sequences from UniProt based on the ‘50/50 rule’ [30]. Functional profiles were summarized from BLAST results by mapping UniProt IDs to Gene Ontology (GO) terms. Mapping GO parental terms were performed using publicly available GO-PERL modules for searching through a graph structure of ontology data [45,46]

Pan-genome trees
The Pan-genome matrix consists of gene clusters (rows) and genomes (columns). The absence and presence of genes across genomes are represented by 0’s and 1’s respectively. The relative Manhattan distance between genomes was calculated and used for hierarchical clustering. The bootstrap values are calculated in order to represent the confidence of branches [20].

Additional material

Acknowledgements
This study was supported by the Center for Genomic Epidemiology (09-067103/DSF) http://www.genomicepidemiology.org and by grant 3304-FVFP-08- from the Danish Food Industry Agency. PL and OL would like to acknowledge funding from the Technical University of Denmark. The authors would like to thank Colleen Ussery for editorial assistance in preparing the manuscript.

Author details
1 National Food Institute, Building 204, The Technical University of Denmark, 2800 Kgs Lyngby, Denmark. 2 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Building 208, DK-2800 Kgs Lyngby, Denmark.
Authors' contributions
PL planned the study, carried out all bioinformatics analysis and drafted the manuscript. OL participated in consensus tree based on core genes. CF participated in the planning of the study, the core genes identification and drafted the manuscript. FMA supervised and planned the study and drafted the manuscript. DWU supported the supervision, participated in the design of the study and drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 30 September 2011 Accepted: 12 March 2012 Published: 12 March 2012

References


5.3 TaxonomyFinder web-server

The pan-genome of a given taxonomic group of genomes (phylum, genus, species) consists of a set of conserved proteins, proteins that are present in some, but not all genomes, or specific for certain strains. Taxonomy is usually predicted using evolutionary conserved genes, such as 16S rRNA, a set of seven ‘housekeeping’ genes in MLST, or the ribosomal proteins for rMLST. However, taxa group-specific proteins can be also used to infer taxonomic identification. To address this assumption, a new approach, TaxonomyFinder, is introduced in this PhD thesis. Taxonomy group-specific proteins are extracted using Pan-FunPro tool, described earlier. Briefly, homologous proteins from all the analysed genomes are grouped into protein families, based on functional profiles (combinations of functional profiles). Later, taxa group-specific profiles are predicted. Profile is considered to be specific, if it is 100% conserved within set of query genomes, and is absent in the rest of analysed genomes. However, it may be not possible if the number of members in taxonomic group is large, such as Proteobacteria, Firmicutes phyla, or Escherichia genus. In this case, the threshold is lowered, meaning that profiles are still specific to that taxonomic group, but can be absent in several genomes within the group.

TaxonomyFinder method is publically available as a web-based tool (http://cge.cbs.dtu.dk/services/TaxonomyFinder/). Taxonomy can be predicted on phylum and species level. The database includes 33 phylum-specific and 1242 species-specific profile sets. Brief instructions are shown on Figure 5.1. The first step is to upload the genome of interest. An input file can be uploaded in three formats: Genbank format, assembled genome, or already predicted protein sequences. After the taxonomy level is specified, the job can be submitted.

The prediction output is shown both as on-screen results and downloadable files. An example is shown in Figure 5.2. The on-screen results output depend on whether prediction of phylum or species was performed. Species prediction output will also include phylum information. Prediction score is the fraction of matched profiles to the total taxa group-specific profiles and ranges between 0 and 100%. A prediction score between 0 and 10 is considered very poor and is coloured in red, while grey to green gradient colour intensity indicates prediction score between 10 and 100%, where 100% is the best prediction.
Downloadable files include the table with all the predictions, sorted by the best score; and input genome annotation, performed by PanFunPro approach. In case of species level, whether phylum prediction is not predicted poorly, species search is narrowed to the species of the predicted phylum.

Figure 5.1: Submission of isolate to the TaxonomyFinder server
Figure 5.2: TaxonomyFinder prediction output. A. Phylum prediction output. B. Species prediction output.
5.4 Paper VIII. (Manuscript). Benchmarking of Methods for Genomic Taxonomy
Benchmarking of Methods for Genomic Taxonomy.

Mette Voldby Larsen¹, Salvatore Cosentino¹, Oksana Lukjancenko¹, Dhany Saputra¹, Simon Rasmussen¹, Henrik Hasman², Thomas Sicheritz Pontén¹, Frank M. Aarestrup², David Wayne Ussery¹,³ and Ole Lund¹

¹Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
²National Food Institute, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
³Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Abstract

One of the first questions that emerge when encountering a prokaryotic organism of interest is what it is and which species it is. The 16S rRNA gene formed the basis of the first method for sequence-based taxonomy and has had a tremendous impact on the field of microbiology. Nevertheless, the method has been found to have a number of shortcomings.

