From genotype to phenotype: inferring relationships between microbial traits and genomic components

Inaugural-Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Aaron Weimann
aus Oberhausen

Düsseldorf, 29.08.16

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathemathisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Alice C. McHardy
Koreferent:	Prof. Dr. Martin J. Lercher
Tag der mündlichen Prüfung:	24.02.17

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation eigenständig und ohne fremde Hilfe angefertig habe. Arbeiten Dritter wurden entsprechend zitiert. Diese Dissertation wurde bisher in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den ....

(Aaron Weimann)

Statement of authorship

I hereby certify that this dissertation is the result of my own work. No other person's work has been used without due acknowledgement. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Summary

Bacteria live in almost any imaginable environment, from the most extreme environments (e.g. in hydrothermal vents) to the bovine and human gastrointestinal tract. By adapting to such diverse environments, they have developed a large arsenal of enzymes involved in a wide variety of biochemical reactions. While some such enzymes support our digestion or can be used for the optimization of biotechnological processes, others may be harmful – e.g. mediating the roles of bacteria in human diseases. Thus, understanding the functional potential of bacteria holds promises for biotechnology and for addressing important questions in the treatment of bacterial infections. This is especially true, as more and more pathogens develop resistances to available antimicrobial drugs, causing deaths and large economic costs.

Due to advances in high-throughput DNA sequencing, the number of sequenced bacterial genomes is growing rapidly. For many of these, their functional or pathogenic potential is not known. Making use of these data sets requires sophisticated interpretation techniques that leverage existing genome annotations, including methods to systematically deduce bacterial genotype-phenotype associations, which can then be used to determine the phenotypic potential of newly sequenced genomes. In this thesis I describe three such approaches that use machine learning techniques for relating bacterial phenotypes to their genotypes, with applications in infection research and biotechnology.

First, we were interested in learning the genetic determinants of bacterial plant biomass degradation. Enzymes encoded in the genomes of plant biomass degrading bacteria can be used in the biotechnological conversion of plant material into biofuels, which could eventually replace climate-damaging fossil fuels. We used an L1-regularized L2-penalized support vector machine to learn an accurate phenotype classifier based on the profiles of protein families from the genomes of a large and manually curated set of plant biomass degraders. Based on feature selection from the obtained phenotype model, we identified protein families of enzymes known by physiological and biochemical tests to be implicated in the degradation of various components of plant biomass, as well as uncharacterized protein families, which represent targets for biochemical characterization.

Second, we developed Traitar, a multi-trait prediction software. The traits available for classification in Traitar cover many aspects of the bacterial metabolism, such as use of various substrates as carbon and energy sources, oxygen requirement, morphology, antibiotic susceptibility, proteolysis and other enzymatic activities. Traitar provides two prediction modes: one based on the profiles of protein families and one also incorporating the evolutionary history of protein families. The phenotype classifiers provided by Traitar were trained and rigorously evaluated on phenotypes and genomes from 572 species of 8 phyla. Traitar only requires features that can be computed directly from the genome sequences and is applicable to the rapidly increasing number of genomes recovered from single cells, isolates, and metagenomes

Last, we learned the determinants of multi-drug resistance of *Pseudmonas aeruginosa* by employing a logistic regression classifier. Here, we used expression and mutational profiles from a large set of clinical isolates. We could accurately predict the resistance to five different antibiotics and identified known resistance markers, as well as uncharacterized proteins that might provide further insights into the resistance acquisition mechanisms.

In summary, the methods presented in this thesis allow to study the phenotypic potential of bacteria based on their genomes. Uncovering the genotype-phenotype associations for biotechnologically important traits may aid in the discovery of novel enzymes that can be employed in industrial processes. Similarly, biomarkers of antibiotic resistance could improve the selection of antibiotics in patient therapy.

Zusammenfassung

Bakterien besiedeln alle vorstellbaren Lebensräume. Solche, in denen sehr extreme Umweltbedingungen vorherrschen (z.B. in heißen Thermalquellen) genauso wie im Verdauungstrakt von Rindern und Menschen. Indem sich Bakterien an solche Lebensräumen angepasst haben, haben sie im Laufe der Evolution ein Arsenal von Enzymen hervorgebracht, das in verschiedensten Stoffwechselwegen zum Einsatz kommt. Manche solcher Enzyme fördern sogar unsere Verdauung oder können zur Optimierung von biotechnologischen Prozessen eingesetzt werden. Andere Enzyme dagegen haben eine schädliche Wirkung, z.B. Virulenzfaktoren von Krankheitserregern. Das Verständnis des funktioniellen Potentials solcher Bakterien ermöglicht es, Fragestellungen in der Biotechnologie und in der Infektionsforschung zu adressieren. Das ist besonders wichtig, da mehr und mehr pathogene Organismen Antibiotikaresistenzen entwickeln, die zu Todesfällen und hohen ökonomische Kosten führen.

Die Fortschritte bei Hochdurchsatz-Sequenziertechnologien haben zu einem rapiden Anstieg der Anzahl der sequenzierten bakteriellen Genome geführt. Für viele diese Genome, ist das funktionelle oder pathogene Potential noch nicht bekannt. Um einen Nutzen aus diesen großen Datenmengen zu ziehen, bedarf es ausgeklügelter Techniken zur Datenauswertung, die die bereits existierenden Genomannotationen gezielt ausnutzen können. Dazu gehören auch Ansätze, die systematisch Genotyp und Phänotyp von Bakterien in Beziehung setzen, die dann auch eingesetzt werden können, um das phänotypische Potential von neu sequenzierten Bakterien zu bestimmmen.

In einem ersten Ansatz haben wir die genetischen Ursprünge des bakteriellen Pflanzen-Biomasseabbaus untersucht. Enzyme, die in den Genomen solcher bakterieller Biomasse-Abbauer codiert sind, können in der biotechnologischen Umsetzung von Pflanzenmaterial in Biotreibstoff eingesetzt werden, der langfristig konventionelle klimaschädliche fossile Treibstoffe ersetzen könnte. Hierzu haben wir ein L1-regularisiertes L2-penalisiertes Stützvektor-Verfahren eingesetzt, um einen akuraten Phänotyp-Klassifikator basierend auf den Proteinfamilien-Profilen der Genome eines großen und kuratierten Datensatzes von Pflanzen-Biomasseabbauern zu entwickeln. Auf Basis dieses Klassifikators konnten wir mittels Methoden zur

Merkmalsidentifizierung Enzym-Proteinfamilien finden. Teils solche, die schon durch physiologische und biochemische Tests mit dem Abbau von pflanzlicher Biomasse in Verbindung gebracht wurden, andererseits aber auch uncharakterisierte Proteinfamilien, die aussichtsreiche Kandidaten für die tiefergehende biochemische Charakterisierung darstellen.
Zweitens haben wir Traitar entwickelt, ein Programm, um gleichzeitig viele Phänotypen anhand von einem Genom vorherzusagen. Die Phänotypen, die mit Traitar klassifiziert werden können, decken viele Aspekte des bakteriellen Metabolismus ab, wie zum Beispiel die Nutzung von verschiedensten Substraten als Kohlenstoff- und Energiequelle, dem Sauerstoffbedarf, Morphologie, Antibiotikaresistenzen, Proteolyse und weitere Enzymaktivitäten. Traitar bietet zwei verschiedene Vorhersagemodi: Einer basierend auf Profilen von Proteinfamilien und ein weiterer, der auch die evolutionäre Geschichte der Proteinfamilien berücksichtigt. Traitar wurde zuerst trainiert und anschließend gründlich auf Phänotypen und Genomen von 572 Spezies aus 8 Phyla evaluiert. Ferner benötigt Traitar zur Vorhersage nur Merkmale, die direkt aus den Genomsequenzen berechnet werden können, und ist dabei für die rapide ansteigende Anzahl von Genomen einsetzbar, egal, ob diese aus einzelnen Zellen, Isolaten oder Metagenomen stammen.
Zuletzt haben wir die genetischen Faktoren von Antibiotikaresistenzen in *Pseudomonas aeruginosa* mit Hilfe von logistischer Regression untersucht. Hierzu haben wir Expressions- und Mutationsprofile einer großen Anzahl von Isolaten verwendet. Wir konnten die Resistenzen gegen fünf Antibiotika akurat vorhersagen und haben dabei bekannte Resistenzmarker aber auch uncharakterisierte Proteine identifiziert, die weitere Einblicke in die Resistenzmechanismen der unterschiedlichen Antibiotika gewähren könnten.
Zusammenfassend erlauben die hier präsentierten Methoden, das phänotypische Potential von Bakterien basierend auf ihrem Genom zu studieren. Indem die Genotyp-Phänotyp-Assoziationen für biotechnologisch wichtige Phänotypen aufgedeckt werden, könnten in Zukunft neue Enzyme mit Einsatzmöglichkeiten in industriellen Prozessen gefunden werden. Genauso haben die Biomarker, die für die Antibiotikaresistenzen gefunden wurden das Potential, die Therapie und Diagnostik von multiresistenten Keimen wie *P. aeruginosa* zu verbessern.

Danksagungen

Ich möchte mich bei allen bedanken, die mich während der Promotion begleitet haben: Mitdoktoranden, wissenschaftliche Hilfskräfte, meine Eltern, Familie und Freunde. Die Liste derer, die ich gerne erwähnen würde ist zu lang, aber ich möchte gerne einige Menschen stellvertretend nennen. Besonders möchte ich meiner Betreuerin Alice McHardy für ihre Unterstützung während meiner Promotion danken. Ihre Ideen und Erfahrung haben diese Arbeit erst ermöglicht. Kooperation ist der Schlüssel zu erfolgreicher Forschung und deswegen möchte ich mich auch bei unserem Kollaborationspartner Phil Pope für die gute Zusammenarbeit im Schwerpunkt Pflanzenbiomasseabbau und sein fortwährendes Interesse an unseren Methoden bedanken. Ebenso bei Susanne Häußler, die als Kooperationspartnerin im Projekt zu den Antibiotikaresistenzen die letzte Phase meiner Promotion am Helmholtz-Institut für Infektionsforschung sehr geprägt hat. Dann möchte ich auch Angela Rennwanz erwähnen, die mit ihrem Engagement mehr als die gute Seele des Lehrstuhls war. Man mag es als selbstverständlich ansehen, aber ich möchte mich auch ganz herzlich bei meinen Mitdoktoranden für wissenschaftliche Diskussionen und Feedback zu Postern, Vorträgen und Manuskripten bedanken und hierbei besonders David Lähnemann, Johannes Dröge und Sebastian Konietzny hervorheben. Der Zusammenhalt der Arbeitsgruppe war eine wichtige Stütze und hat über so manche Durststrecke hinweg geholfen. Die Bolognese-Abende, Hutbau-Nachmittage und Kaffeepausen am See werden in Erinnerung bleiben.

Contents

List of Figures xv

List of Tables xvii

1 Introduction 1

1.1 Bacteria, their traits, and their potential 1
1.2 Microbial genome sequencing 4
1.3 Next generation sequencing for unculturable bacteria 5
1.3.1 Metagenomics 5
1.3.2 Single-cell genome sequencing 7
1.4 Protein function prediction 8
1.4.1 Guilt by association 9
1.4.2 From proteins to bacterial metabolism 10
1.5 Supervised learning for genotype-phenotype inference 11
1.6 Outline 13

2 Personal bibliography 17

3 Plant biomass degradation prediction 19

3.1 Abstract . . . 21
3.1.1 Background . . . 21
3.1.2 Results . . . 21
3.1.3 Conclusions . . . 21
3.2 Background . . . 22
3.3 Results . . . 25
3.3.1 Distinctive Pfam domains of microbial plant biomass degraders . . . 25
3.3.2 Distinctive CAZy families of microbial plant biomass degraders . . . 29
3.3.3 Identification of plant biomass degraders from a cow rumen metagenome . . . 32
3.3.4 Timing experiments . . . 35
3.4 Discussion . . . 35
3.5 Conclusions . . . 38
3.5.1 Methods . . . 39
3.5.2 Phenotype annotation of lignocellulose-degrading and non-degrading microbes . . . 40
3.5.3 Classification with an ensemble of support vector machine classifiers . . . 41
3.5.4 Performance evaluation . . . 41
3.5.5 Feature selection . . . 43
3.6 Supplementary material . . . 43

4 From genomes to phenotypes **45**
4.1 Abstract . . . 47
4.2 Introduction . . . 47
4.3 Results . . . 50
4.3.1 The Traitar software . . . 50
4.3.2 Evaluation . . . 53
4.3.3 Performance per taxon at different ranks of the taxonomy 56
4.3.4 Phenotyping incomplete genomes . . . 58
4.3.5 Traitar as a resource for gene target discovery . . . 61

4.3.6 Phenotyping biogas reactor population genomes 64
4.4 Discussion 65
4.5 Methods 68
4.5.1 The Traitar software 68
4.5.2 Cross-validation 73
4.5.3 Evaluation metrics 74
4.5.4 Majority feature selection 75
4.5.5 Acknowledgements 75
4.6 Supplementary material 75

5 P. aeruginosa antibiotic resistance prediction 77
5.1 Introduction 78
5.2 Results 79
5.3 Discussion 82
5.4 Materials and Methods 83

6 Synopsis 87

A Journal versions of the published articles 119

List of Figures

1.1 Extreme environments inhabited by microbial communities 2
1.2 Overview of microbial community analysis via metagenome sequencing 6
1.3 Maximum-margin hyperplanes for classification 12

3.1 Frequencies of the selected Pfam families in the individual genomes and metagenomes. 28
3.2 Frequencies of selected glycoside hydrolase (GH) families and carbohydrate binding modules (CBMs) in the (meta-) genome sequences. 31

4.1 Traitar application scenario 49
4.2 Work flow of Traitar 53
4.3 Macro-accuracy for each phenotype for the Traitar phypat and phypat+PGL phenotype classifiers 55
4.4 Classification accuracy for each taxon at different ranks of the NCBI taxonomy 57
4.5 Single-cell phenotyping with Traitar 59

4.6 Phenotyping simulated draft genomes and single cell genomes . . 60
4.7 Phenotype gain and loss dynamics match protein family dynamics 62

5.1 Venn diagram of cross-resistance to five antibiotics 85
5.2 Performance estimates for classification of five antibiotics 86

List of Tables

3.1 Misclassified species in the SVM analyses 26
3.2 Accuracy of classifying microbes as lignocellulose-degraders or non-degraders 30
3.4 Prediction of the plant biomass degradation capabilities for 15 draft genomes 34

4.1 Overview of the 67 traits available in Traitar for phenotyping . . . 51
4.2 Performance overview of Traitar evaluated on different data sets . 54
4.3 The most relevant Pfam families for classification of three important phenotypes 63
4.4 Phenotype predictions for two novel Clostridiales species 66

5.1 Antibiotic resistance frequency across 467 *P. aeruginosa* isolates . 80

CHAPTER 1

Introduction

1.1 Bacteria, their traits, and their potential

"I took a little white matter, which is as thick as if it were batter. I then most always saw, with great wonder, that in the said matter there were many very little living animalcules, very prettily a-moving."[1]. This is how Antoni van Leeuwenhoek – using a technically advanced microscope – probably as the first human being observed bacteria. He made this discovery more than 350 years ago, and 200 years ahead of time before the pioneering work of Louis Pasteur and Robert Koch in microbiology. Today we know that we live in a microbial world. Bacteria are ubiquitous and important for biotechnology and human health. They have evolved traits, such as cold resistance or the ability to utilize inorganic material as energy source, to survive in the most extreme environments, e.g. in the drainage of acid coal mines, hydrothermal vents in the deep sea, or the permafrost around the arctic. (Figure 1.1). In general microbes are described, characterized and distinguished by their traits as in Bergey's Manual of Systematic Bacteri-

[1]Letter written to the Royal Society on September 17 1683 by Antoni Leeuwenhoek

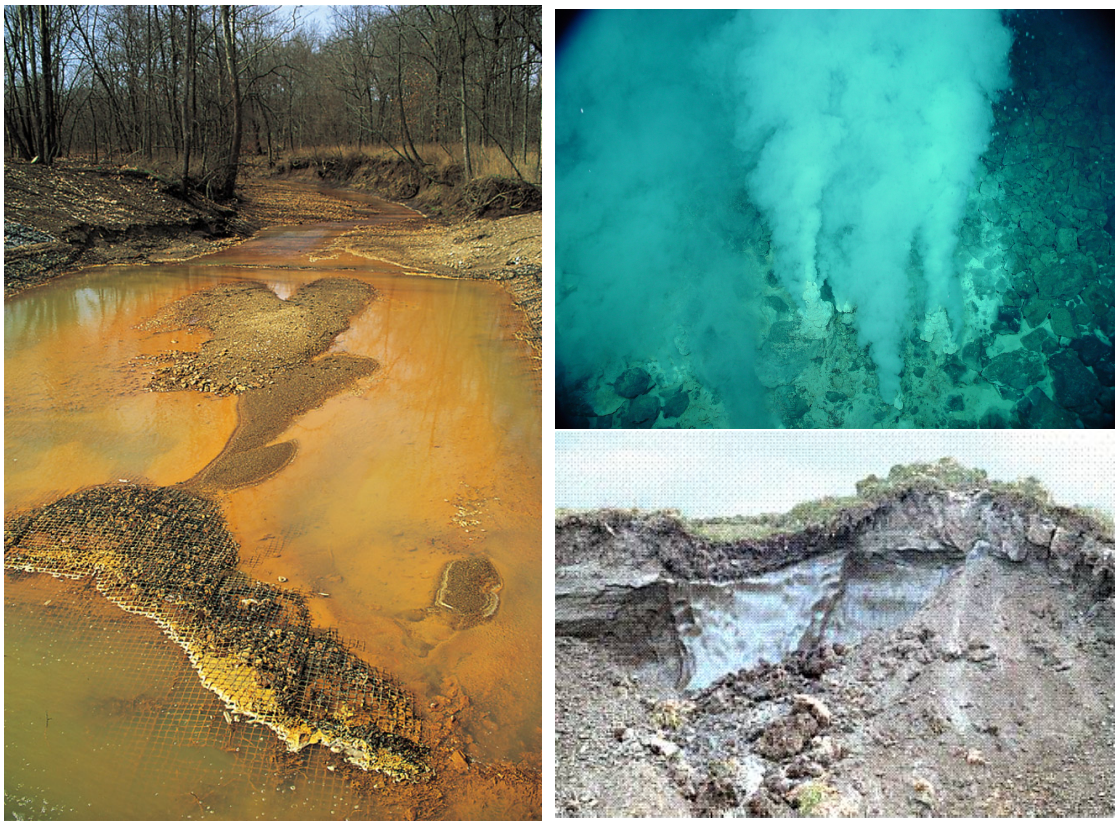


Figure 1.1: Extreme environments inhabited by microbial communities. Acid mine drainage (left; public domain) (Hardesty n.d.), hydrothermal vent (top right; public domain) (U.S. National Oceanic and Atmospheric Administration 2004), permafrost soil (bottom right; public domain) (NASA n.d.).

ology (Goodfellow et al. 2012). A trait or phenotype (used interchangeable in this thesis) can vary in complexity. For example, it can refer to the degradation of a specific substrate or the activity of an enzyme inferred in a lab assay, the respiratory mode of an organism, the reaction to Gram staining or antibiotic resistances. Traits, such as antibiotic resistance, are also driving factors for microbial community composition (J. L. Martinez 2008). Antibiotic treatment shapes a microbial community by killing off antibiotic susceptible bacteria, whereas resistant bacteria survive the treatment (Sommer et al. 2011).

Bacteria in human health Bacteria have also adapted to human environments, e.g. in the human gut they account for two kilograms of biomass (Gevers

et al. 2012). By fermentation of dietary fibers into short chain fatty acids, they support intestinal digestion (Andoh et al. 2003). The microbiota in obese mice has an elevated functional potential for collecting energy from the nutrition compared to lean mice (Turnbaugh et al. 2006). The human microbiome – the entirety of the microbial genomes of human-associated microbes – has been called our second genome (Grice et al. 2012). Large scale initiatives like the Human Microbiome Project were initiated to study the genomic composition of the complete healthy human microbiota, whereas the MetaHit project focused on the intestinal microbiota (Gevers et al. 2012; Qin et al. 2010).

Changes in the human gut microbiota have also been associated with different disorders, such as diabetes (Everard et al. 2013) and irritable bowel syndrome (IBD) (Kostic et al. 2014). For instance, the decrease of certain bacteria and an overall reduced bacterial diversity has been linked to IBD (Manichanh et al. 2006). Some human-associated bacteria may cause serious health problems for people with a compromised immune system. Anti bacterial drugs – so called antibiotics – are available for treatment of bacterial infections, but the occurrence of multi-drug resistant strains has become a global problem and causes many deaths yearly (Wright 2012).

Bacteria in Biotechnology Microbial community members with varying traits can help in waste water treatment, bioremediation of soils and promotion of plant growth (Narihiro et al. 2007; Olapade et al. 2015; Bai et al. 2015); plant biomass degrading bacteria influence the ability to process the recalcitrant plant material in the cow rumen microbiota (Hess et al. 2011); the dominant bacterial strain in the Tammar wallaby foregut microbiome is associated with lower methane emissions produced by the wallabies compared to ruminants (P. B. Pope, Smith, et al. 2011). Bacterial enzymes already drive the production of over 500 industrial products and there is a high demand for novel enzymes since biotechnological processes are often more sustainable and economic than their chemical counterparts (Adrio et al. 2014). Bacteria in microbial communities with traits related to such biotechnological processes have a large potential to serve as a reservoir of such enzymes.

From genomes to traits Inferring genotype-phenotype relationships for bacterial traits may contribute to understanding the functional roles of bacterial community members and provide insights into medically and biotechnologically relevant processes, such as the mechanisms of antibiotic resistances or microbial cellulose degradation. Thus, methods are required to suggests bacterial strains with relevant phenotypes from microbial communities and to identify protein families encoded in their genomes that can be linked to relevant traits or metabolic capabilities.

1.2 Microbial genome sequencing

DNA sequencing has revolutionized biology since the invention of the chain termination method by Sanger in 1975 (Sanger et al. 1975). Since the mid-2000s, high-throughput next generation sequencing (NGS) methods have led to an exponential increase of the sequenced DNA – although at the cost of shorter read lengths. In contrast, recently developed long-read sequencing platforms deliver reads with sizes of several kilobases and can also span repetitive genomic regions, but have high error rates and lower throughput (Goodwin et al. 2016; Laehnemann et al. 2016).

Genotyping methods based on microbial genome sequencing have the potential to replace phenotypic tests with faster and cheaper genotyping methods. For example, in a recent study, Quick *et al.* applied a phylogenetic placement method using MinION long-read real-time sequencing for profiling a *Salmonella* strain from a hospitalized patient. Within 2 h, they found that this strain belonged to the main outbreak in the hospital (Quick et al. 2015). Each year, thousands of microbial genome sequences are deposited in public databases (Land et al. 2015). This large number of available genomes paves the way to systematically associate microbial genotypes with phenotypes.

1.3 Next generation sequencing for unculturable bacteria

In 1985, Staley and Konopka made the observation – later on termed "the Great Plate Count Anomaly" – that the size of microbial populations estimated by dilution plating and microscopy differed significantly from another (Staley et al. 1985). Today, we know that this is because many bacteria do not grow under available culturing conditions. Hugenholtz *et al.* estimated that as little as 1% of bacteria can be cultured. Bacteria are usually isolated by selectively growing them in a culture medium – a prerequisite for most diagnostic tests, for example to determine the cause of an infection (Hugenholtz et al. 1998).

1.3.1 Metagenomics

NGS methods gave rise to a new research direction in microbial community analysis called metagenomics. I will explain this approach in more detail in this section as the first two methods described in this thesis mainly target metagenomic data. Metagenomics avoids the culturing step by directly sequencing DNA from an environmental sample, with the ultimate goal to reconstruct full genomes from the metagenomic sequencing data. A metagenomic experiment requires a series of experimental, as well as bioinformatics analysis steps (Figure 1.2). In contrast to standard genome sequencing, the sequenced DNA fragments are not derived from a clonal culture, but from hundreds or thousands of differentially abundant bacterial strains – creating a complicated puzzle to be solved.

Assembly As a first step, metagenome reads are often merged to reconstruct the original DNA sequences, which is also known as assembly. Metagenomic *de novo* assembly methods usually do not recover complete genomes but rather assemble reads into longer contiguous sequences (contigs), that are used in downstream analyses. Strain variation and non-uniform coverage make it a highly complex problem, which requires substantial computing power (A. Howe et al. 2015). Howe *et al.* proposed a pre-filtering approach for metagenomes from complex communities such as soil, which after assembly gave similar results compared

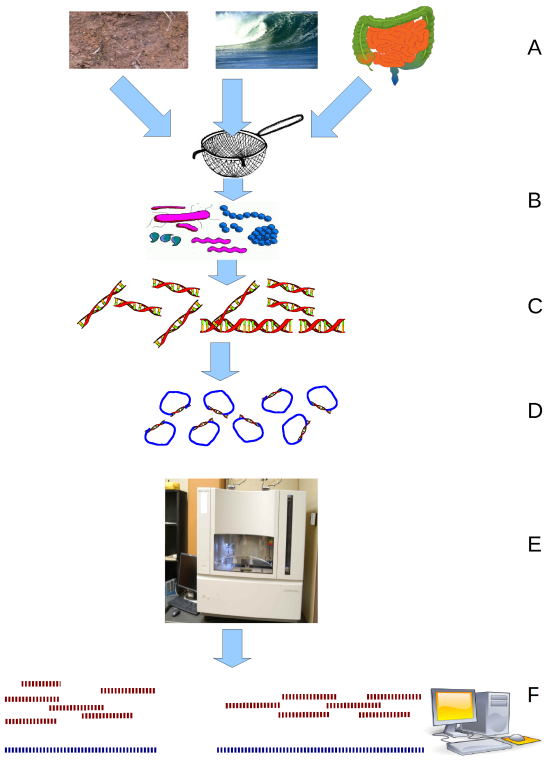


Figure 1.2: Overview of microbial community analysis via metagenome sequencing: samples from an environment are taken (A); bacterial cells are isolated (B); cells are lysed and DNA is extracted (C); DNA is amplified and sequence library is prepared (D); environmental DNA is sequenced (E) Sequence assembly, genome reconstruction and other bioinformatics downstream analyses (F) (from Wooley and Godzik (Wooley et al. 2010)); Creative Commons Attribution license .

to assembling the unfiltered metagenome, requiring less computing power (A. C. Howe et al. 2014). State-of-the-art assembling methods for short-reads from NGS commonly employ de Bruijn graphs to solve the assembly problem. A de Bruijn graph represents overlaps between oligomers of a given length – also called k-mers, which are derived from the sequence reads (A. Howe et al. 2015). IDBA-UD is a metagenomic assembler that attempts to partition the de Bruijn graph into isolated components for each species and tries to resolve strain variation within these components (Peng et al. 2011). Ray Meta was developed to scale to very large data sets by distributing the computations across many processors with moderate memory usage (Boisvert, Raymond, et al. 2012). Recently, the assembly of synthetic long read data sets of complex soil microbial communities was

shown to significantly improve the metagenomic assembly quality (Bankevich and Pevzner 2016). Synthetic long reads, which are derived from barcoded short reads, represent an alternative to short-read NGS, as well as to the error-prone and low-throughput true long-read sequencing platforms. Since in general, the quality of genome or metagenome assemblies can differ considerable across different method, QUAST and metaQUAST were developed to compare different assemblies of genomes or metagenomes with each other (Gurevich et al. 2013; Mikheenko et al. 2015).

Genome recovery The assembled metagenomic contigs are subjected to recovery of genomes from metagenomes (GFMs), also known as binning (Dröge et al. 2012). Differential read coverage and k-mer usage have proven to be the most successful features for genomic deconvolution and inferring the taxonomic origin of the contigs (Alneberg et al. 2014; Cleary et al. 2015; Gregor et al. 2016; Imelfort et al. 2014; Kang et al. 2015; Nielsen et al. 2014). Methods for genome recovery have been applied to large metagenomic data sets, delivering hundreds of draft genomes with different degrees of completeness and quality (Brown et al. 2015; Hess et al. 2011). Clade-specific marker genes can provide an estimate of the completeness of such genome recoveries (Parks et al. 2015).
For each bioinformatic analysis step of a metagenome, an increasing number of tools are available – many evaluated on different data sets and each with unique strengths and weaknesses. This leaves the user with the difficult task to select the most appropriate choice for the analysis. To address this shortcoming, the Critical Assessment of Metagenome Interpretation (CAMI) challenge was initiated to provide an objective and comprehensive comparison of metagenomic software [2].

1.3.2 Single-cell genome sequencing

The sequencing of genomes derived from single bacterial cells represents another culture-independent technique to recover genomes – also from less abundant taxa in microbial communities. After isolation of single cells, the ϕ29 polymerase – a

[2] http://www.cami-challenge.org/

bacterial phage DNA polymerase that is very sensitive and has low error rates – is employed to amplify the single-copy DNA using multiple displacement amplification (Gawad et al. 2016). Algorithms developed for the assembly of single-cell genomes take the non-uniform coverage of single-cell sequencing data into account, in contrast to assemblers for multi-cell genome assembly (Boisvert, Laviolette, et al. 2010; Bankevich, Nurk, et al. 2012). MeCorS leverages the uniform coverage of metagenomic data for error correction of single-cell sequencing assemblies in regions of low coverage, given a metagenome is available from the same environment as the sequenced single cells (Bremges et al. 2016). Large-scale single-cell sequencing studies have led to several hundreds of recovered single-cell draft genomes (Lasken et al. 2014; Rinke et al. 2013).

1.4 Protein function prediction

The protein-encoding genes derived in sequence data set derived from single-cells and microbial communities represent a large reservoir of functionally uncharacterized proteins – potentially, many with functions relevant for biotechnology and clinical applications. The functional annotation of such proteins is the fundamental step to understand and determine their functional roles and the metabolic capabilities of microbial community members. However, only a small fraction of the proteins in public databases have a functional annotation and much less were experimentally characterized (Jaroszewski et al. 2009). As a first step, targets for functional annotation need to be identified. To this end, specialized software such as Prodigal and FragGeneScan can detect protein-coding sequences in prokaryotic genomic sequences. Whereas Prodigal works on assembled sequences (Hyatt et al. 2010), FragGeneScan has been optimized to find genes in short error-prone reads common in metagenomes (Rho et al. 2010). A popular strategy is to functionally annotate coding sequences encoding proteins of unknown function by searching for characterized proteins that have a similar structure or sequence (homologous proteins)(D. Lee et al. 2007), for instance using the famous Basic Local Alignment Tool (BLAST) (Altschul et al. 1990).

1.4.1 Guilt by association

Homology transfer of function is not without its caveats. Gene duplication is the process by which a region of DNA containing one or several genes gets duplicated. One copy of a gene is often rendered dysfunctional in the course of subsequent evolution or undergoes neo-functionalization, while retaining a similar sequence or structure; the other copy keeps the original functional role (Roth et al. 2007). Orthology refers to a group of proteins, which descended from a single gene of the last common ancestor. An orthologous group of proteins is a better indicator of functional conservation than than sequence homology alone (Sonnhammer et al. 2002). The annotation of proteins with the help of information from the genomic context is known as guilt by association (Aravind 2000). For instance, clusters of Orthologous Genes (COGs) can be constructed using all-against-all sequence comparisons with BLAST across bacterial genomes (Tatusov et al. 2001). Such a cluster must contain at least three proteins that are more similar to each other than to any other protein from the same genomes. eggNOG is the unsupervised extension of COG, which does not require manual curation (Huerta-Cepas et al. 2016).

COGs or other groups of functionally related protein sequences can be represented by a multiple sequence alignment (MSA) that carries information about insertions and deletions at specific positions of the alignment. Profile hidden Markov models (HMM) have been developed to model a MSA (Eddy 1998). A protein can be annotated against a database of pre-built profile HMMs with BLAST-like accuracy and speed. For instance, the protein family database (Pfam) contains more than 15000 profile HMMs derived from protein sequences from all domains of life (Finn, Coggill, et al. 2016), but there are also more specific databases like dbCAN, which contains only families of carbohydrate active enzymes (Yin et al. 2012).

Another example of assigning function based on guilt by association is to detect gene fusion events. Sets of genes that appear separate in one genome frequently appear joined together in a second genome. Evolution suggests that there is a selective pressure for these events to happen, and indeed fused genes are likely to be functionally associated (Enright et al. 1999).

Association by co-occurence Protein families that have a similar phylogenetic distribution across genomes from many organisms tend to be involved in a similar biological process (Kensche et al. 2008). Clustering methods that operate on co-occurence patterns can be used to detect functional modules – sets of functionally coupled proteins (Yamada et al. 2006; Konietzny, Dietz, et al. 2011). A functional module can represent a pathway, which is a set of biochemical reaction steps or a protein complex like the flagellum, the bacterial motor complex. Examples of databases of curated functional modules, such as pathways, include the Kyoto Encyclopedia of Genes and Genomes (KEGG) and MetaCyc (Caspi et al. 2016; Kanehisa et al. 2015). However, approaches that only take the co-occurence patterns of the proteins into account and ignore the phylogenetic relationships between the organisms may recover false positive functional links between these proteins. Other methods explicitly make use of a phylogenetic tree, which represents the evolutionary history of the analyzed genomes (Barker et al. 2005; Cohen, Ashkenazy, et al. 2013).

Association by phenotype Bacterial genome-wide association studies (GWAS) attempt to link genetic features (like single nucleotide polymorphisms or protein families) across bacterial genomes from strains with a common phenotype of interest (Y. Liu et al. 2006; P. E. Chen et al. 2015; Dutilh et al. 2013). Machine learning methods can be employed to learn a statistical model of a phenotype, by exploiting co-occurrence patterns of phenotype and protein families to find a new group of protein families that are predictive of a trait. (Khaledi et al. 2016; MacDonald et al. 2010; Lingner et al. 2010). Such genotype-phenotype models may also be re-used to identify bacterial strains that encode these phenotypes of interest from newly sequenced metagenomic, single-cell or isolate genomes. But, so far, only few genotype-phenotype models for biotechnologically or medically relevant traits have been established.

1.4.2 From proteins to bacterial metabolism

Metabolic reconstruction aims at reconstructing the bacterial metabolism based on the individual enzymes found in a given genome. These methods model the

bacterial cell as a network of pipes through which substrates and products flow. However, these models are time-consuming to construct and require a high quality functional annotation, which is only available for model organisms (Durot et al. 2009). Metabolic reconstruction has also been attempted for metagenomic data but not for single genome recovered from metagenomes. For instance, Abubucker *et al.* developed HUMAnN, to determine the functional potential of microbial community metagenomes, and used it to identify pathways enriched in the human microbiome at specific body sites (Abubucker et al. 2012). Liu and Pop developed MetaPath, which also aims at identifying differentially abundant pathways across metagenomes (B. Liu et al. 2011). Methods that can provide insights into the metabolism of non-model organisms from genome-level information, are needed to complement such metabolic reconstruction approaches.