In the current study we trained and benchmarked five methods for whole genome sequence based prokaryotic species identification on a common dataset of complete genomes; 1) SpeciesFinder, which is based on the complete 16S rRNA gene, 2) Reads2Type that searches for species-specific 50-mers in either the 16S rRNA gene, the GyrB gene (for the Enterobacteraceae family) or the ITS gene (for the Mycobacterium genus), 3) The rMLST method that samples up to 53 ribosomal genes, 4) TaxonomyFinder, which is based on species-specific functional protein domain profiles, and finally 5) KmerFinder, which examines the number of co-occurring k-mers. The performances of the methods were subsequently evaluated on three datasets of short sequence reads or draft genomes from public databases. In total, the evaluation sets constituted more than 11,000 isolates covering 159 genera and 243 species. Our results indicate that methods that only sample chromosomal, core genes have difficulties in distinguishing closely related strains, which only recently diverged. The KmerFinder method had the overall highest accuracy and identified from 93%-97% of the isolates in the evaluations sets correctly to the species level.

Importance: The 16S rRNA locus has served as the backbone of prokaryotic taxonomy for more than 30 years, but has been recognized to be less than optimal for a number of species. The current advent of whole genome sequencing provides the opportunity to surpass 16S rRNA typing by including a larger fraction of the genome. Meanwhile, the ample amounts of WGS data in public databases enable us to perform educated proposals on how to optimally use this type of data.

INTRODUCTION

Rapid identification of isolated bacterial species is essential for surveillance for human and animal health and for choosing the optimal treatment and control measures. Since the be-

*Corresponding author, e-mail: mette@cbs.dtu.dk
ginning of microbiology more than a century ago, this has to a large extent been based on morphology and biochemical testing. However, for more than 30 years, 16S rRNA sequence data has served as the backbone for the classification of prokaryotes (1) and tremendous amounts of 16S rRNA sequences are available in public repositories (2; 3; 4). However, due to the conserved nature of the 16S rRNA gene, the resolution is often too low to adequately resolve different species and sometimes not even adequate for genus delineation (5; 6). Furthermore, many prokaryotic genomes contain several copies of the 16S rRNA gene with substantial inter-gene variation (7; 8). It is also considered problematic that this gene represents only a tiny fraction, roughly about 0.1% or less, of the coding part of a microbial genome (9).

Second- and third generation sequencing techniques have the potential to revolutionize the classification and characterization of prokaryotes. However, so far no consensus on how to utilize the vast amount of information in Whole Genome Sequence (WGS) data has emerged. Nevertheless, a number of different methods have been proposed. Roughly, they can be divided into those that require annotation of genes in the data and those that employ the nucleotide sequences directly.

One of the first attempts to employ WGS data for taxonomic purposes was carried out in 1999 (10). At the time, 13 completely sequenced genomes of unicellular organisms were available and distance-based phylogeny was constructed on the basis of presence and absence of suspected orthologous (direct common ancestry) gene pairs. Later it was recognized that methods that take into account gene content can be greatly influenced by Horizontal Gene Transfer (HGT) and alternative methods were developed that used homologous groups (gene family content) (11) or protein domains (12). Functional protein domains also form the basis of a recent approach developed by our group (13). Here, the protein domains are combined into functional profiles of which some are species-specific and can thus be used for inferring taxonomy.

As an extension of 16S rRNA analysis, which focuses on a single locus, Super Multilocus Sequence Typing (SuperMLST) has been proposed (14). It relies on the selection of a set of genes that are highly conserved and hence can be used with any organism. In a publication from 2012, Jolley et al. suggested that 53 genes encoding ribosomal proteins are used for bacterial classification in an approach called ribosomal MLST (rMLST) (15). Not all 53 genes were found in all bacterial genomes, but due to the relatively high number of sampled loci, this is not considered as problematic. The rMLST method forms the basis of a proposed reclassification of Neisseria species (16) and has also been used for analyzing human Campylobacter isolates (17).

It is also possible to employ the sequence data directly without pre-annotation of genes. This can, for instance, be done by looking at k-mers (substrings of k nucleotides in DNA sequence data) that are sufficiently long to avoid co-occurrence in two random genomes. As an example, there are more than 4 billion different possible 16-mers, making their co-occurrence in two unrelated bacterial genomes unlikely. The number of co-occurring k-mers in two bacterial genomes can thus be considered a measure of evolutionary relatedness, and used to construct a phylogeny. Using this approach, all regions of the genome are considered, not only core genes. Furthermore, a gene segment will score highly despite the transposition of a gene segment within the genome, since only the flanking regions will be mismatched.

In the current study we have trained five different methods for species identification on a common dataset of complete prokaryotic genomes. 1) SpeciesFinder serves as the baseline, as it is based solely upon the 16S rRNA gene. 2) Reads2Type is a variant hereof, searching for species-specific 50-mers, predominantly within the 16S rRNA gene, with the help of non-species-specific 50-mers to quickly narrow down the search. 3) rMLST, which predicts species by examining 53 ribosomal genes. 4) TaxonomyFinder, which is based on species-specific functional protein domain profiles, and finally 5) KmerFinder, which predicts species
by examining the number of overlapping 16-mers.