1.5 Supervised learning for genotype-phenotype inference

A central idea of this thesis was to use supervised machine learning methods to infer links between a target phenotype and genetic features. A feature represents a single measurement, for example the expression level of a specific gene. Supervised learning, also known as classification, derives some rules between input features and a target class, whereas unsupervised methods look for some structure inherent to the input data i.e. clusters. Classification methods can be further divided into linear and non-linear models. Linear methods separate the input samples by a linear decision boundary, but, in contrast to non-linear methods, they cannot take interactions between features into account. However, they provide a direct way to interpret the contributions of the individual features to the classification. This is particularly important in biological settings, where high-throughput assays can probe hundreds of thousands of features at the same. In addition, the analysis of such high-dimensional data sets is challenging, because classifier can easily be overfitted. An overfitted classifier has a good fit with the sample data but a poor fit for unseen samples. Methods, which only select a few input features as relevant for the classification often have superior generalization performance (Hastie et al. 2009).

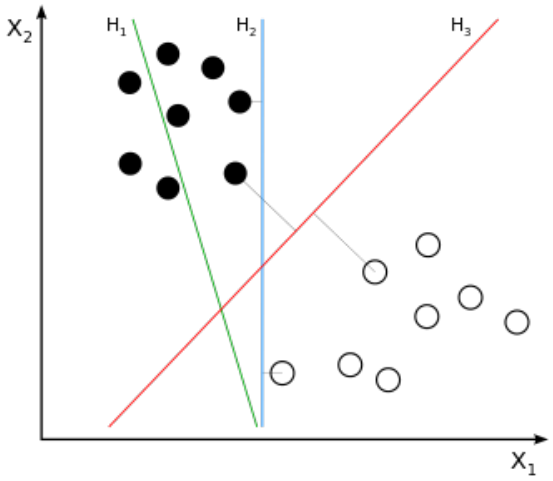


Figure 1.3: Maximum margin hyperplanes for classification: H_1 does not separate the two point clouds. H_2 does, but H_3 is the separating hyperplane with the largest margin (Creative Commons Attribution-Share Alike 3.0 Unported) (Weinberg 2012).

Support vector machines We heavily relied on support vector machines (SVM) in our studies. SVMs are geometric classifiers. They are trained by finding a hyperplane that separates the phenotype-positive and phenotype-negative class with the largest possible margin. The margin is defined by the distance between the closest point of the positive and the negative class to the hyperplane (Figure 1.3). To accomodate cases where the problem is not linearly separable, the soft-margin SVM allows samples of the positive class to lie on the other side of the margin, i.e. it seeks a trade-off between a large margin and placing a small number of samples on the wrong side of the margin (Boser et al. 1992). In particular, we used a L1-regularized L2-loss SVM, which is a linear SVM. L1 regularization allows to shrink the number of non-zero weights given to individual features, enabling sparse models (Fan et al. 2008).

Logistic regression We further made use of the logistic regression classification method in our genotype-phenotype association studies. Logistic regression is also a linear classifier, but, in contrast to SVMs, it is a probabilistic classifier and models the posterior probabilities of the phenotype-positive and phenotype-negative classes as linear functions. In practice logistic regression has similar gen-

eralization performance as the SVM and will always find the optimal hyperplane in the linear separable case. Samples classified with a logistic regression classifier will receive an actual probability to belong to the positive or negative class, which provides a degree of certainty about the classification decision (Hastie et al. 2009). In contrast, SVM output would need first to be transformed into a probability distribution over the two classes using, e.g. Platt's scaling (Platt 1999). Again, we used a L1-regularized version of the classifier to obtain sparse models (Fan et al. 2008).

1.6 Outline

This thesis is a cumulative work consisting of a series of scientific articles I co-authored during my PhD project. In Chapter 1, I place the individual contributions of this dissertation into a larger context. Chapter 2 provides my personal bibliography including the articles presented in this thesis as well as further publications I co-authored, with a short description of my individual contributions. Chapter 3 describes a method for finding the genetic determinants of plant biomass degradation, published as a peer-reviewed article at the journal Biotechnology for Biofuels. Chapter 4 presents a method for multi-trait prediction from microbial genomes, published as a pre-print, which is at the time of writing this thesis under review. Both articles are identical to the published versions, but were typeset to allow a uniform appearance. The orginal versions of the articles are also provided in the appendix. Chapter 5 describes an approach to predicting antibiotic resistance of *Pseudmonas aeruginosa* and is currently in preparation for publishing. I conclude with a short synopsis of this dissertation.

The genomic components of plant biomass degradation Microbial plant biomass degradation is a phenotype of considerable biotechnological interest. Bacteria efficiently break down complex plant carbohydrates into simpler sugars by enzymes encoded in their genomes. These enzymes can be employed in biotechnological conversion of plant material into biofuels, which have lower greenhouse emissions than climate-damaging fossil fuels (Rubin 2008). We used a L1-regularized L2-penalized SVM to learn the plant biomass degradation phe-

notype based on the profiles of protein families of a large and manually curated set of plant biomass degraders (Chapter 3). From the classifier, using feature selection, we could identify Pfam and CAZy families known to be implicated with plant biomass degradation, but also uncharacterized protein families that represent potential new carbohydrate-active enzymes (CAZyme). We also employed our genomic model of lignocellulose degradation for assessing the degradative capabilities encoded in the draft genomes of uncultured bacteria from a cow rumen metagenome (Hess et al. 2011). Our model identified Bacteroidetes-affiliated phylotypes to be involved in plant biomass degradation, which was supported by biochemical analysis of enzymatically active CAZymes (Naas et al. 2014).
In a complementary approach, we identified functional modules associated with plant biomass degradation, using latent Dirichlet allocation (Konietzny, P. B. Pope, et al. 2014). Each of five modules detected in this way contained protein families related to the degradation of different components of plant material such as cellulose, hemicellulose and pectin, but also covered the building blocks of the cellulosome – a multi-enzyme complex produced by many plant biomass degrading microbes. Using these modules, we could identify a considerable number of previously characterized plant biomass degradation associated operons – genomic units, that typically contain proteins with related functions – in genomes of known lignocellulose degraders. Importantly, we also detected many uncharacterized gene clusters that contain both known CAZy families, as well as hypothetical proteins. Several of such clusters have become targets for on-going biochemical characterization studies in the lab of our collaborator Phillipp B. Pope at the Norwegian University of Life Sciences.

Traitar, the microbial trait analyzer Since our approach to determine the genetic determinants of an individual phenotypes proved to be very accurate for the plant biomass degradation phenotype, we decided to devise a fully automated method for multi-trait prediction that only requires as input a microbial genome sequence (Chapter 4). Previous studies have attempted to predict phenotypes from the phyletic patterns of protein families, but only for a limited number of phenotypes (Feldbauer et al. 2015; Kastenmuller et al. 2009; Konietzny, P. B. Pope, et al. 2014; Kastenmuller et al. 2009; Clark et al. 2007; Lingner et al. 2010;

Weimann, Trukhina, et al. 2013). PICA – originally developed by Beiko *et al.* and extended by Feldbauer *et al.* – currently is the only available software for microbial genotype-phenotype inference, but it requires as input microbial genomes with a pre-computed annotation of COGs and only supports the prediction of eleven traits. In comparison, 67 traits are currently available in Traitar (Feldbauer et al. 2015; MacDonald et al. 2010). As before, we used regularized SVMs as phenotype classifiers, but Traitar also includes a new prediction mode that provides classifiers trained by incorporating a statistical model of protein family evolution. Our software was trained and rigorously evaluated on phenotype data from a manually curated database, as well as on a large data set compiled from a microbiological encyclopedia (Goodfellow et al. 2012; Berger 2005). Traitar can reliably predict a total of 67 phenotypes, including 60 entirely novel ones. Additionally, the phenotype classifiers in Traitar represent a large resource of candidate links between protein families and phenotypes, and Traitar also suggests associations between phenotypes and protein families for newly sequenced genomes.

Antibiotic resistance prediction in *Pseudomonas aeruignosa* Multidrug resistance of human pathogens causes many deaths worldwide (Wright 2012). *Pseudomonas aeruginosa* is a particularly dangerous bacterium, due to high levels of virulence and natural antibiotic resistance (Rumbaugh 2014). In our study, we systematically associated genomic features of several hundred clinical isolates of *P. aeruginosa* with resistance to five commonly administered antibiotics (Chapter 5). Notably, and different from the approaches presented before, the clinical isolates stem from the same bacterial species and not from a diverse set of taxa. We trained logistic regression classifiers on the transcriptional and mutational profiles derived from RNAseq data. Both, mutational and transcriptional profiles allowed to train classifiers with high accuracies. Using these classifiers, we discovered known resistance markers and several uncharacterized proteins, which are currently experimentally characterized in the lab of our collaborator and might provide further insights into the resistance mechanisms. Our approach showed that genotype-phenotype association studies could soon reveal the genetic markers required to establish cheap and rapid antibiotic resistance profiling for treating *P. aeruginosa* infections.

CHAPTER 2

Personal bibliography

Publications of the thesis

- A. Weimann, K. Mooren, J. Frank, P. B. Pope, A. Bremges, and A. C. McHardy (2016b). "From genomes to phenotypes: Traitar, the microbial trait analyzer". *bioRxiv*, p. 043315. DOI: 10.1101/043315 (published as pre-print at the time of writing, but has since been published in the journal mSystems (Weimann et al. 2016a))

- A. Weimann, Y. Trukhina, P. B. Pope, S. G. Konietzny, and A. C. McHardy (2013). "De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes". *Biotechnology for Biofuels* 6.24. DOI: 10.1186/1754-6834-6-24.

Other publications

- S. Hacquard, B. Kracher, K. Hiruma, P. C. Münch, R. Garrido-Oter, M. R. Thon, A. Weimann, U. Damm, J.-F. Dallery, M. Hainaut, B. Henrissat, O. Lespinet, S. Sacristán, E. Ver Loren van Themaat, E. Kemen, A. C.

McHardy, P. Schulze-Lefert, and R. J. O'Connell (2016). "Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi". *Nature Communications* 7, p. 11362. DOI: 10.1038/ncomms11362.

I conducted an ancestral reconstruction of the gene gain and loss events based on the genomes of species in the *Colletotrichum* lineage and ran functional enrichment tests that revealed signficantly altered functional categories across the species. I prepared figures and text describing my analyses and read and commented the full length paper.

- D. Bulgarelli, R. Garrido-Oter, P. C. Münch, A. Weiman, J. Dröge, Y. Pan, A. C. McHardy, and P. Schulze-Lefert (2015). "Structure and function of the bacterial root microbiota in wild and domesticated barley". *Cell Host & Microbe* 17.3, pp. 392–403. DOI: 10.1016/j.chom.2015.01.011.

  I was responsible for the functional annotation of the barley metagenome data and provided assistance in the dN/dS ratio analysis. I prepared text describing my analyses and read and commented the full length paper.

- S. G. A. Konietzny, P. B. Pope, A. Weimann, and A. C. McHardy (2014). "Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders". *Biotechnology for Biofuels* 7.124. DOI: 10.1186/s13068-014-0124-8.

  I curated a collection of sequenced microbial plant biomass degraders and non-degraders. I implemented a functional annotation pipeline for the uniform annotation of the dataset. I prepared text describing my analyses and read and commented the full length paper.

CHAPTER 3

Predicting the genomic components and capabilities for plant biomass degradation

Status	**published**
Journal	Biotechnology for Biofuels (Impact factor 6.44)
Citation	A. Weimann, Y. Trukhina, P. B. Pope, S. Konietzny and A. C. McHardy (2011). De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-) genomes *Biotechnology for Biofuels 2013,* **6**: 24.
URL	http://www.biotechnologyforbiofuels.com/content/6/1/24.
Own contribution	38% Wrote the manuscript (with YT, ACM, PBP) Conceived and designed the experiments (with YT, ACM) Implemented and conducted the analyses Curated the phenotype labels (with SGAK) Inferred Pfam and CAZy annotation of the training genomes and cow rumem bins (with SGAK) Interpreted the classification results, the Pfam and CAZy families relevance for the phenotype, etc. (with ACM, YT, PBP)

3.1 Abstract

3.1.1 Background

Understanding the biological mechanisms used by microorganisms for plant biomass degradation is of considerable biotechnological interest. Despite of the growing number of sequenced (meta)genomes of plant biomass-degrading microbes, there is currently no technique for the systematic determination of the genomic components of this process from these data.

3.1.2 Results

We describe a computational method for the discovery of the protein domains and CAZy families involved in microbial plant biomass degradation. Our method furthermore accurately predicts the capability to degrade plant biomass for microbial species from their genome sequences. Application to a large, manually curated data set of microbial degraders and non-degraders identified gene families of enzymes known by physiological and biochemical tests to be implicated in cellulose degradation, such as GH5 and GH6. Additionally, genes of enzymes that degrade other plant polysaccharides, such as hemicellulose, pectins and oligosaccharides, were found, as well as gene families which have not previously been related to the process. For draft genomes reconstructed from a cow rumen metagenome our method predicted Bacteroidetes-affiliated species and a relative to a known plant biomass degrader to be plant biomass degraders. This was supported by the presence of genes encoding enzymatically active glycoside hydrolases in these genomes.

3.1.3 Conclusions

Our results show the potential of the method for generating novel insights into microbial plant biomass degradation from (meta-)genome data, where there is an increasing production of genome assemblages for uncultured microbes.

3.2 Background

Lignocellulosic biomass is the primary component of all plants and one of the most abundant organic compounds on earth. It is a renewable, geographically distributed and a source of sugars, which can subsequently be converted into biofuels with low greenhouse gas emissions, such as ethanol. Chemically, it primarily consists of cellulose, hemicellulose and lignin. Saccharification – the process of degrading lignocellulose into the individual component sugars – is of considerable biotechnological interest. Several mechanical and chemical procedures for saccharification have been established; however, all are relatively expensive, slow and inefficient (Rubin 2008). An alternative approach is realized in nature by various microorganisms, which use enzyme-driven lignocellulose degradation to generate sugars as sources of carbon and energy. The search for novel enzymes allowing an efficient breakdown of plant biomass has therefore attracted considerable interest (Kaylen et al. 2000; J. Lee 1997; Mitchell 1998; Wheals et al. 1999). In particular, the discovery of novel cellulases for saccharification is considered crucial in this context (Himmel et al. 2007). However, the complexity of the underlying biological mechanisms and the lack of robust enzymes that can be economically produced in larger quantities currently still prevent industrial application.

For some lignocellulose-degrading species, carbohydrate-active enzymes (CAZymes) and protein domains implicated in lignocellulose degradation are well known. Many of these have been recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as the enzymes of the glycoside hydrolase (GH) families GH6 and GH9 and the endoglucanase-containing family GH5. Two well-studied paradigms are currently known for microbial cellulose degradation: The 'free-enzyme system' is realized in most aerobic microbes and entails secretion of a set of cellulases to the outside of the cell. In anaerobic microorganisms large multi-enzyme complexes, known as cellulosomes, are assembled on the cell surface and catalyze degradation. In both cases, the complete hydrolysis of cellulose requires endoglucanases (GH5 and GH9), which are believed to target non-crystalline regions, and exo-acting cellobiohydrolases, which attack crystalline structures from either the reducing (GH7 and GH48) or non-reducing (GH6) end of the beta-glucan chain. However, in the

genomes of some plant biomass-degrading species, homologs of such enzymes have not been found. Recent genome analyses of the lignocellulose-degrading microorganisms, such as the aerobe *Cytophaga hutchinsonii* (Xie et al. 2007), the anaerobe *Fibrobacter succinogenes* (Brumm, Mead, et al. 2010; Morrison et al. 2009) and the extreme thermophile anaerobe *Dictyoglomus turgidum* (Brumm, Hermanson, et al. 2010) have revealed only GH5 and GH9 endoglucanases. Genes encoding exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures (dockerins and cohesins) are absent.

Metagenomics offers the possibility of studying the genetic material of difficult-to-culture (i.e. uncultured) species within microbial communities with the capability to degrade plant biomass. Recent metagenome studies of the gut microbiomes of the wood-degrading higher termites (*Nasutitermes*), the Australian Tammar wallaby (*Macropus eugenii*) (P. Pope et al. 2010; Warnecke et al. 2007) and two studies of the cow rumen metagenome (Brulc et al. 2009; Hess et al. 2011) have revealed new insights into the mechanisms of cellulose degradation in uncultured organisms and microbial communities. Microbial communities of different herbivores have been shown to be dominated by lineages affiliated to the Bacteroidetes and Firmicutes, of which different Bacteroidetes lineages exhibited endoglucanse activity (P. Pope et al. 2010; P. B. Pope, Mackenzie, et al. 2012). Notably, exo-acting families and cellulosomal structures have a low representation or are entirely absent from gut metagenomes sequenced to date. Thus, current knowledge about genes and pathways involved in plant biomass degradation in different species, particularly uncultured microbial ones, is still incomplete.

We describe a method for the *de novo* discovery of protein domains and CAZy families associated with microbial plant biomass degradation from genome and metagenome sequences. It uses protein domain and gene family annotations as input and identifies those domains or gene families, which in concert are most distinctive for the lignocellulose degraders. Among the gene and protein domains identified with our method were known key genes of plant biomass degradation. Additionally, it identified several novel protein domains and gene families as being relevant for the process. These might represent novel leads towards elucidating the mechanisms of plant biomass degradation for the currently less well understood microbial species. Our method furthermore can be used to identify plant

biomass-degrading species from the genomes of cultured or uncultured microbes. Application to draft genomes assembled from the metagenome of a switchgrass-adherent microbial community in cow rumen predicted genomes from several Bacteroidales lineages which encode active glycoside hydrolases and a relative to a known plant biomass degrader to represent lignocellulose degraders.

In technical terms, our method selects the most informative features from an ensemble of L1-regularized L2-loss linear Support Vector Machine (SVM) classifiers, trained to distinguish genomes of cellulose-degrading species from non-degrading species based on protein family content. Protein domain annotations are available in public databases and new protein sequences can be rapidly annotated with Hidden Markov Models (HMMs) or – somewhat slower - with BLAST searches of one protein versus the NCBI-nr database (Sayers et al. 2012). Co-occurrence of protein families in the biomass-degrading fraction of samples and an absence of these families within the non-degrading fraction allows the classifier to link these proteins to biomass degradation without requiring sequence homology to known proteins involved in lignocellulose degradation. Classification with SVMs has been previously used successfully for phenotype prediction from genetic variations in genomic data. In Beerenwinkel *et al.* (Beerenwinkel et al. 2003), support vector regression models were used for predicting phenotypic drug resistance from genotypes. SVM classification was used by Yosef *et al.* (Yosef et al. 2010) for predicting plasma lipid levels in baboons based on single nucleotide polymorphism data. In Someya *et al.* (Someya et al. 2010), SVMs were used to predict carbohydrate-binding proteins from amino acid sequences. The SVM (Boser et al. 1992; Cortes et al. 1995) is a discriminative learning method that infers, in a supervised fashion, the relationship between input features (such as the distribution of conserved gene clusters or single nucleotide polymorphisms across a set of sequence samples) and a target variable, such as a certain phenotype, from labeled training data. The inferred function is subsequently used to predict the value of this target variable for new data points. This type of method makes no *a priori* assumptions about the problem domain. SVMs can be applied to datasets with millions of input features and have good generalization abilities, in that models inferred from small amounts of training data show good predictive accuracy on novel data. The use of models that include an L1-regularization term favors solutions in which

few features are required for accurate prediction. There are several reasons why sparseness is desirable: the high dimensionality of many real datasets results in great challenges for processing. Many features in these datasets are usually non-informative or noisy, and a sparse classifier can lead to a faster prediction. In some applications, like ours, a small set of relevant features is desirable because it allows direct interpretation of the results.

3.3 Results

We trained an ensemble of SVM classifiers to distinguish between plant biomass-degrading and non-degrading microorganisms based on either Pfam domain or CAZY gene family annotations (see Methods section for the training and evaluation of the SVM classification ensemble). We used a manually curated data set of 104 microbial (meta-)genome sequence samples for this purpose, which included 19 genomes and 3 metagenomes of lignocellulose degraders and 82 genomes of non-degraders (Figure 3.1, Figure 3.2, Additional file 1: Table S1). Fungi are known to use several enzymes for plant biomass degradation for which the corresponding genes are not found in prokaryotic genomes and vice versa, while other genes are shared by prokaryotic and eukaryotic degraders. To investigate similarities and differences detectable with our method, we included the genome of lignocellulose degrading fungus *Postia placenta* into our analysis. After training, we identified the most distinctive protein domains and CAZy families of plant biomass degraders from the resulting models. We compared these protein domains and gene families with known plant biomass degradation genes. We furthermore applied our method to identify plant biomass degraders among 15 draft genomes from the metagenome of a microbial community adherent to switch grass in cow rumen.

3.3.1 Distinctive Pfam domains of microbial plant biomass degraders

For the training of a classifier which distinguishes between plant biomass-degrading and non-degrading microorganisms we used Pfam annotations of 101 microbial genomes and two metagenomes. This included metagenomes of mi-

Table 3.1: Shown are species which were misclassified with the $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers. Contrary to previous beliefs (Ivanova et al. 2011), recent literature indicates in agreement with our predictions that *T. curvata* is a non-degrader. Furthermore, recent evidence supports that *A. mirum* is a lignocellulose degrader, which has not been previously described (Anderson et al. 2012).

	$\mathbf{eSVM_{bPFAM}}$	$\mathbf{eSVM_{CAZY\_B}}$
False negatives	*Postia placenta* Mad-698-R *Xylanimonas cellulosilytica* DSM 15894 *Thermomonospora curvata* DSM 43183	*Thermonospora curvata* DSM 43183
False positives	*Actinosynnema mirum* 101 *Arthrobacter aurescens* TC1 *Thermotoga lettingae* TMO	*Actinosynnema mirum* 101

crobial communities from the gut of a wood-degrading higher termite and from the foregut of the Australian Tammar Wallaby as examples for plant biomass-degrading communities. Furthermore, 19 genomes of microbial lignocellulose degraders were included – of the phyla Firmicutes (7 isolate genome sequences), Actinobacteria (5), Proteobacteria (3), Bacteroidetes (1), Fibrobacteres (1), Dictyoglomi (1) and Basidiomycota (1). Eighty-two microbial genomes annotated to not possess the capability to degrade lignocellulose were used as examples of non-lignocellulose-degrading microbial species (Additional file 1: Table S1).

We assessed the value of information about the presence or absence of protein domains for distinguishing lignocellulose degraders from non-degraders. With the respective classifier, $eSVM_{bPFAM}$, each microbial (meta-)genome sequence was represented by a feature vector with the features indicating the presence or absence of Pfam domains (see methods). The nested cross-validation macro-accuracy of $eSVM_{bPFAM}$ in distinguishing plant biomass-degrading from non-degrading microorganisms was 0.91. This corresponds to 94% (97 of 103) of the (meta-)genome sequences being classified correctly. Only three of the 21 cellulose-degrading samples and three of the non-degraders were misclassified (Table 3.1). Among these were four Actinobacteria and one genome affiliated with the Basidiomycota and Theromotogae each.

We identified the Pfam domains with the greatest importance for assignment to the lignocellulose-degrading class by $eSVM_{bPFAM}$ (Figure 3.1; see Methods for the feature selection algorithm). Among these are several protein domains known

to be relevant for plant biomass degradation. One of them is the GH5 family, which is present in all of the plant biomass-degrading samples. Almost all activities determined within this family are relevant to plant biomass degradation. Because of its functional diversity, a subfamily classification of the GH5 family was recently proposed (Aspeborg et al. 2012). The carbohydrate-binding modules CBM_6 and CBM_4_9 were also selected. Both families are Type B carbohydrate-binding modules (CBMs), which exhibit a wide range of specificities, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose (Boraston et al. 2004). Type A CBMs (e.g. CBM2 and CBM3), which are more commonly associated with binding to insoluble, highly crystalline cellulose, were not identified as relevant by $\text{eSVM}_{\text{bPFAM}}$. Furthermore, numerous enzymes that degrade non-cellulosic plant structural polysaccharides were identified, including those that attack the backbone and side chains of hemicellulosic polysaccharides. Examples include the GH10 xylanases and GH26 mannanases. Additionally, enzymes that generally display specificity for oligosaccharides were selected, including GH39 β-xylosidases and GH3 enzymes.

We subsequently trained a classifier - $\text{eSVM}_{\text{fPFAM}}$ - with a weighted representation of Pfam domain frequencies for the same data set. The macro-accuracy of $\text{eSVM}_{\text{fPFAM.}}$was 0.84; lower than that of the $\text{eSVM}_{\text{bPFAM}}$; with nine misclassified samples (4 Actinobacteria, 2 Bacteroidetes, 1 Basidiomycota, 1 Thermotogae phyla and the Tammar Wallaby metagenome). Again, we determined the most relevant protein domains for identifying a plant biomass-degrading sequence sample from the models by feature selection. Among the most important protein families were, as before, GH5, GH10 and GH88 (PF07221: N-acylglucosamine 2-epimerase) (Figure 3.1).

GH6, GH67 and CE4 acetyl xylan esterases ("accessory enzymes" that contribute towards complete hydrolysis of xylan) were only relevant for prediction with the $\text{eSVM}_{\text{fPFAM}}$ classifier. Additionally, both models specified protein domains not commonly associated with plant biomass degradation as being relevant for assignment, such as the lipoproteins DUF4352 and PF00877 (NlpC/P60 family) and binding domains PF10509 (galactose-binding signature domain) and PF03793 (PASTA domain) (Figure 3.1).

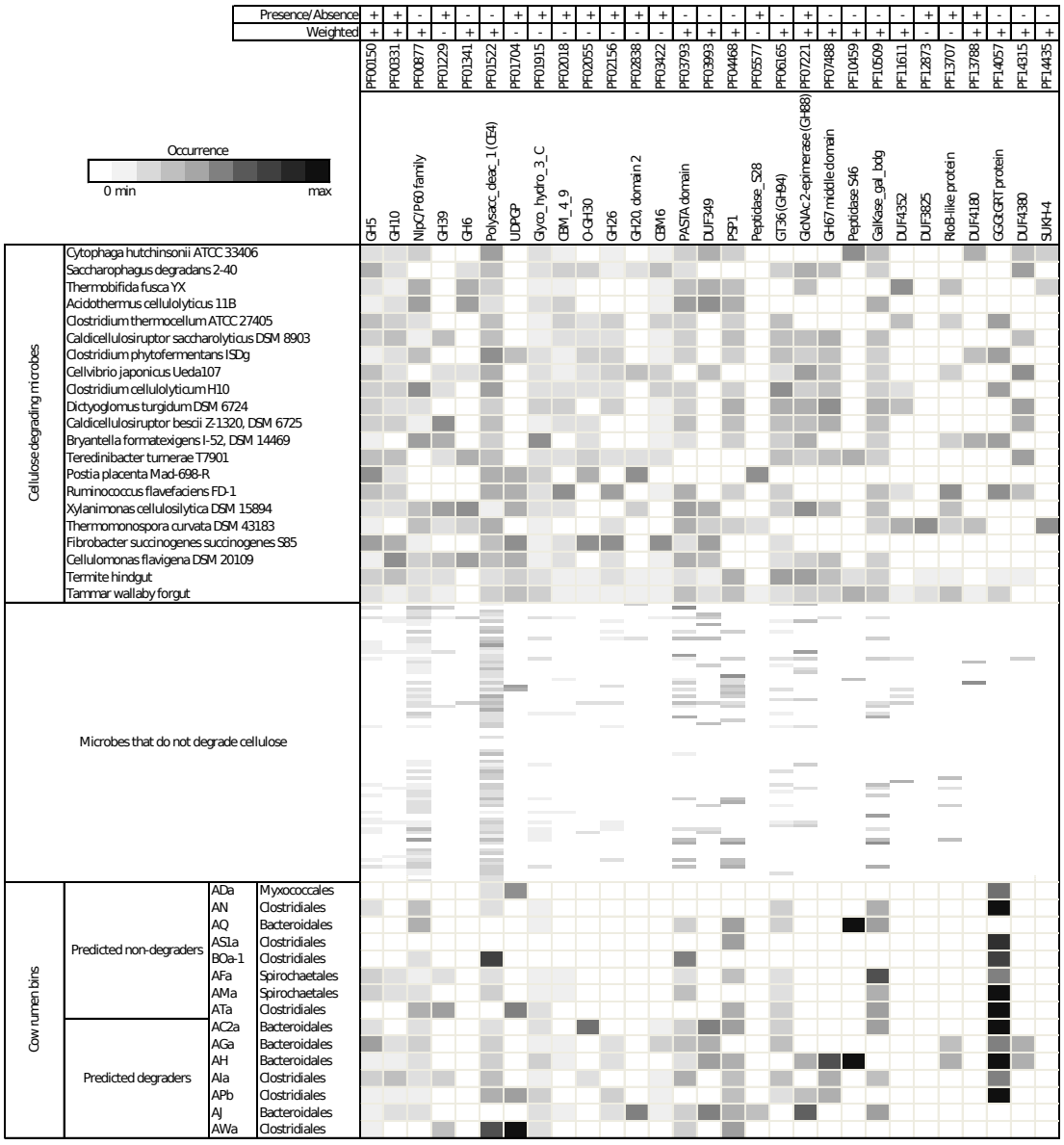


Figure 3.1: Frequencies of the selected Pfam families in the individual genomes and metagenomes. The data for each entry are rescaled by the total number of Pfam domains annotated to the microbial genome or metagenome. The color scale from grey to black indicates domain families that are present in low to high amounts, respectively. White indicates absent protein domains. The signs "+" and "-" indicate whether a protein domain was chosen in the respective experiment.

3.3.2 Distinctive CAZy families of microbial plant biomass degraders

We searched for distinctive CAZy families of microbial plant biomass degraders with our method. CAZy families include glycoside hydrolases (GH), carbohydrate-binding modules (CBM), glycosyltransferases (GT), polysaccharide lyases (PL) and carbohydrate esterases (CE). The annotations from the CAZy database comprised 64 genomes of non-lignocellulose-degrading species and 16 genomes of lignocellulose-degraders. There were no CAZy annotations available for the remaining genomes. In addition, we included the metagenomes of the gut microbiomes of the Tammar wallaby (TW), the wood-degrading higher termite and of the cow rumen microbiome (Additional file 1: Table S1). We evaluated the value of information about the presence or absence of CAZy domains, or of their relative frequencies for identification of lignocellulose-degrading microbial (meta-)genomes in the following experiments:

1. By training of the classifiers $\text{eSVM}_{\text{CAZY\_A}}$ (presence/absence) and $\text{eSVM}_{\text{CAZY\_a}}$ (counts), based on genome annotations with all CAZy families.

2. By training of the classifiers $\text{eSVM}_{\text{CAZY\_B}}$ (presence/absence) and $\text{eSVM}_{\text{CAZY\_b}}$ (counts), based on the annotations of the genomes and the TW sample with all CAZy families, except for the GT family members, which were not annotated for the TW sample.

3. By training of the classifiers $\text{eSVM}_{\text{CAZY\_C}}$ (presence/absence) and $\text{eSVM}_{\text{CAZY\_c}}$ (counts) with the entire data set based on GH family and CBM annotations, as these were the only ones available for the three metagenomes.

The macro-accuracy of these classifiers ranged from 0.87 to 0.96, similar to the Pfam-domain-based models (Table 3.2). Notably, almost exclusively Actinobacteria were misclassified by the $\text{eSVM}_{\text{CAZY}}$ classifiers, except for the Firmicute *Caldicellulosiruptor saccharolyticus*. The best classification results were obtained with the presence-absence information for all CAZy families except for the GT

Table 3.2: L1-regularized SVMs were trained with Pfam domain or CAZY family (meta-)genome annotations. Capital letters denote classifiers trained based on the presence or absence of CAZy families and small letters indicate classifiers trained based on the relative abundances of CAZy families in annotations. Abbreviations "A", "a"," B", "b", "C", "c" denote the following: Classifiers "A","a" were trained with annotations of all CAZy families for 16 microbial genomes; Classifiers "B","b" were trained with annotations for all CAZy families, except for the GT family members (which were not annotated for the Tammar Wallaby metagenome), for 16 genomes and the TW metagenome of plant biomass degraders; Classifiers "C","c" were trained with annotations for the GH families and CBMs for the 16 microbial genomes and three metagenomes of plant biomass degraders, as only these were annotated for the metagenomes. All CAZy-based classifiers were trained with available annotations for 64 genomes of non-biomass degraders. The Pfam-based classifiers were trained with 21 (meta-)genomes of biomass-degraders and 82 microbial genomes of non-degraders. For more details on the experimental set-up and the evaluation measures shown see the Methods section on performance evaluation.

	Presence/absence of Pfam domains	**Weighted Pfam domain representation**	**Presence/absence CAZy family representation**			**Weighted CAZy family representation**		
			A	**B**	**C**	**a**	**b**	**c**
nCV macro-accuracy	0.91	0.84	0.90	0.96	0.94	0.91	0.93	0.87
nCV recall	0.86	0.73	0.81	0.94	0.90	0.88	0.88	0.79
nCV true negative rate	0.96	0.96	0.98	0.98	0.98	0.95	0.98	0.95

families of the microbial genomes and the TW sample. In this setting only two species (*Thermomonospora curvata* and *Actinosynnema mirum*) were misclassified. These species remained misclassified with all six classifiers.

Using feature selection, we determined the CAZy families from the six $\text{eSVM}_{\text{CAZy}}$ classifiers that are most relevant for identifying microbial cellulose-degraders. Many of these GH families and CBMs are present in all (meta-)genomes (Figure 3.2).