The public available databases contain ample amounts of WGS data from prokaryotes, enabling us to conducting a large-scale benchmark study of the proposed methods. Hence, the process of reaching a consensus on how the WGS data should optimally be used for prokaryotic taxonomy is initiated.

MATERIALS & METHODS

Dataset

Training Data

In August 2011 a total of 1,647 complete genomes originating from Bacteria (1,535) and Archaea (112) were downloaded from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov-genome). For each genome, the annotated taxonomy according to GenBank was compared to the taxonomy according to Entrez, which was retrieved using the taxonomy module of BioPerl. Discrepancies were checked and corrected manually. For each genome, it was also examined if the annotated name was in accordance to the List of Prokaryotic names with Standing in Nomenclature (http://www.bacterio.cict.fr/allnames.html). When possible, names that were not in accordance were corrected to valid ones. In this way, 1,426 genomes were assigned to 847 approved genus and species names. The remaining 221 genomes, which were either only assigned to a genus, e.g., Vibrio spp., or assigned to species with informal names, e.g., Synechoccus islandicus, were left in the training data under the assumption that they will influence the different methods for species identification equally. An overview of the training data is available in Supplementary Table 1.

Evaluation Data

Three datasets were generated for the purpose of evaluating the methods. The first consisted of assembled complete of draft genomes with assigned species, which were downloaded from NCBI in September 2012 and not already part of the training data. Only genomes assigned to species that were also present in the training data were included. The set is called NCBI drafts and consists of genomes from 695 isolates covering 81 genera and 149 species. The set includes three Archaea; two Methanobrevibacter smithii and one Sulfolobus solfataricus. An overview of the data can be seen in Supplementary Table 2.

Furthermore, In January 2012, 11,768 sets of Illumina raw reads were downloaded from the NCBI Sequence Reads Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) with assigned species (18). 10,517 of them had been sequenced by the Illumina Genome Analyzer II sequencer, while the remaining 1,251 had been sequenced by the Illumina HiSeq 2000 sequencer. Reads that could not be assembled to a draft genome were removed as were reads from species that were not present in the training. The final SRA reads dataset consists of 8,798 sets of paired-end reads and 1,609 sets of single reads, 10,407 sets in total.

The short reads of the SRA reads set were de novo assembled using velvet 1.1.04 (19). For of the draft assemblies the optimal k-mer length was estimated and used as described previously (20). The resulting set of draft genomes constitutes the SRA drafts evaluation set. To measure the qualities of the draft assemblies, the N50 values were calculated (21). The draft assemblies had an average N50 of 77,018, ranging from 101 to 779,945 (see Supplementary Figure 1), an average number of scaffolds of 697, and an average size of 3,301
kilobases. The SRA\textsubscript{reads} and SRA\textsubscript{drafts} sets both cover 167 different species from 120 genera with more than 5,000 strains from the \textit{Streptococcus}, \textit{Staphylococcus} and \textit{Salmonella} genera. There are no species from Archeae. An overview of the SRA\textsubscript{reads} and SRA\textsubscript{drafts} sets is available in Supplementary Table 3.

### Methods for species identification

#### SpeciesFinder

SpeciesFinder predicts the prokaryotic species based on the 16S rRNA gene. A 16S database was built from the genomes of the common training data using RNAmmer (22). The species predictions were performed differently depending on the input type. If the input was short reads, the prediction was done as follows:

I. The reads were mapped against the 16S database using the Burrows-Wheeler aligner (BWA)(23).

II. The BWA output was assembled using Trinity (24) to obtain the 16S rRNA sequences.

III. The BLAST algorithm (25) was used to search the output from Trinity against the 16S database.

IV. The best BLAST hit (see below) was chosen and the species associated with the best hit was given as the final prediction.

When the input sequence was a draft or complete genome, the prediction was performed as follows:

I. The 16S rRNA gene was predicted from the input sequences using RNAmmer.

II. Using the BLAST algorithm, the predicted sequence was aligned against the 16S database.

III. The best BLAST hit (see below) was chosen and the species associated with it given as the final prediction.

The best BLAST hit was chosen by ranking the output from the BLAST alignment by a combination of coverage, percent identity, bitscore, number of mismatches, and number of gaps. The highest ranked hit was chosen for the prediction.

SpeciesFinder is available at \url{http://cge.cbs.dtu.dk/services/SpeciesFinder/}.

#### rMLST

The rMLST method predicts bacterial species based on 53 ribosomal genes originally defined by Jolley \textit{et al.} (15). The set of genes can either be used in an approach similar to Multilocus Sequence Typing (MLST), where each locus in the query genome is considered identical or non-identical to alleles of the corresponding locus in the reference database, and an allelic profile based on random numbers assigned to each of the alleles in the database is generated accordingly. Since the strains that we compare are more diverse than the ones compared in MLST, it is likely that many loci would have no identical matches in the database, making a simple cluster analysis based on allelic profiles problematic. To improve the resolution of the method, in our implementation of rMLST, the nucleotide sequence of each locus is aligned to the alleles in the reference database and a measure of the similarity of the locus and the
best matching allele is used subsequently, as described below.

Briefly, for each of the genomes in the training data, the 53 ribosomal genes were provided by Keith Jolley, Department of Zoology, University of Oxford, UK. In this way, for each genome, a gene collection of up to 53 ribosomal genes was assigned. To predict the species of a query genome, the query genome was first aligned to each gene collection using BLAT (26). Only hits with at least 95% identity and 95% coverage were considered as a potential match. If there were several potential matches, the best match was selected based on the best cumulative rank of coverage, percent identity, bitscore, number of mistmatches, and number of gaps in the alignments. The final prediction was given as the organism with the highest number of best hits across all genes. Our implementation of rMLST performs predictions for draft or complete genomes, but not short reads.

**TaxonomyFinder**

The TaxonomyFinder method is based on taxonomy group-specific protein profiles (ref). It performs predictions for draft or complete genomes, but not for short reads. The common training data was used to create the taxonomy-specific profile database. Briefly, for each genome functional profiles were assigned based on three collections of Hidden Markov Models (HMMs) databases: PfamA (27), TIGRFAM (28), and Superfamily (29). Genes that did not match any entry in the HMM databases were clustered using CD-HIT (30). Further, genomes were grouped according to the taxonomy level, either phylum or species, and profiles that were specific to each taxonomic group were extracted. Profiles were considered specific to a taxonomic group, if they were conserved in most of the genomes within a phylum/species group and absent in all genomes outside of the group. The workflow of the TaxonomyFinder method is a four-step process, which includes:

I Open-reading frame prediction using Prodigal (31).

II Construction of functional profiles from protein-coding sequences.

III Assignment of functional profiles.

IV Functional profile comparison to the taxonomy-specific profile database. The number of architectures, matched to each of the taxonomy groups, is recorded, and the fraction of taxa-specific genes (score) is calculated. The best-matching taxonomy group is selected based on a consensus of the best score and highest number of matched architectures.

TaxonomyFinder is available at [http://cge.cbs.dtu.dk/services/TaxonomyFinder/](http://cge.cbs.dtu.dk/services/TaxonomyFinder/).

**KmerFinder**

The KmerFinder method predicts prokaryotic species based on the number of overlapping (co-occuring) k-mers, i.e. 16-mers between the query genome and genomes in a reference database. Initially, all genomes in the common training data were split into overlapping 16-mers with step-size one, meaning that if the first 16-mer is initiated at position N and ends at position N+15, the next 16-mer is initiated at position N+1 and ends at position N+16, and so on. To reduce the size of the final 16-mer database only 16-mers with the prefix ATGAC were kept. These 16-mers were stored in a hash table with links to the original genomes. When performing the prediction, the species of the query genome is predicted to be identical to the species of the genome in the training data with which
it has the highest number of 16-mers in common regardless of position. The input for KmerFinder can be draft or complete genomes as well as short reads. KmerFinder is available at http://www.cbs.dtu.dk/services/KmerFinder/.

**Reads2Type**

Reads2Type identified the prokaryotic species based on a database of 50-mers generated from chosen marker genes (Saputra D., Rasmussen S., Larsen M.V., Haddad N., Aarestrup F.M., Lund O., and Sicheritz-Pontén T., submitted for publication). The version of Reads2Type evaluated in this study requires short reads as input. For bacterial species not belonging to the *Enterobacteriaceae* family or the *Mycobacterium* genus, the 50-mer database relies on the 16S rRNA locus, while for *Enterobacteriaceae*, the gyrB locus is used, and for *Mycobacterium* the ITS locus. Briefly, the following steps were applied for building the 50-mer probe database:

I 16S rRNA sequences of the complete bacterial genomes of the common training set were predicted using RNamm (22).

II For species belonging to the *Enterobacteriaceae* family or the *Mycobacterium* genus, gyrB sequences and ITS sequences, respectively, were downloaded from NCBI.

III The above sequences were pooled and all possible 50-bp fragments were generated from that pool.

IV 16S rRNA probes unique for *Enterobacteriaceae* and *Mycobacteria* were removed from the pool of 50-mers.

V All 50-mer duplicates associated to the conserved regions of different strains but the same species were removed.

VI To further reduce the size of the final 50-mers database, 25 consecutive 50-mers previously fragmented from one 50 bp stretch of 16S rRNA belonging to the same list of organism were removed.

The resulting 50-mers probe database consists of a number of sequences found uniquely in one species, as well as other sequences shared between several species. Subsequently, each read was compressed into a suffix tree, which is a data structure for fast string matching. The compressed short reads were aligned to the 50-mer probe database using a "narrow-down approach" strategy, i.e. when a compressed read matched a probe belonging to a group of species, a much smaller probe database excluding other species was created on the fly, causing the read progress to be faster and the species to be identified much faster.