This analysis identified further gene families known to be relevant for plant biomass degradation. Among them are cellulase-containing families (GH5, GH6, GH12, GH44, GH74), hemicellulase-containing families (GH10, GH11, GH26, GH55, GH81, GH115), families with known oligosaccharide/side-chain-degrading activities (GH43, GH65, GH67, GH95) and several CBMs (CBM3, -4, -6, -9, -10, -16, -22, -56). Several of these (GH6, GH11, GH44, GH67, GH74, CBM4, CBM6, CBM9) were consistently identified by at least half of the six classifiers as distinctive for plant biomass degraders. These might be considered signature genes of

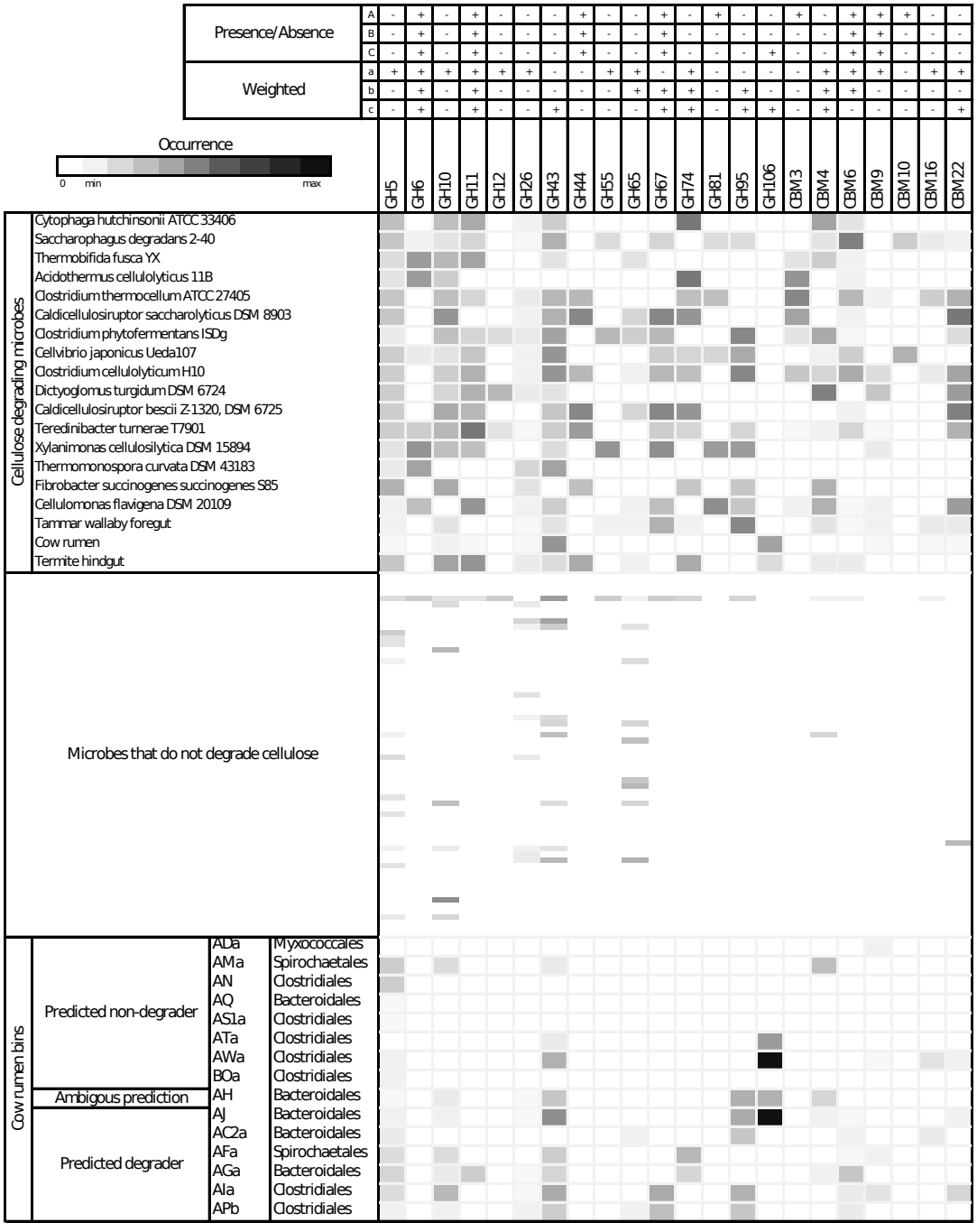


Figure 3.2: Frequencies of selected glycoside hydrolase (GH) families and carbohydrate binding modules (CBMs) in the (meta-) genome sequences. The data for each entry are rescaled by the total number of GH and CBM domains annotated to the microbial genome or metagenome. The coloring from black to grey indicates domains that are present in high to low amounts, respectively. White indicates absent domain families ("A", "a", "B", "b", "C", "c" as described in Table 3.2).

the plant biomass-degrading microorganisms we analyzed. Additionally, several GT, PL and CE domains were identified as relevant ($eSVM_{CAZY\_A}$: PL1, PL11 and CE5, "$eSVM_{CAZY\_B}$: CE5; $eSVM_{CAZY\_a}$: GT39, PL1 and CE2, $eSVM_{CAZY\_b}$: none). These CAZy families, as well as GH115 and CBM56, are not included in Figure 3.2, as they are not annotated for all sequences.

3.3.3 Identification of plant biomass degraders from a cow rumen metagenome

We used our method to predict the plant biomass-degrading capabilities for 15 draft genomes of uncultured microbes reconstructed from the metagenome of a microbial community adherent to switchgrass in cow rumen (Hess et al. 2011) (see Methods for the classification with an ensemble of SVM classifiers). The draft genomes represent genomes with more than 50% of the sequence reconstructed by taxonomic binning of the metagenome sample. The microbial community adherent to switchgrass is likely to be enriched with plant biomass degraders, as it was found to differ from the rumen fluid community in its taxonomic composition and degradation of switch grass after incubation in cow rumen had occurred. For identification of plant biomass-degrading microbes, we classified each draft genome individually with the $eSVM_{bPFAM}$ and $eSVM_{CAZY\_B}$ models, which had the highest macro-accuracy based on Pfam domain or CAZy family annotations, respectively. The $eSVM_{bPFAM}$ classifier assigned seven of the draft genomes to plant biomass degraders (Table 3.4). One of these, genome *APb*, was found by 16S rRNA analysis to be related to the fibrolytic species *Butyrivibrio fibrisolvens.* Four others (*AC2a*, *AGa*, *AJ* and *AH*) are of the order of Bacteroidales, and include all but one draft genomes affiliated to the Bacteroidales. The 6^{th} and 7^{th} predicted degrader, represented by genome *AIa* and *AWa*, belong to the Clostridiales, like genome *APb*. The $eSVM_{CAZY\_B}$ classifier also assigned five of these genomes to the plant biomass degraders. Additionally it classified genome *AH* as plant biomass-degrading, while being ambiguous in the assignment of *AFa* (Table 3.4). To validate these predictions, we searched the draft genomes for genes encoding 51 enzymatically active glycoside hydrolases characterized from the same rumen dataset (for the results of these experiments see Figure 3 in Hess

et al. (Hess et al. 2011). Genomes *AGa*, *AC2a*, *AJ* and *AIa* were all linked to different enzymes of varying specificities (Table 3.4). *AC2a* was linked to cellulose degradation, specifically to a carboxymethyl cellulose (CMC)-degrading GH5 endoglucanase as well as GH9 enzyme capable of degrading insoluble cellulosic substrates such as Avicel®. *AIa* demonstrated capabilities towards xylan and soluble cellulosic substrates with affiliations to four GH10 xylanases. Both *AGa* and *AJ* demonstrated broader substrate versatility and were linked to enzymes with capabilities towards cellulosic substrates CMC and Avicel® (GH5, GH9 and GH26), hemicellulosic substrates lichenan (β-1,3, β-1,4 β-glucan) and xylan (GH5, GH9 and GH10), as well as the natural feedstocks miscanthus and switchgrass (GH5 and GH9). Importantly, no carbohydrate-active enzymes were affiliated to draft genomes that were predicted to not possess plant biomass-degrading capabilities (Table 3.4). Overall, assignments were largely consistent between the two classifiers and supporting evidence for the capability to degrade plant biomass was found for five of the predicted degraders.

	AC2a	**AGa**	**AIa-2**	**AJ**	**APb**	**AFa**	**AH**	**AWa**	**ADa**	**AMa**	**AN**	**AQ**	**AS1**	**ATa**	**BOa**
$eSVM_{CAZY\_B}$	++	++	++	+	++	++	0	--	--	--	--	--	--	--	--
$eSVM_{bPFAM}$	++	++	++	++	++	-	++	+	--	-	--	--	--	-	--
CMC	GH5 (TW-33)	GH5 (TW-40) GH5 (MH-2)	GH10 (TW-34)	GH5 (TW-39) GH26 (TW-10) GH10 (TW-8)											
XYL		GH10 (TW-25)	GH10 (TW-30) GH10 (TW-31) GH10 (TW-37)	GH10 (TW-8)											
SWG		GH5 (TW-40) GH5 (MH-2)													
MIS	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
AVI	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
LIC		GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											

Table 3.4: Genome reconstructions from the metagenome of a microbial community adherent to switchgrass in the cow rumen were obtained by taxonomic binning of assembled sequences in the original study. Symbols depict the prediction outcome of a voting committee of the 5 $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers with the best macro-accuracy (see text for the description of the classifiers). ++: genome classified as plant biomass degrader by all classifiers; +: genome classified as plant biomass degrader by 4 out of 5 classifiers; 0: ambiguous prediction; -: genome classified as not plant biomass degrader by 4 out of 5 classifiers; --: genome classified as not plant biomass degrader by all classifiers. For every draft genome, the presence of genes encoding glycoside hydrolases with verified enzymatic activity for different substrates in this study (Hess et al. 2011) is indicated. The genome and substrate names correspond to those of Figure 3 and Table S6 of the study.
Hydrolytic activity detected on:
(CMC) 1% (w/v) carboxymethyl cellulose agar.
(XYL) 1% (w/v) Xylan.
(SWG) 1% (w/v) IL-Switchgrass.
(MIS) 1% (w/v) IL-Miscanthus.
(AVI) 1% (w/v) IL-Avicel.

3.3.4 Timing experiments

Our method uses annotations with Pfam domains or CAZy families as input. Generating these by similarity-searches with profile HMMs rather than with BLAST provides a better scalability for next-generation sequencing data sets. HMM databases such as dbCAN contain a representation of entire protein families rather than of individual gene family members, which largely decreases the number of entries one has to compare against. For example, searching the ORFs of the *Fibrobacter succinogenes* genome (Suen et al. 2011) for similarities to CAZy families with the dbCAN HMM models took 23 seconds on an Intel® Xeon® 1.6 GHz CPU. In comparison, searching for similarities to CAZy families by BLASTing the same set of ORFs against all sequences with CAZy family annotation of the NCBI non-redundant protein database (downloaded from ftp://ftp.ncbi.nih.gov/blast/db/FASTA on April 19th 2011) on the same machine required approximately 1 hour and 55 minutes, a difference of two orders of magnitude. Because of their better scalability and also because they are well-established for identifying protein domains or gene families (Haft et al. 2001; Punta et al. 2012; Schultz et al. 2000), we recommend the use of HMM-based similarities and annotations as input to our method.

3.4 Discussion

We investigated the value of information about the presence-or-absence of CAZy families and Pfam protein domains, as well as information about their relative abundances, for the identification of lignocellulose degraders. Classifiers trained with CAZy family or Pfam domain annotations allowed an accurate identification of plant biomass degraders and determined similar domains and CAZy families as being most distinctive. Many of these are recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as GH6, members of the GH5 family and to a lesser extent GH44 and GH74. In contrast to widely accepted paradigms for microbial cellulose degradation, recent genome analysis of cellulolytic bacteria has identified examples (i.e. *Fibrobacter*) where there is an absence of genes encoding

exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures (Wilson 2009). In addition, these exo-acting families and cellulosomal structures have had a low representation or are entirely absent from sequenced gut metagenomes. Our method also finds the exo-acting cellobiohydrolases GH7 and GH48 to be less important. GH7 represents fungal enzymes, so its absence makes sense; however, the lower importance assigned to GH48 is interesting. The role of GH48 is believed to be of high importance, although recent research has raised questions. Olson *et al.* (Olson et al. 2010) have found that a complete solubilization of crystalline cellulose can occur in *Clostridium thermocellum* without the expression of GH48, albeit at significantly lower rates. Furthermore, genome analysis of cellulose-degrading microbes *Cellvibrio japonicus* (DeBoy et al. 2008) and *Saccharophagus degradans* (Taylor et al. 2006) have determined the presence of only non-reducing end enzymes (GH6) and an absence of a reducing end cellobiohydrolase (GH48), suggesting that the latter are not essential for all cellulolytic enzyme systems.

While we have focused on cellulose degradation, our method has also identified enzymes that degrade other plant polysaccharides as being relevant, such as hemicellulose (GH10, GH11, GH12, GH26, GH55, GH81, CE4), pectins (PL1, GH88 and GH43), oligosaccharides (GH3, GH30, GH39, GH43, GH65, GH95) and the side-chains attached to noncellulosic polysaccharides (GH67, GH88, GH106). This was expected, since many cellulose-degrading microbes produce a repertoire of different glycoside hydrolases, lyases and esterases (see, for example, (DeBoy et al. 2008; Taylor et al. 2006)) that target the numerous linkages that are present within different plant polysaccharides, which often exist in tight cross-linked forms within the plant cell wall. The results from our method add further weight to this. The observation of numerous CBMs being relevant in the CAZy analysis also agrees with previous findings that many different CBM–GH combinations are possible in bacteria. Moreover, recent reports have demonstrated that the targeting actions of CBMs have strong proximity effects within cell wall structures, i.e. CBMs directed to a cell wall polysaccharide (e.g. cellulose) other than the target substrate of their appended glycoside hydrolase (e.g. xylanase) can promote enzyme action against the target substrate (e.g. xylan) within the cell wall (Hervé et al. 2010). This provides explanations as to why cellulose-directed

CBMs are appended to many non-cellulase cell wall hydrolases.

Several Pfam domains of unknown function (DUFs) or protein domains which have not previously been associated with cellulose degradation are predicted as being relevant. These include transferases (PF01704) and several putative lipoproteins (DUF4352), some of which have predicted binding properties (NlpC/P60 family: PF00877, PASTA domain: PF03793). The functions of these domains in relation to cellulose degradation are not known, but possibilities include binding to cellulose, binding to other components of the cellulolytic machinery or interaction with the cell surface.

Another result of our study are the classifiers for identifying microbial lignocellulose-degraders from genomes of cultured and uncultured microbial species reconstructed from metagenomes. Classification of draft genomes reconstructed from switchgrass-adherent microbes from cow rumen with the most accurate classifiers predicted six or seven of these to represent plant biomass-degrading microbes, including a close relative to the fibrolytic species *Butyrivibrio fibrisolvens*. Cross-referencing of all draft genomes against a catalogue of enzymatically active glycoside hydrolases provided a degree of method validation and was in majority agreement with our predictions. Four genomes (*AGa*, *AC2a*, *AJ* and *AIa*) predicted positive were linked to cellulolytic and/or hemicellulolytic enzymes, and importantly no genomes that were predicted negative were linked to carbohydrate-active enzymes from that catalogue of enzymatically active enzymes. Also, no connections to carbohydrate-active enzymes from that catalogue were observed for the three genomes (*AFa*,*AH* and *AWa*) where ambiguous predictions were made. As both draft genomes as well as the catalogue of carbohydrate active enzymes in cow rumen are incomplete, in addition to our training data not covering all plant-biomass-degrading taxa, such ambiguous assignments might be better resolvable with more information in the future.

We trained a previous version of our classifier with the genome of *Methanosarcina barkeri fusaro* incorrectly labeled as a plant biomass degrader, according to information provided by IMG. In cross-validation experiments, our method correctly assigned *M. barkeri* to be a non-plant biomass-degrading species. We labeled *Thermonospora curvata* as a plant biomass degrader and *Actinosynnema mirum* as non-degrader according to information from the literature (see Additional file

1: Table S1). Both were misassigned by all classifiers in the cross-validation experiments. However, in a recent work by Anderson *et al.* (Anderson et al. 2012) it was shown that in cellulose activity assays *A. mirum* could degrade various cellulose substrates. In the same study, *T. curvata* did not show cellulolytic activity against any of these substrates, contrary to previous beliefs (Ivanova et al. 2011). The authors found out that the cellulolytic *T. curvata* strain was in fact a *T. fusca* strain. Thus, our method could correctly assign both strains despite of the incorrect phenotypic labeling. The genome of *Postia placenta*, the only fungal plant biomass degrader of our data set was misassigned in the Pfam-based SVM analyses. Fungi possess cellulases not found in prokaryotic species (Duan et al. 2010) and might employ a different mechanism for plant biomass degradation (Lynd et al. 2002; Wilson 2009). Indeed, in our data set, *Postia placenta* is annotated with the cellulase-containing GH5 family and xylanase GH10, but the hemicellulase family GH26 does not occur. Furthermore, the (hemi-)cellulose binding CBM domains CBM6 and CBM_4_9, which were identified as being relevant for assignment to lignocellulose degraders with the $\text{eSVM}_{\text{bPFAM}}$ classifier, are absent. All of the latter ones, GH26, CBM6 and especially CBM4 and CBM9, occur very rarely in eukaryotic genome annotations, according to the CAZy database.

3.5 Conclusions

We have developed a computational technique for the identification of Pfam protein domains and CAZy families that are distinctive for microbial plant biomass degradation from (meta-)genome sequences and for predicting whether a (draft) genome of cultured or uncultured microorganisms encodes a plant biomass-degrading organism. Our method is based on feature selection from an ensemble of linear L1-regularized SVMs. It is sufficiently accurate to detect errors in phenotype assignments of microbial genomes. However, some microbial species remained misclassified in our analysis, which indicates that further distinctive genes and pathways for plant biomass degradation are currently poorly represented in the data and could therefore not be identified.