The Reads2Type method is available as a web server (http://cge.cbs.dtu.dk/services/Reads2Type/) and as a console. The web-based Reads2Type is unique in not requiring the short read file to be uploaded to the server. Instead, the 4.6 MB 50-mers probe database is automatically transferred into the client computers memory before initiating the species identification. All computations needed for the species identification is fully performed on the clients computer, minimizing the data transfer and avoiding the network bottleneck on the server.
Testing the speed

The speed of the methods was evaluated on non-published internal data from up to 450 strains covering eight species (Enterococcus faecalis, Enterococcus faecium, Eschericia coli, Escherichia fergusonii, Klebsiella pneumoniae, Salmonella enterica, Staphylococcus aureus, and Vibrio cholera) that had been sequenced by the Illumina sequencing method. Draft genomes were de novo assembled as described above for the SRA_{drafts} set. The speed was tested on a Cluster with x86_64 architecture, 128 nodes, 4 tasks per node, 30 or 7G per node.

RESULTS

Performances on NCBI draft genomes

The SpeciesFinder, rMLST, TaxonomyFinder, and KmerFinder methods are able to perform species predictions on draft or completed prokaryotic genomes. Their performances were evaluated on the NCBI_{drafts} set of 695 draft genomes covering 149 species. Supplementary File 1 lists all predictions, while Figure 1A summarizes the results. Overall, SpeciesFinder, which is based on the 16S rRNA gene, had the poorest performance, only correctly identifying 76% of the isolates down to species level. KmerFinder, which is based on co-occurring 16-mers, had the highest performance and correctly identified 93% of the isolates. For only three isolates (0.43%), KmerFinder did not even get the genus correct. These three isolates were two Escherichia coli predicted as Shigella sonnei and one Providencia alcalifaciens predicted as Yersinia pestis.

The NCBI_{drafts} set contains three Archaeal isolates; two M. smithii and one S. solfataricus. SpeciesFinder, TaxonomyFinder, and KmerFinder predicted the species of all three isolates correctly, while rMLST, which was only intended for characterization of Bacteria (15) predicted the M. smithii correctly, but was unable to make a prediction for the S. solfataricus.

The overlap in predictions of the four methods was examined and illustrated in Figure 2A. All four methods correctly identified 428 out of 695 isolates (62%), and all methods misidentified the same six isolates. Table 1 lists these six isolates. Since all four methods agreed on these predictions, the isolates are likely to be wrongly annotated. Alternatively, the annotations of the isolates in the training data that the predictions were based on are incorrect.

As seen in Figure 2A, isolate predictions agreed upon by several methods are more accurate that predictions unique to a particular method. However, the KmerFinder method made unique predictions for 36 isolates of which 20 were in concordance with the annotation.

Predictions for the most common species in the dataset were examined more closely and illustrated in Figure 3 and in Supplementary Figure 2-5. In general, the wrong predictions by SpeciesFinder (that is, the ones that were in disagreement with the NCBI annotation) were typically scattered, often consisting of a few wrong predictions of each type. The rMLST method was, on the other hand, more consistent in its incorrect predictions. As an example, the rMLST method wrongly annotated all 14 Bacillus anthracis isolates as Bacillus thuringiensis, all 8 Brucella abortus as Brucella suis, and all 6 Burkholderia mallei as Burkholderia pseudomallei. In general, all four methods had difficulties identifying species within the Bacillus genus, such as isolates annotated as B. thuringiensis, but predicted to be Bacillus cereus or vice versa. Another mistake common to all methods was Streptococcus miltis being predicted as Streptococcus oralis or Streptococcus pneumoniae. Also, none of
<table>
<thead>
<tr>
<th>RefSeqID</th>
<th>Strain name</th>
<th>Annotated species</th>
<th>Predicted species</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ_ACLX00000000</td>
<td>AH621 uid55161</td>
<td>Bacillus cereus</td>
<td>Bacillus weihenstephanensis</td>
</tr>
<tr>
<td>NZ_ACMD00000000</td>
<td>BDRD ST196 uid55169</td>
<td>Bacillus cereus</td>
<td>Bacillus weihenstephanensis</td>
</tr>
<tr>
<td>NZ_ABDQ00000000</td>
<td>C Eklund uid54841</td>
<td>Clostridium botulinum</td>
<td>Clostridium novyi</td>
</tr>
<tr>
<td>NZ_ABXZ00000000</td>
<td>FTG uid55313</td>
<td>Francisella novicida</td>
<td>Francisella tularensis</td>
</tr>
<tr>
<td>NZ_AHIE00000000</td>
<td>DC283 uid86627</td>
<td>Pantoea stewartii</td>
<td>Pantoea ananatis</td>
</tr>
<tr>
<td>NZ_AEPO00000000</td>
<td>ATCC 49296 uid61461</td>
<td>Streptococcus sanguinis</td>
<td>Streptococcus oralis</td>
</tr>
</tbody>
</table>
the methods were able to correctly identify all annotated *E. coli* isolates, but identified at least some of them as *Shigella* spp. SpeciesFinder and TaxonomyFinder both had problems identifying the *Borrelia burgorferi* isolates, while SpeciesFinder and rMLST had problems distinguishing *Yersinia pestis* from *Yersinia pseudotuberculosis*. SpeciesFinder was the only method that had difficulties identifying *Mycobacterium tuberculosis* isolates, often predicting them to be *Mycobacterium bovis*.