To identify a lignocellulose degrader from the currently available data, the presence of a few domains, many of which are already known, is sufficient. The iden-

tification of several protein domains which have so far not been associated with microbial plant biomass degradation in the Pfam-based SVM analyses as being relevant may warrant further scrutiny. A difficulty in our study was to generate a sufficiently large and correctly annotated dataset to reach reliable conclusions. This means that the results could probably be further improved in the future, as more sequences and information on plant biomass degraders become available. The method will probably also be suitable for identifying relevant gene and protein families of other phenotypes.
The prediction and subsequent validation of three Bacteroidales genomes to represent cellulose-degrading species demonstrates the value of our technique for the identification of plant biomass degraders from draft genomes from complex microbial communities, where there is an increasing production of genome assemblages for uncultured microbes. These to our knowledge represent the first cellulolytic Bacteroidetes-affiliated lineages described from herbivore gut environments. This finding has the potential to influence future cellulolytic activity investigations within rumen microbiomes, which has for the greater part been attributed to the metabolic capabilities of species affiliated to the bacterial phyla Firmicutes and Fibrobacteres.

3.5.1 Methods

Annotation

We annotated all protein coding sequences of microbial genomes and metagenomes with Pfam protein domains (Pfam-A 26.0) and Carbohydrate-Active Enzymes (CAZymes) (Cantarel et al. 2009; Lynd et al. 2002; Punta et al. 2012; Wilson 2009). The CAZy database contains information on families of structurally related catalytic modules and carbohydrate binding modules (CBMs) or (functional) domains of enzymes that degrade, modify or create glycosidic bonds. HMMs for the Pfam domains were downloaded from the Pfam database. Microbial and metagenomic protein sequences were retrieved from IMG 3.4 and IMG/M 3.3 (Markowitz, I. M. Chen, et al. 2012; Markowitz, Ivanova, et al. 2008). HMMER 3 (Finn, Clements, et al. 2011) with gathering thresholds was used to annotate the samples with Pfam domains. Each Pfam family has a manually de-

fined gathering threshold for the bit score that was set in such a way that there were no false-positives detected. For annotation of protein sequences with CAZy families, the available annotations from the database were used. For annotations not available in the database, HMMs for the CAZy families were downloaded from dbCAN (http://csbl.bmb.uga.edu/dbcan) (Yin et al. 2012). To be considered a valid annotation, matches to Pfam and dbCAN protein domain HMMs in the protein sequences were required to be supported by an e-value of at least 1e-02 and a bit score of at least 25. Additionally, we excluded matches to dbCAN HMMs with an alignment longer than 100 bp that did not exceed an e-value of 1e-04. Multiple matches of one and the same protein sequence against a single Pfam or dbCAN HMM exceeding the thresholds were counted as one annotation.

3.5.2 Phenotype annotation of lignocellulose-degrading and non-degrading microbes

We defined genomes and metagenomes as originating from either lignocellulose-degrading or non-lignocellulose-degrading microbial species based on information provided by IMG/M and in the literature. For every microbial genome and metagenome, we downloaded the genome publication and further available articles (Additional file 1: Table S1). We did not consider genomes for which no publications were available. For cellulose-degrading species annotated in IMG, we verified these assignments based on these publications. We used text search to identify the keywords "cellulose", "cellulase", "carbon source", "plant cell wall" or "polysaccharide" in the publications for non-cellulose-degrading species. We subsequently read all articles that contained these keywords in detail to classify the respective organism as either cellulose-degrading or non-degrading. Genomes that could not be unambiguously classified in this manner were excluded from our study.

3.5.3 Classification with an ensemble of support vector machine classifiers

The SVM is a supervised learning method that can be used for data classification (Boser et al. 1992; Cortes et al. 1995). Here, we use an L1-regularized L2-loss SVM, which solves the following optimization problem for a set of instance-label pairs $(\vec{x_i}, y_i), \vec{x_i} \in \mathbf{R^n}, y_i \in \{-1, +1\}, i = 1, \ldots, l$:

$$\min_{\vec{w}} \|\vec{w}\|_1 + C \sum_{i=1}^{l} (max(0, 1 - y_i \vec{w}^T \vec{x_i}))^2,$$

where $C \geq 0$ is a penalty parameter. This choice of the classifier and regularization term results in sparse models, where non-zero components of the weight vector $\vec{w}$ are important for discrimination between the classes (Yaun et al. 2010). SVM classification was performed using the LIBLINEAR package (Fan et al. 2008). The components of $\vec{x_i}$ are either binary valued and represent the presence or absence of protein domains, or continuous-valued and represent the frequency of a particular protein domain or gene family relative to the total number of annotations. All features were normalized by dividing by the sum of all vector entries and subsequently scaled, such that the value of each feature was within the range [0,1]. The label +1 was assigned to genomes and metagenomes of plant biomass-degrading microorganisms, the label -1 to all sequences from non-degrading ones. Classification of the draft genomes assembled from the fiber-adherent microbial community from cow rumen was performed with a voting committee of multiple models with different settings for the penalty parameter C that performed comparably well. A majority vote of the 5 most accurate models was used here obtained in a single cross-validation run with different settings of the penalty parameter C.

3.5.4 Performance evaluation

The assignment accuracy of a classifier was determined with a standard nested cross-validation (nCV) setup (Ruschhaupt et al. 2004). In nCV, an outer cross-validation loop is organized according to the leave-one-out principle: In each step,

one data point is left out. In an inner loop, the optimal parameters for the model (here, the penalty parameter C) are sought, in a second cross-validation experiment with the remaining data points. For determination of the best setting for the penalty parameter C, values for $C = 10^x$, x = -3.0, -2.5, -2.25, ..., 0 were tried. Values of the parameter C larger than 1 were not tested extensively, as we found that they resulted in models with similar accuracies. This is in agreement with the Liblinear tutorial in the appendix of (Fan et al. 2008) which states that once the parameter C exceeds a certain value, the obtained models have a similar accuracy. The SVM with the penalty parameter setting yielding the best assignment accuracy was used to predict the class membership of the left out data point. The class membership predictions for all data points were used to determine the assignment accuracy of the classifier, based on their agreement with the correct assignments. For this purpose, the result of each leave-one-out experiment was classified as either a true positive (TP – correctly predicted lignocellulose degraders), true negative (TN – correctly predicted non-degraders), false positive (FP – non-degraders predicted to be degraders) or a false negative assignment (FN – degraders predicted to be non-degraders). The recall of the positive class and the true negative rate of the classifier were calculated according to the following equations:

$$Recall = \frac{TP}{TP + FN}$$

$$True\ negative\ rate = \frac{TN}{TN + FP}$$

The average of the recall and the true negative rate, the macro-accuracy, was used as the assignment accuracy to assess the overall performance:

$$MACC = \frac{Recall + True\ negative\ rate}{2}$$

Subsequently, we identified the settings for the penalty parameter C with the best macro-accuracy by leave-one-out cross-validation. The parameter settings resulting in the most accurate models were used to each train a separate model on the entire data set. Prediction of the five best models were combined to form a voting committee and used for the classification of novel sequence samples such as

the partial genome reconstructions from the cow rumen metagenome of switchgrass adherent microbes (see Additional file 2: Table S2 for an evaluation and meta-parameter settings of these ensembles of classifiers).

3.5.5 Feature selection

An SVM model can be represented by a sparse weight vector $\vec{w}$. The positive and negative components of $\vec{w}$, the 'feature weights', specify the relative importance of the protein domains or CAZy families for discrimination between plant biomass-degrading and non-plant biomass-degrading microorganisms. To determine the most distinctive features for the positive class (that is, the lignocellulose degraders), we selected all features that received a positive weight in weight vectors of the majority of the five most accurate models. This ensemble of models was also used for classification of the cow rumen draft genomes of uncultured microbes (see Classification with a SVM).

3.6 Supplementary material

The supplementary material can be found in the original version of the paper in the appendix and online at `https://doi.org/10.1186/1754-6834-6-24`.

CHAPTER 4

From genomes to phenotypes: Traitar, the microbial trait analyzer

Status	**published as pre-print at the time of writing, but has since been published in the journal mSystems (Weimann et al. 2016a))**
Journal	bioRxiv
Citation	Aaron Weimann, Kyra Mooren, Jeremy Frank, Phillip B Pope, Andreas Bremges, Alice C McHardy (2016). From genomes to phenotypes: Traitar, the microbial trait analyzer. *bioRxiv*, **043315**.
URL	http://biorxiv.org/content/early/2016/07/26/043315
Own contribution	75% Wrote the manuscript (with ACM, AB) Conceived and designed the experiments (with ACM) Implemented the Traitar software and web service Mapping of Bergey and GIDEON species and phenotypes with public strains and genomes (with KM) Implemented and conducted the experiments (with JF, AB) Interpreted the classification results, found biomarkers, etc. (with AB, PBP, ACM)

4.1 Abstract

The number of sequenced genomes is growing exponentially, profoundly shifting the bottleneck from data generation to genome interpretation. Traits are often used to characterize and distinguish bacteria, and are likely a driving factor in microbial community composition, yet little is known about the traits of most microbes. We describe Traitar, the microbial trait analyzer, which is a fully automated software package for deriving phenotypes from the genome sequence. Traitar provides phenotype classifiers to predict 67 traits related to the use of various substrates as carbon and energy sources, oxygen requirement, morphology, antibiotic susceptibility, proteolysis and enzymatic activities. Furthermore, it suggests protein families associated with the presence of particular phenotypes. Our method uses L1-regularized L2-loss support vector machines for phenotype assignments based on phyletic patterns of protein families and their evolutionary histories across a diverse set of microbial species. We demonstrate reliable phenotype assignment for Traitar to bacterial genomes from 572 species of 8 phyla, also based on incomplete single-cell genomes and simulated draft genomes. We also showcase its application in metagenomics by verifying and complementing a manual metabolic reconstruction of two novel Clostridiales species based on draft genomes recovered from commercial biogas reactors. Traitar is available at `https://github.com/hzi-bifo/traitar`.

4.2 Introduction

Microbes are often characterized and distinguished by their traits, for instance, in *Bergey's Manual of Systematic Bacteriology* (Goodfellow et al. 2012). A trait or phenotype can vary in complexity; for example, it can refer to the degradation of a specific substrate or the activity of an enzyme inferred in a lab assay, the respiratory mode of an organism, the reaction to Gram staining or antibiotic resistances. Traits are also likely driving factor for microbial community composition (Martiny et al. 2015). Microbial community members with varying metabolic capabilities can aid in waste water treatment, bioremediation of soils and promotion of plant growth (Bai et al., 2015; Narihiro and Sekiguchi, 2007; Olapade and Ronk, 2015); in the cow rumen microbiota, bacterial cellulose degraders influence the ability to process plant biomass material (Hess et al.

2011). In the Tammar wallaby foregut microbiome, the dominant bacterial species is implicated in the lower methane emissions produced by wallaby compared to ruminants (P. Pope et al. 2010).

In addition to the exponential growth of available sequenced microbial genome isolates, metagenome and single cell genome sequencing further contributes to the increasing number of available genomes. For the recovery of genomes from metagenomes (GFMs), computational methods based on e.g. differential read coverage and k-mer usage were developed (Alneberg et al. 2014; Cleary et al. 2015; Gregor et al. 2016; Imelfort et al. 2014; Kang et al. 2015; Nielsen et al. 2014). In addition, single-cell genomics provides another culture-independent analysis technique and also allows, although often fragmented, genome recovery for less abundant taxa in microbial communities (Lasken et al. 2014; Rinke et al. 2013). Together, these developments profoundly shift the analytical bottleneck from data generation to interpretation.

The genotype–phenotype relationships for some microbial traits have been well studied. For instance, bacterial motility is attributed to the proteins of the flagellar apparatus (Macnab 2003). We have recently shown that delineating such relationships from microbial genomes and accompanying phenotype information with statistical learning methods enables the accurate prediction of the plant biomass degradation phenotype and the *de novo* discovery of both known and novel protein families that are relevant for the realization of the plant biomass degradation phenotype (Konietzny, P. B. Pope, et al. 2014; Weimann, Trukhina, et al. 2013). However, a fully automated software framework for prediction of a broad range of traits from only the genome sequence is currently missing. Additionally, horizontal gene transfer, a common phenomenon across bacterial genomes, has not been utilized to improve trait prediction so far. Traits with their causative genes may be transferred from one bacterium to the other (Ochman et al. 2000; Pal et al. 2005) (e.g. for antibiotic resistances (J. L. Martinez 2008)) and the vertically transferred part of a bacterial genome might be unrelated to the traits under investigation (Barker et al. 2005; Harvey et al. 1991; Martiny et al. 2015).

Here we present Traitar, the microbial trait analyzer: an easy-to-use, fully automated software framework for the accurate prediction of currently 67 phenotypes directly from the genome sequence (Figure 4.1).

We used phenotype data from the microbiology section of the Global Infectious Disease and Epidemiology Network (GIDEON) – a resource dedicated to the diagnosis, treatment and teaching of infectious diseases and microbiology (Berger 2005) – for training phenotype classification models on the protein family annotation of a large number

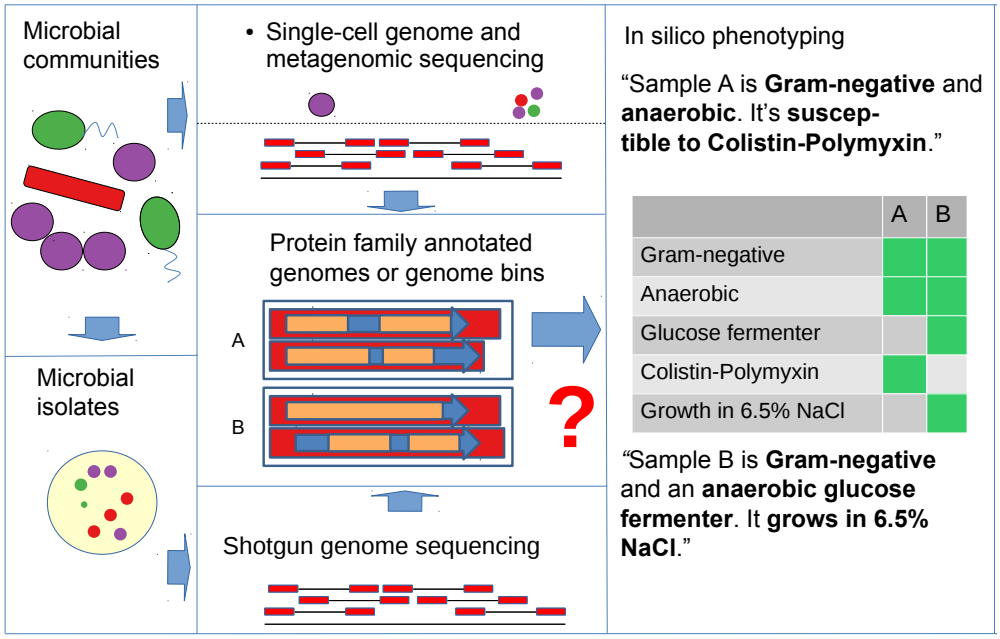


Figure 4.1: Traitar can be used to phenotype microbial community members based on genomes recovered from single-cell sequencing or (metagenomic) environmental shotgun sequencing data or of microbial isolates. Traitar provides classification models based on protein family annotation for a wide variety of different phenotypes related to the use of various substrates as source of carbon and energy for growth, oxygen requirement, morphology, antibiotic susceptibility and enzymatic activity.

of sequenced genomes of microbial isolates (predominantly bacterial pathogens). We investigated the effect of incorporating ancestral protein family gain and losses into the model inference on classification performance, to allow consideration of horizontal gene transfer events in inference of phenotype-related protein families and phenotype classification. We rigorously tested the performance of our software in cross-validation experiments, on further test data sets and for different taxonomic ranks. To test Traitar's applicability beyond the bacteria represented in GIDEON, we subsequently applied it to several hundred bacteria described in Bergey's systematic bacteriology (Goodfellow et al. 2012). We used Traitar to phenotype bacterial single amplified genomes (SAGs) and simulated incomplete genomes to investigate its potential for phenotyping microbial samples with incomplete genome sequences. We characterized two novel Clostridiales species of a biogas reactor community with Traitar, based on their genomes recovered

with metagenomics. This verified and complemented a manual metabolic reconstruction. As Traitar furthermore suggests protein families associated with the presence of a particular phenotype, we discuss the protein families Traitar identified for several phenotypes, namely for 'Motility', 'Nitrate to nitrite' conversion and 'L-arabinose' fermentation.

Traitar is implemented in Python 2.7. It is freely available under the open-source GPL 3.0 license at `https://github.com/hzi-bifo/traitar` and as a Docker container at `https://hub.docker.com/r/aweimann/traitar`. A Traitar web service can be accessed at `https://research.bifo.helmholtz-hzi.de/traitar`.

4.3 Results

4.3.1 The Traitar software

We begin with a description of the Traitar software and phenotype classifiers. Traitar predicts the presence or absence of a phenotype, i.e. assigns a phenotype label, for 67 microbial traits to every input sequence sample (Table 4.1, Supplementary Table 1). For each of these traits, Traitar furthermore suggests candidate protein families associated with its realization, which can be subject of experimental follow-up studies.

For phenotype prediction, Traitar uses one of two different classification models. We trained the first classifier – the phypat classifier – on the protein and phenotype presence & absence labels from 234 bacterial species (Methods – Phenotype models). The second classifier – the phypat+PGL classifier – was trained using the same data and additionally information on evolutionary protein family and phenotype gains and losses. The latter were determined using maximum likelihood inference of their ancestral character states on the species phylogeny (Methods – Ancestral protein family and phenotype gains and losses).

The input to Traitar is either a nucleotide sequence FASTA file for every sample, which is run through gene prediction software, or a protein sequence FASTA file. Traitar then annotates the proteins with protein families. Subsequently, it predicts the presence or absence of each of the 67 traits for every input sequence. Note that Traitar doesn't require a phylogenetic tree for the input samples.