**Performances on SRA draft genomes**

The SpeciesFinder, rMLST, TaxonomyFinder, and KmerFinder methods were next evaluated on the SRAdrafts set of 10,407 draft genomes covering 167 species. The performances on the draft genomes, for which the methods were able to make a prediction, are depicted in Figure 1B, while the overlap in predictions is illustrated in Figure 2B. Again, SpeciesFinder had the lowest performance with only 84% correct predictions. The rMLST, TaxonomyFinder, and KmerFinder methods had almost equal performances of 94%, 95%, and 95%, respectively. There was, however, a difference in the percentage of draft genomes for which each of the methods failed to make any prediction. SpeciesFinder and KmerFinder were the most robust methods, failing to make predictions for only 0.2% and 0.4% of the draft genomes, respectively. TaxonomyFinder was not able to make a prediction for 1.8% of the draft genomes, and rMLST not for 3.5%. That rMLST was the least robust method was at least partly due to our implementation of the method, where only hits with at least 95% identity and
Figure 2: Overlap in predictions by the five methods for species identification. Numbers written in regular font indicate the number of isolates for which the predicted species corresponds to the annotated species. Numbers written in italics indicate the number of isolates for which the predicted and annotated species differ. A: The 16S, rMLST, KmerFinder and TaxonomyFinder methods evaluated on the NCBI drafts set. B: The 16S, rMLST, and KmerFinder methods evaluated on the SRA drafts set. C: The 16S, KmerFinder, and Reads2Type methods evaluated on the SRA reads set.

95% coverage were considered a potential match. On the other hand, the N50 values for the draft genomes that SpeciesFinder and KmerFinder could not make a prediction for, were approximately half the size of the corresponding values for rMLST and TaxonomyFinder (data not shown), meaning that the quality of the draft genomes had to be higher for rMLST and TaxonomyFinder to be able to make a prediction. This is in accordance with these methods relying on the presence of many complete genes in the draft genomes.

Predictions for the most common species in the dataset are shown in Figure 4 and in Supplementary Figure 6-9. As seen previously when evaluating on the NCBI drafts set, the rMLST method was more consistent in its predictions for a given species than the other methods. For instance, rMLST predicted all 15 Mycobacterium bovis isolates to be M. tuberculosis. As also seen when evaluating on the NCBI drafts set, it is evident that all methods had difficulties distinguishing E. coli from species within the Shigella genus. Furthermore, species within the Brucella genus were often wrongly identified. In particular, it was only TaxonomyFinder that was able to correctly identify most Brucella abortus isolates. Some of the common problems that were obvious when evaluating on the NCBI drafts set, were not obvious when evaluating on the SRA drafts set, since the problematic species were too scarcely represented here. For instance, there are only five species from the Bacillus genus and only one S. mitis in SRA drafts. The difference in species distribution between the NCBI drafts and SRA drafts set...
Figure 3: Predictions for the most common species of the NCBI\textsubscript{drafts} set. For each method, the results for a given species is only shown if the method made a prediction for five or more isolates annotated as this species (e.g., if there are five isolates annotated as species A in the dataset, but the method was not able to make a prediction for one of the isolates, the species is not shown), or two or more isolates are predicted as this species (e.g., there are no isolates annotated as species B in the dataset, but two isolates annotated as species C are predicted to be species B, then species B is shown). A: Predictions by SpeciesFinder. B: Predictions by rMLST. C: Predictions by TaxonomyFinder. D: Predictions by KmerFinder.
also explain why SpeciesFinder, TaxonomyFinder and rMLST all have increased performance on the SRA\textsubscript{drafts} set: While more than half of the isolates in the SRA\textsubscript{drafts} set belong to the Salmonella, Staphylococcus or Streptococcus genera, which none of the methods have particular problems identifying, these genera constitute less than 20% of NCBI\textsubscript{drafts}. Conversely, the NCBI\textsubscript{drafts} set contains a high proportion of the problematic species \textit{E. coli} (8.8%) and the genus \textit{Bacillus} (10%). The corresponding proportions for SRA\textsubscript{drafts} are 3.5\% \textit{E. coli} and 0.05\% isolates of the \textit{Bacillus} genus. Furthermore, the NCBI\textsubscript{drafts} set is proportionally more diverse consisting of 149 species, while the almost 15 times larger SRA\textsubscript{drafts} set consists of only 168 different species.

Performances on short reads from SRA

Only three of the methods were able to perform species predictions directly on short reads, without first assembling the reads. These methods were SpeciesFinder, KmerFinder, and Reads2Type. Their performances on the SRA\textsubscript{reads} set of 10,407 sets of short reads representing 168 species are shown in Figure 1C.

Again, the SpeciesFinder method had the poorest performance with 86\% of the isolates being correctly predicted. Reads2Type performed a bit better (87\%), while KmerFinder achieved 97\% correct.

Figure 2C illustrates the overlap in predictions between the three methods, while predictions for the most common species are shown in Supplementary Figure 10. In general, the results correspond to those observed for the SRA\textsubscript{drafts} set.

Speed

The speed of the methods was evaluated on a subset of draft genomes and short reads as described in the Material and Methods. Since the actual speed experienced by the user will depend on a number of factors, for instance, the network bandwidth capacity of the client computer and the number of jobs queued at the server, the relative speed of the different methods in comparison to each other is more relevant than the absolute speed.

Table 2: Speed of the tested methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Speed on draft genomes</th>
<th>Speed on short reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpeciesFinder</td>
<td>00:13</td>
<td>3:14</td>
</tr>
<tr>
<td>Reads2Type</td>
<td>NA</td>
<td>1:20</td>
</tr>
<tr>
<td>rMLST</td>
<td>00:45</td>
<td>NA</td>
</tr>
<tr>
<td>TaxonomyFinder</td>
<td>11:33</td>
<td>NA</td>
</tr>
<tr>
<td>KmerFinder</td>
<td>00:09</td>
<td>03:10</td>
</tr>
</tbody>
</table>

DISCUSSION

In the present study we trained five different methods for prokaryotic species identification on a common dataset and evaluated their performances on three datasets of draft genomes
Figure 4: Predictions for the most common species in the SRA\textsubscript{drafts} dataset. For each method, the results for a given species is only shown if the method made a prediction for ten or more isolates annotated as this species, or two or more isolates are predicted as this species A: Predictions by SpeciesFinder. B: Predictions by rMLST. C: Predictions by TaxonomyFinder. D: Predictions by KmerFinder.
or short sequence reads.

The SpeciesFinder method is based on the 16S rRNA gene, which has served as the backbone of prokaryotic systematics since 1977 (1). Accordingly, sequencing of the 16S rRNA gene is a well-established method for identification of prokaryotes and has in all likelihood been used for annotating some of the isolates in the training and evaluation sets. In the light of this potential advantage of the SpeciesFinder method over the other methods, it is noteworthy that it had the lowest performance on all evaluation sets. Previous studies have, however, also pointed to the many limitations of the 16S rRNA gene for taxonomic purposes. Examples, which are also observed in this study, include its inadequacy for the delineation of species within the *Borrelia burgdorferi* sensu lato complex and the *Mycobacterium tuberculosis* complex (32). Similarly, *in silico* studies of the applicability of the 16S rRNA gene for the identification of medically important bacteria led to the authors concluding that although the method is useful for identification to the genus level, it is only able to identify 62% of anaerobic bacteria (33) and less than 30% of aerobic bacteria (34) confidently to the species level.

The performance of SpeciesFinder was surpassed only marginally by Reads2Type. This is not surprising, since the two methods are conceptionally very similar: SpeciesFinder utilizes the entire 16S rRNA gene of approximately 1,540 nucleotides, while for most species, Reads2Type looks for species-specific 50-mers in the same gene. In terms of its future usability, Reads2Type has, however, one advantage over the other methods: Like most of the other methods it is available as a web-server, but uniquely it does not require the read data to be uploaded to the server. Instead, a small 50-mer database is transferred to the user's computer and all computations performed here. As a result, bottleneck problems on the server are avoided and the data transfer is minimized, which may be particularly advantageous for users with limited Internet access.

While SpeciesFinder and Reads2Type only sample one locus, the rMLST method samples up to 53 loci – all ribosomal genes located to the chromosome of the bacteria. Evaluating on the dataset of SRA draft genomes, rMLST, TaxonomyFinder, and KmerFinder performed equally well. However, on the more diverse and difficult set of NCBI draft genomes, the rMLST method performed only marginally better than SpeciesFinder and significantly worse than TaxonomyFinder and KmerFinder. In particular, the rMLST method consistently made incorrect identifications of a number of closely related species, e.g., *Y. pestis* versus *Y. pseudotuberculosis* (35) and *M. tuberculosis* versus *M. bovis* (36). Also, rMLST consistently predicted the human pathogen *B. anthracis* to be *B. thuringiensis*. The later is used extensively as a biological pesticide and is generally not considered harmful for humans. *B. anthracis* and *B. thuringiensis* are both members of the *B. cereus* group and genetically very similar, with most of the disease and host specificity being attributable to their content of plasmids (37; 38). It has even been suggested that all members of the *B. cereus* group should be considered to be *B. cereus* and only subsequently be differentiated by their plasmids (39). Hence, in concordance with rMLST sampling only chromosomal, core genes, it is not surprising that the method fails to distinguish these isolates. A similar example is given by the rMLST method identifying all *E. coli* isolates as *Shigella sonnei*. Although *Shigella* spp. isolates have been rewarded their own genus, its separation from *Escherichia* spp. is mainly historical (40; 41; 42). To be sure, some of the mistakes commonly made by rMLST as well as the other methods highlight taxonomic taxa that are intrinsically difficult to distinguish due to a sub-optimal initial classification: Although *Shigella* spp. has for several years been considered a sub-strain of *E. coli*, the practical implications of renaming it is considered insurmountable.

The TaxonomyFinder method was the second most accurate method on the set of NCBI draft genomes and performed in the top for the SRA drafts set. In contrary to the other methods it does not work directly on the nucleotide sequence of the isolates, but rather on
the proteome, utilizing functional protein domain profiles for the species prediction. It was the slowest of the tested methods, but in return for the extra time, the user is rewarded with an annotated genome.