Finally, it associates the predicted phenotypes with the protein families that contributed to these predictions (Figure 4.2). A parallel execution of Traitar is supported by GNU parallel (Tange, 2011). The Traitar annotation procedure and the training of the phe-

notype models are described in more detail below (Methods – Traitar software).

Table 4.1: The 67 traits available in Traitar for phenotyping. We grouped each of these phenotypes into a microbiological or biochemical category.

Phenotype$_{(a)}$	Category$_{(b)}$
Alkaline phosphatase	Enzyme
Beta hemolysis	
Coagulase production	
Lipase	
Nitrate to nitrite	
Nitrite to gas	
Pyrrolidonyl-beta-naphthylamide	
DNase	
Bile-susceptible	Growth
Colistin-Polymyxin susceptible	
Growth at 42°C	
Growth in 6.5% NaCl	
Growth in KCN	
Growth on MacConkey agar	
Growth on ordinary blood agar	
Mucate utilization	
Arginine dihydrolase	Growth: Amino Acid
Indole	
Lysine decarboxylase	
Ornithine decarboxylase	
Acetate utilization	Growth:Carboxylic Acid
Citrate	
Malonate	
Tartrate utilization	
Gas from glucose	Growth:Glucose
Glucose fermenter	
Glucose oxidizer	
Methyl red	

Voges Proskauer	
Cellobiose	Growth:Sugar
D-Mannitol	
D-Mannose	
D-Sorbitol	
D-Xylose	
Esculin hydrolysis	
Glycerol	
Lactose	
L-Arabinose	
L-Rhamnose	
Maltose	
Melibiose	
myo-Inositol	
ONPG (beta galactosidase)$_{(d)}$	
Raffinose	
Salicin	
Starch hydrolysis	
Sucrose	
Trehalose	
Urea hydrolysis	
Bacillus or coccobacillus	Morphology
Coccus	
Coccus - clusters or groups predominate	
Coccus - pairs or chains predominate	
Gram negative	
Gram positive	
Motile	
Spore formation	
Yellow pigment	
Aerobe	Oxygen
Anaerobe	
Capnophilic	

Facultative	
Catalase	Oxygen:Enzyme
Oxidase	
Hydrogen sulfide	Product
Casein hydrolysis	Proteolysis
Gelatin hydrolysis	

(a) GIDEON phenotypes with at least 10 presence and 10 absence labels

(b) Phenotypes assigned to microbiological / biochemical categories

(c) ONPG: o-Nitrophenyl-β-D-galatopyranosid

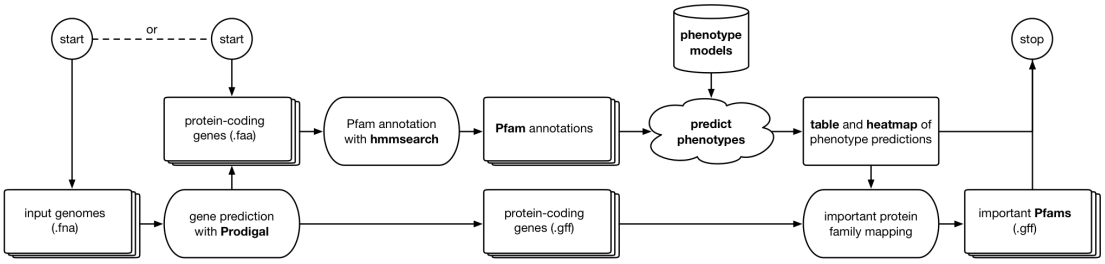


Figure 4.2: Work flow of Traitar. Input to the software can be genome sequene samples in nucleotide or amino acid FASTA format. Traitar predicts phenotypes based on pre-computed classification models and provides graphical and tabular output. In the case of nucleotide input, the protein families that are important for the phenotype predictions will be further mapped to the predicted protein-coding genes.

4.3.2 Evaluation

We evaluated the two Traitar classifiers using ten-fold nested cross-validation on 234 bacterial species found in GIDEON (GIDEON I). The determined macro-accuracy (the accuracy balanced over all phenotypes) for the 67 GIDEON phenotypes was 82.6% for the phypat classifier and 85.5% for the phypat+PGL classifier; the accuracy (fraction of correct assignments averaged over all tested samples) for phypat was 88.1%, in comparison to 89.8% for phypat+PGL (Methods – Evaluation metrics; Table 4.2). Notably, Traitar classified 53 phenotypes with more than 80% macro-accuracy and

Table 4.2: We evaluated the Traitar phypat and phypat+PGL phenotype classifiers and a consensus vote of both classifiers for 234 bacteria described in the Global Infectious Disease and Epidemiology Online Network (GIDEON) in a 10-fold nested cross-validation using different evaluation measures (Methods – Evaluation). Subsequently, we tested another 42 bacteria from GIDEON and 296 bacteria described in Bergey's manual of systematic bacteriology for an independent performance assessment of the two classifiers. We only report the overall accuracy for the evaluation of the classifiers on the data from Bergey's, as insufficient phenotype labels (less than 5 with a negative and positive label, respectively) were available for several phenotypes, to enable a comparable macro-accuracy calculation to the other data sets (Supplementary Table 1).

Data set (# bacteria)	Classifier	Macro-accuracy	Accuracy	Recall Phenotype+	Recall Phenotype-
GIDEON I (234)	phypat	82.6	88.1	86.1	91.4
	phypat+PGL	**85.5**	**89.8**	**87.8**	90.9
	consensus	83.0	88.8	82.2	**95.4**
GIDEON II (42)	Phypat	85.3	87.5	84.9	90.2
	phypat+PGL	**86.7**	**87.9**	**86.3**	89.7
	consensus	85.7	87.2	80.8	**93.7**
Bergey's (296)	phypat	NA	**72.9**	**74.6**	71.2
	phypat+PGL	NA	72.4	74	70.8
	consensus	NA	**72.9**	66.6	**79.2**

26 phenotypes with at least 90% macro-accuracy with one of the two classifiers (Figure 4.2, Supplementary Table 2). Phenotypes that could be predicted with very high confidence included the outcome of a 'Methyl red' test, 'Spore formation', oxygen requirement (i.e. 'Anaerobe' and 'Aerobe'), 'Growth on MacConkey agar' or 'Catalase'. Some phenotypes proved to be difficult to predict (60-70% macro-accuracy), which included 'DNAse', 'myo-Inositol' or 'Yellow pigment' and 'Tartrate utilization', regardless of which classifier was used. This might be caused by the relatively small number (<20) of positive (phenotype present) examples that were available.

For an independent assessment of Traitar's classification performance we next tested Traitar on 42 bacterial species that had phenotype information available in GIDEON (GIDEON II), but were not used for learning the phenotype models (The Traitar software – Annotation). For calculation of the macro-accuracy, we considered only phenotypes represented by at least five phenotype-positive and five phenotype-negative bacteria. On these data, Traitar predicted the phenotypes with a macro-accuracy of

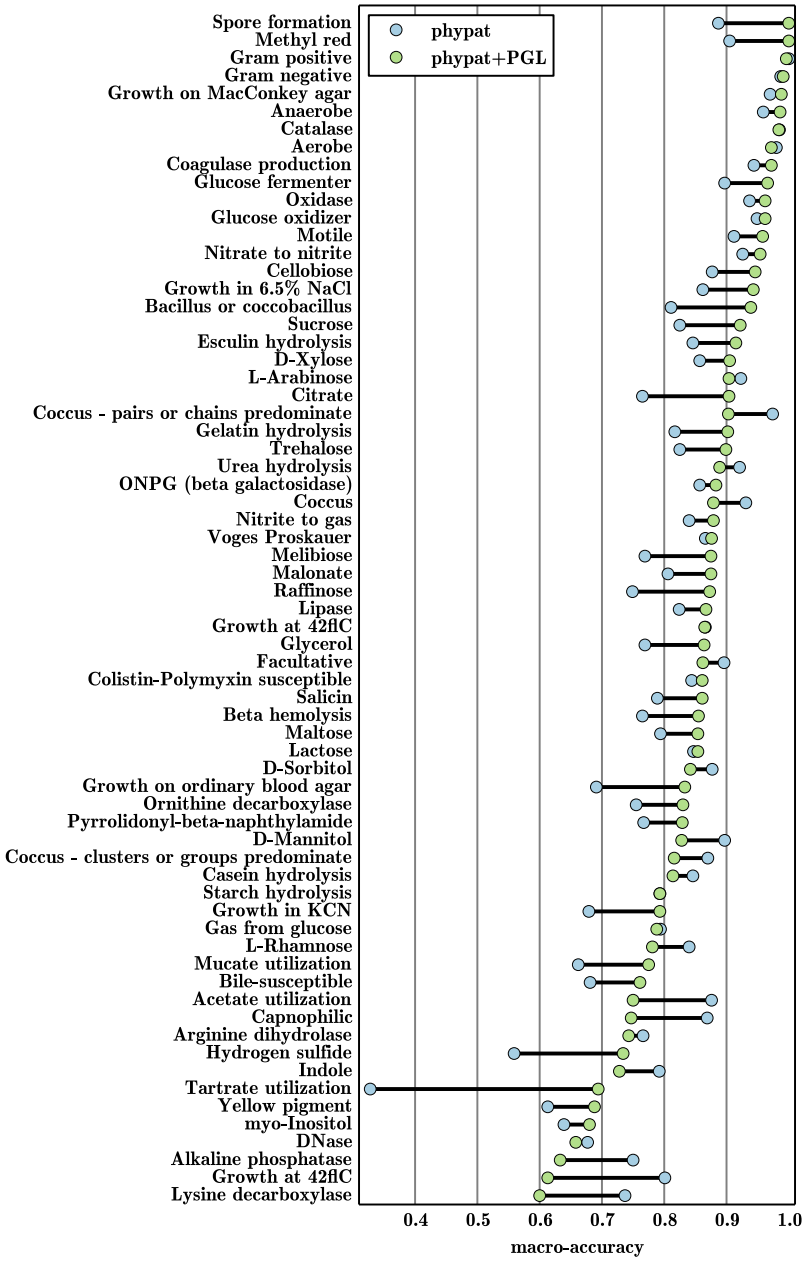


Figure 4.3: Macro-accuracy for each phenotype for the Traitar phypat and phypat+PGL phenotype classifiers determined in nested cross-validation on 234 bacterial species described in the Global Infectious Disease and Epidemiology Online Network (Methods – Evaluation metrics, Table 4.1, Supplementary Table 1).

85.3% with the phypat classifier and 86.7% with the phypat+PGL classifier, and accuracies of 87.5% and 87.9%, respectively (Table 4.2). To investigate the performance of Traitar for bacterial genomes from a different data source, we next determined from two volumes of Bergey's Manual of Systematic Bacteriology, namly 'The Proteobacteria' and 'The Firmicutes', the phenotypes of further sequenced bacteria that were not in our GIDEON I and II data sets (Supplementary Table 1, 4). In total, we thus identified phenotypes for another 296 sequenced bacterial species (The Traitar software – Annotation). Also for these bacteria, Traitar performed well but was less reliable than before, with accuracies for the phypat classifier of 72.9% and 72.1% for the phypat+PGL classifier (Table 4.2). This is likely due to the taxonomic differences of bacteria listed in GIDEON and Bergey's and also because most of the bacteria in Bergey's have only draft genomes available for phenotyping.

When combining the predictions of the phypat and phypat+PGL classifiers into a consensus vote, Traitar assigns phenotypes more reliably, while predicting less phenotype labels compared to the individuals classifiers (Table 2). Depending on the use case, Traitar can be used with performance characterized by different trade-offs between the recall of the phenotype-positive and the phenotype-negative classes.

4.3.3 Performance per taxon at different ranks of the taxonomy

We investigated the performance of Traitar across the part of the bacterial tree of life represented in our data set. For this purpose, we evaluated the nested cross-validation performance of the phypat and phypat+PGL classifiers at different ranks of the NCBI taxonomy. For a given GIDEON taxon, we pooled all bacterial species that are descendants of this taxon. Figure 4.4 shows the accuracy estimates projected on the NCBI taxonomy from the domain level down to individual families. Notably, the accuracy of the phypat+PGL (phypat) classifier for the phyla covered by at least five bacterial species showed low variance and was high across all phyla, ranging from 84% (81%) for Actinobacteria over 90% (89%) for Bacteroidetes, 89% (90%) for Proteobacteria, 91% (90%) for Firmicutes to 91% (86%) for Tenericutes.

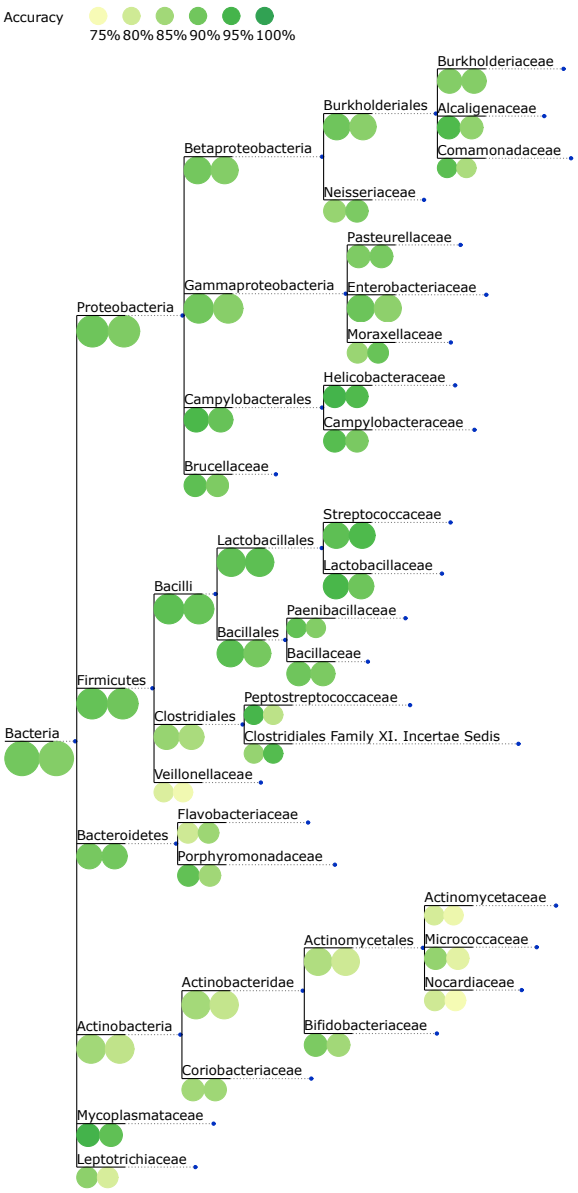


Figure 4.4: Classification accuracy for each taxon at different ranks of the NCBI taxonomy. For better visualization of names for the internal nodes, the taxon names are displayed on branches leading to the respective taxon node in the tree. The nested cross-validation accuracy obtained with Traitar for 234 bacterial species described in the Global Infectious Disease and Epidemiology Online Network was projected onto the NCBI taxonomy down to the family level. Colored circles at the tree nodes depict the performance of the phypat+PGL classifier (left-hand circles) and the phypat classifier (right-hand circles). The size of the circles reflects the number of species per taxon.

4.3.4 Phenotyping incomplete genomes

GFMs or SAGs are often incomplete and thus we analyzed the effect of missing genome assembly parts onto the performance of Traitar. Rinke *et al.* used a single-cell sequencing approach to analyze poorly characterized parts of the bacterial and archaeal tree of life, the so-called 'microbial dark matter' (Rinke et al. 2013). They pooled 20 SAGs from the candidate phylum Cloacimonetes, formerly known as WWE1, to generate joint – more complete – genome assemblies that had at least a genome-wide average nucleotide identity of 97% and belonged to a single 16S-based operational taxonomic unit, namely *Cloacamonas acidaminovorans (Pelletier et al., 2008).*

According to our predictions based on the joint assembly of the single-cell genomes, *C. acidaminovorans* is Gram-negative and is adapted to an anaerobic lifestyle, which agrees with the description by Rinke *et al.* (Figure 4.5). Traitar further predicted 'Arginine dihydrolase' activity, which is in line with the characterization of the species as an amino acid degrader (Rinke et al. 2013). Remarkably, the prediction of a bacillus or coco-bacillus shape agrees with the results of Limam *et al.* (Limam et al. 2014), who used a WWE1-specific probe and characterized the samples with fluorescence *in situ* hybridization. They furthermore report that members of the Cloacimonetes candidate phylum are implicated in anaerobic digestion of cellulose, primarily in early hydrolysis, which is in line with the very limited carbohydrate degradation spectrum found by Traitar.

Subsequently, we compared the predicted phenotypes for the SAGs to the predictions for the joint assembly. The phypat classifier recalled more of the phenotype predictions of the joint assembly based on the SAGs than the phypat+PGL classifier. However, the phypat+PGL classifier made fewer false positive predictions (Figure 4.6 a).

In the next experiment, we inferred phenotypes based on simulated GFMs, by subsampling from the coding sequences of each of the 42 bacterial genomes (GIDEON II). Starting with the complete set of coding sequences we randomly deleted genes from the genomes. For the obtained draft genomes with different degrees of completeness, we re-ran the Traitar classification and computed the accuracy measures, as before. We observed that the average fraction of phenotypes identified (macro-recall for the positive class) of the phypat+PGL classifier dropped more quickly with more missing coding sequences than that of the phypat classifier (Figure 4.6 b). However, at the same time, the recall of the negative class of the phypat+PGL classifier improved with a decreasing number of coding sequences, meaning that fewer but more reliable predictions were made.