The KmerFinder method performs its predictions on the basis of co-occurring k-mers, regardless of their location in the chromosome. It had the overall highest accuracy, works on complete or draft genomes as well as short reads, was found to be very robust as well as fast. Furthermore, the KmerFinder method holds promise for future improvements, as the implementation used for this study was very simple: Only the raw number of co-occurring k-mers between the query and reference genome was considered, although a parallel analysis indicates that the performance could be improved even further if more sophisticated measures were used, also taking into account the total number of k-mers in the query and reference genome.

It has previously been noted that some of the isolates present in public databases, and hence used in this study, are wrongly annotated (16; 43; 44). Based on the current study, it is likely that at least the six isolates from the NCBI\textsubscript{drafts} set that all methods identified as something different than the annotated species, are wrongly annotated. In agreement with this, one of the isolates has indeed been re-annotated, since we initially downloaded the data. Of the remaining five isolates, two \textit{B. cereus} isolates were found to be most closely related to the \textit{B. weihenstephanensis} strain KBAB4 of the common training set. This strain is the single representative of the species in the public database and not the type strain. Hence there is no guarantee that the sequenced strain represents the named taxon (45). The same is the case for the \textit{C. botulinum} strain C Eklund, which is predicted to be a \textit{Clostridium novyi} based on its close resemblance to \textit{C. novyi} strain NT of the training set. \textit{Clostridium novyi} strain NT is the only representative of this species in the database and not the type strain.

While some taxonomists consider the goal of bacterial taxonomy to mirror the order of nature and describe the evolutionary order back to the origin of life (5; 46), a more pragmatic and applied view is likely to be advantageous for epidemiological purposes, where most outbreaks last less than six months. The number of prokaryotic genomes in public databases is currently sufficiently high to substitute theoretical views of which loci to sample for optimal species identification by actual testing how different approaches perform. One locus (the 16S rRNA gene) was initially used for sequenced-based examination of relationships between bacteria, and when the approach was found to have limitations, more loci were added in MLST and MLSA (47; 48). The addition of still more loci has been suggested for improving MLSA even further (32; 15). This study suggests that an optimal approach should not be limited to a finite number of genes, but rather look at the entire genome.

CONCLUSION

The 16S rRNA gene has served prokaryotic taxonomy well for more than 30 years, but the emergence of second- and third generation sequencing technologies enables the use of WGS data with the potential of higher resolution and more phylogenetically accurate classifications. Methods that sample the entire genome, not just core genes located to the chromosome, seems particularly well suited for taking up the baton.
ACKNOWLEDGEMENTS

This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

We are grateful to John Damm Sørensen for excellent technical assistance. We are grateful to Keith Jolley, Department of Zoology, University of Oxford, UK for providing us with the rMLST genes for the genomes in the training data.
References


Chapter 6

Conclusions and Future prospects

With the development of sequencing technologies, thousands of microbial sequences have become accessible in the past 20 years. Availability of multiple strains from the same genera and species provide possibility to explore microbial environmental adaptation and to determine the size and content of pan-genome. Sequence similarity search is the important step in the pan-genome analysis and comparative genomics in general. In this PhD thesis, applications of two homology search algorithms are demonstrated. BLAST-based approach is widely used pairwise comparison algorithm, which provides a good overview of the differences and similarities between closely related organisms. However, comparison results of the diverse set of genomes are less accurate. A novel, profile HMM-based approach for sequence similarity search was introduced. Similar to BLAST-based methods, this method finds applications in pan-genome analysis and microbial identifications. However HMM-based approach is more sensitive and performs better in comparison between diverse organisms.

The PanFunPro method was applied to determine the number of shared
proteins within a set of 2110 complete genomes; to investigate differences and similarities between two chromosomes of *Vibrio* species, as well as genomic content comparison of newly sequenced MAP genome to the publicly available strains of the same genus. Furthermore, PanFunPro approach was employed to predict specific functional profiles for more than 1000 species; which were used to establish the novel method for microbial identification. Comparison of the TaxonomyFinder approach to the standard microbial identification methods demonstrated good approach performance and showed that proteins, representing accessory genome can also be used as targets for taxonomy prediction. Additionally, TaxonomyFinder provides *in silico* functional annotation for the unknown isolates in a short amount of time, which can be helpful in epidemiological characterization and outbreak investigation.

In the future, these species-specific sequences can be used in microarray or primer design. Moreover, the idea of both PanFunPro and TaxonomyFinder can be extended to the metagenomics area. Combinations of species-specific functional profiles can be used in metagenomic sample characterization. Pan-genome analysis of the large set of genomes can be performed to investigate the stable pan-genome for different taxonomic groups.
Bibliography


BIBLIOGRAPHY


158


Appendix A

Supplementary Material

Supplementary Material is available online via http://www.cbs.dtu.dk/~oksana/PhD_Thesis/Supplementary_Material/. It contains supplementary figures for each manuscript, included in this thesis. Additionally, figures, demonstrated in the main part of the article are accessible in high resolution.