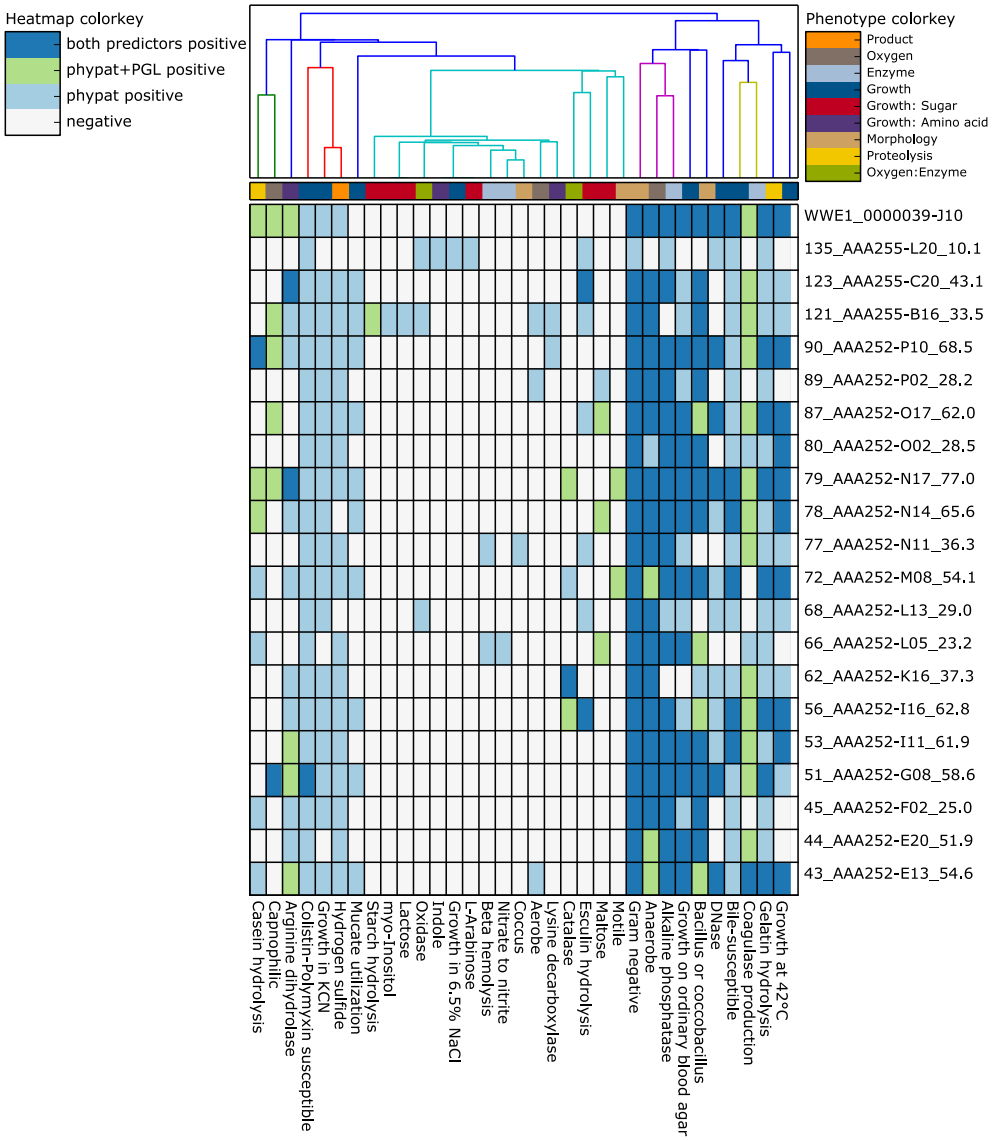


Figure 4.5: Single-cell phenotyping with Traitar. We used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with Traitar. Shown is a heatmap of assembly samples vs. phenotypes, which is the standard visualization for phenotype predictions in Traitar. The origin of the phenotype's prediction (Traitar phypat and/or Traitar phypat+PGL classifier) determines the color of the heatmap entries. The sample labels have their genome completeness estimates as suffixes. The colors of the dendrogram indicate similar phenotype distributions across samples, as determined by a hierarchical clustering with SciPy[1].

Overall, the tradeoffs in the recall of the phenotype-positive and the phenotype-negative classes of the two classifiers resulted in a similar overall macro-accuracy across the range of tested genome completeness.

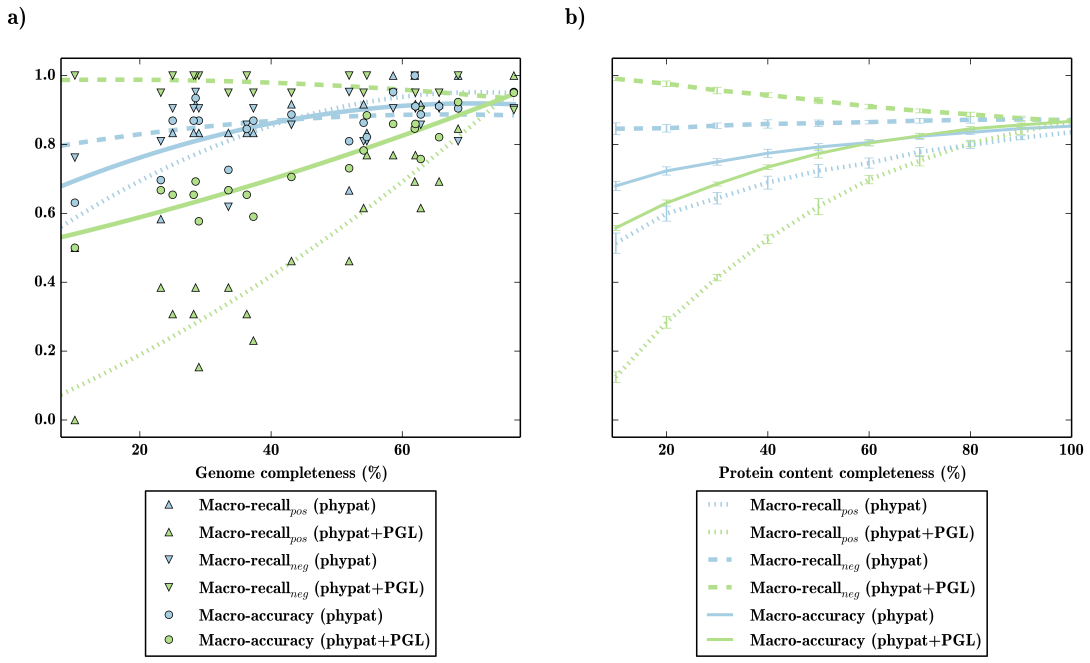


Figure 4.6: Phenotyping simulated draft genomes and single cell genomes. In (a) we used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. Shown is the performance of the phenotype prediction vs. the genome completeness of the single cells with respect to the joint assembly. In (b) we simulated draft genomes based on an independent test set of 42 microbial (pan)genomes. The coding sequences of these genomes were down-sampled (10 replications per sampling point) and the resulting simulated draft genomes were used for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. We plotted various performance estimates (mean center values and and s.d. error bars shown) against the protein content completeness.

Thus, depending on the intended usage, a particular classifier can be chosen: we expect that the reliable predictions inferred with the phypat+PGL classifier and the more abundant, but less reliable predictions made with the phypat classifier will complement one another in different use cases for partial genomes recovered from metagenomic data. By analyzing the protein families with assigned weights and the bias terms of the two classifiers, we found the phypat+PGL classifier to base its predictions primarily on the

presence of protein families that were typical for the phenotypes. In contrast, the phypat classifier also took typically absent protein families from phenotype-positive genomes into account in its decision. More technically, the positive weights in models of the phypat classifier are balanced out by negative weights, whereas for the phypat+PGL classifier, they are balanced out by the bias term. By down-weighting the bias term for the phypat+PGL classifier by the protein content completeness, we could show that the accuracy of the phypat classifier could be increased over that of the phypat+PGL, regardless of the protein content completeness (data not shown). However, this requires knowledge of the protein content completeness for each genomic sample, which could be indirectly estimated using methods such as checkM (Parks et al. 2015).

4.3.5 Traitar as a resource for gene target discovery

In addition to phenotype assignment, Traitar suggests the protein families relevant for the assignment of a phenotype (Methods – Majority feature selection, Table 4.3). We exemplarily demonstrate this capability here for three phenotypes that are already well-studied, namely 'Motile', 'Nitrate to nitrite' conversion and 'L-arabinose' metabolism. These phenotypes represent one each from the phenotype categories morphology, enzymatic activity and growth on sugar.

In general, we observed that the protein families important for classification can be seen to be gained and lost jointly with the respective phenotypes within the microbial phylogeny (Figure 4.7). Among the selected Pfam families that are important for classifying the motility phenotype were proteins of the flagellar apparatus and chemotaxis-related proteins (Table 4.3). Motility allows bacteria to colonize their preferred environmental niches. Genetically, it is mainly attributed to the flagellum, which is a molecular motor, and is closely related to chemotaxis, a process that lets bacteria sense chemicals in their surroundings. Motility also plays a role in bacterial pathogenicity, as it enables bacteria to establish and maintain an infection. For example, pathogens can use flagella to adhere to their host and they have been reported to be less virulent if they lack flagella (Josenhans et al. 2002). Of 48 flagellar proteins described in (R. Liu et al. 2007), four proteins (FliS, MotB, FlgD and FliJ) were sufficient for accurate classification of the motility phenotype and were selected by our classifier, as well as FlaE, which was not included in this collection. FliS (PF02561) is a known export chaperone that inhibits early polymerization of the flagellar filament FliC in the cytosol (Lam et al. 2010). MotB (PF13677), part of the membrane proton-channel complex, acts as

the stator of the bacterial flagellar motor (Hosking et al. 2006). Traitar also identified further protein families related to chemotaxis, such as CZB (PF13682), a family of chemoreceptor zinc-binding domains found in many bacterial signal transduction proteins involved in chemotaxis and motility (Draper et al. 2011), and the P2 response regulator-binding domain (PF07194). The latter is connected to the chemotaxis kinase CheA and is thought to enhance the phosphorylation signal of the signaling complex (Dutta et al. 1999).

Nitrogen reduction in nitrate to nitrite conversion is an important step of the nitrogen cycle and has a major impact on agriculture and public health. Two types of nitrate reductases are found in bacteria: the membrane-bound Nar and the periplasmic Nap nitrate reductase (Moreno-Vivian et al. 1999), which we found both to be relevant for the classification of the phenotype: we identified all subunits of the Nar complex as being relevant for the 'Nitrate to nitrite' conversion phenotype (i.e. the gamma and delta subunit (PF02665, PF02613)), as well as Fer4_11 (PF13247), which is in the iron–sulfur center of the beta subunit of Nar. The delta subunit is involved in the assembly of the Nar complex and is essential for its stability, but probably is not directly part of it (Pantel et al. 1998). Traitar also identified the Molybdopterin oxido-

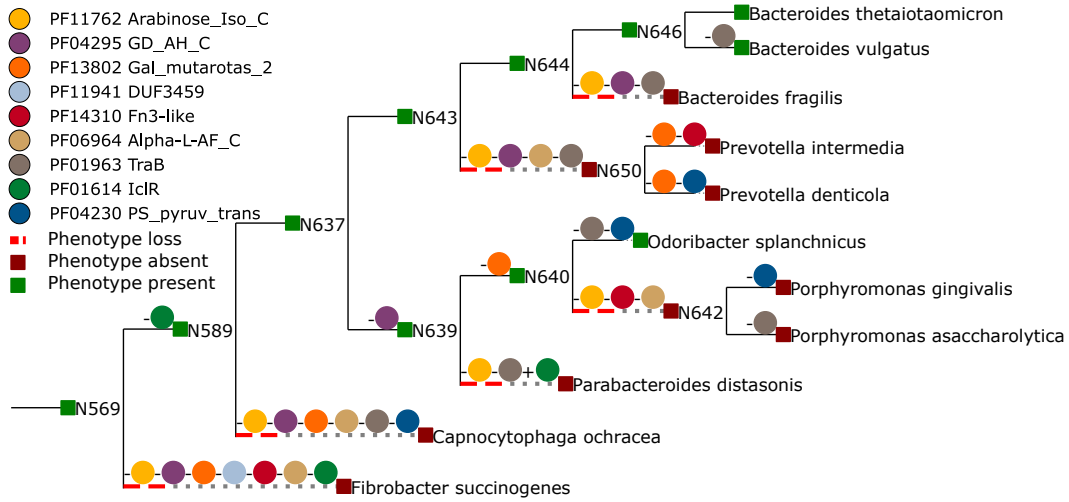


Figure 4.7: Phenotype gain and loss dynamics match protein family dynamics. We show the phenotype–protein family gain and loss dynamics for families identified as important by Traitar for the L-arabinose phenotype. Signed colored circles along the tree branches depict protein family gains (+) or losses (-). Taxon nodes are colored according to their inferred (ancestral) phenotype state.

reductase Fe4S4 domain (PF04879), which is bound to the alpha subunit of the nitrate

reductase complex (Pantel et al. 1998). Traitor furthermore suggested NapB (PF03892) as relevant, which is a subunit of the periplasmic Nap protein and NapD (PF03927), which is an uncharacterized protein implicated in forming Nap (Moreno-Vivian et al. 1999).

Table 4.3: The most relevant Pfam families for classification of three important phenotypes: 'Nitrate to Nitrite', 'Motility' and 'L-Arabinose'. We ranked the Pfam families with positive weights in the Traitar SVM classifiers by the correlation of the Pfam families with the respective phenotype labels across 234 bacteria described in the Global Infectious Disease and Epidemiology Online Network. Shown are the 10 highest ranking Pfam families along with their descriptions and a description of their phenotype-related function, where we found one.

Accession	Phenotype	Pfam description	Remarks
PF13677	Motile	Membrane MotB of proton-channel complex MotA/MotB	Flagellar protein
PF03963	Motile	Flagellar hook capping protein N-terminal region	Flagellar protein
PF02561	Motile	Flagellar protein FliS	Flagellar protein
PF02050	Motile	Flagellar FliJ protein	Flagellar protein
PF07559	Motile	Flagellar basal body protein FlaE	Flagellar protein
PF13682	Motile	Chemoreceptor zinc-binding domain	Chemotaxis-related
PF03350	Motile	Uncharacterized protein family, UPF0114	
PF05226	Motile	CHASE2 domain	Chemotaxis-related
PF07194	Motile	P2 response regulator binding domain	Chemotaxis-related
PF04982	Motile	HPP family	
PF03927	Nitrate to nitrite	NapD protein	Involved in Nar formation
PF13247	Nitrate to nitrite	4Fe-4S dicluster domain	Iron-sulfur cluster center of the beta subunit of Nar
PF03892	Nitrate to nitrite	Nitrate reductase cytochrome c-type subunit (NapB)	Periplasmic Nap subunit
PF02613	Nitrate to nitrite	Nitrate reductase delta subunit	Nap subunit
PF01127	Nitrate to nitrite	Succinate dehydrogenase/Fumarate reductase transmembrane subunit	
PF01292	Nitrate to nitrite	Prokaryotic cytochrome b561	
PF03459	Nitrate to nitrite	TOBE domain	
PF03824	Nitrate to nitrite	High-affinity nickel transport protein	
PF04879	Nitrate to nitrite	Molybdopterin oxidoreductase Fe4S4 domain	Bound to the alpha subunit of Nar
PF02665	Nitrate to nitrite	Nitrate reductase gamma subunit	Nar subunit
PF11762	L-Arabinose	L-arabinose isomerase C-terminal domain	Catalyzes first reaction in L-arabinose metabolism
PF04295	L-Arabinose	D-galactarate dehydratase / Altronate hydrolase, C terminus	
PF13802	L-Arabinose	Galactose mutarotase-like	
PF11941	L-Arabinose	Domain of unknown function (DUF3459)	
PF14310	L-Arabinose	Fibronectin type III-like domain	

PF06964	L-Arabinose	Alpha-L-arabinofuranosidase C-terminus	Acts on L-arabinose side chains in pectins
PF01963	L-Arabinose	TraB family	
PF01614	L-Arabinose	Bacterial transcriptional regulator	
PF06276	L-Arabinose	Ferric iron reductase FhuF-like transporter	
PF04230	L-Arabinose	Polysaccharide pyruvyl transferase	

L-arabinose is major constituent of plant polysaccharides, which is located, for instance, in pectin side chains and is an important microbial carbon source (D. Martinez et al. 2008). Traitar identified the L-arabinose isomerase *C*-terminal domain (PF11762), which catalyzes the first step in L-arabinose metabolism – the conversion of L-arabinose into L-ribulose (Sa-Nogueira et al. 1997), as being important for realizing the L-arabinose metabolism. It furthermore suggested the *C*-terminal domain of Alpha-L-arabinofuranosidase (PF06964), which cleaves nonreducing terminal alpha-L-arabinofuranosidic linkages in L-arabinose-containing polysaccharides (Gilead et al. 1995) and is also part of the well-studied L-arabinose operon in *Escherichia coli* (Sa-Nogueira et al. 1997).

4.3.6 Phenotyping biogas reactor population genomes

We used Traitar to phenotype two novel Clostridiales species (unClos_1, unFirm_1) based on their genomic information reconstructed from metagenome samples. These were taken from a commercial biogas reactor operating with municipal waste (Frank et al. 2015). The genomes of unClos_1 and unFirm_1 were estimated to be 91% complete and 60% complete based on contigs $\geq$ 5 kb, respectively. Traitar predicted unClos_1 to utilize a broader spectrum of carbohydrates than unFirm_1 (Table 4.4). We cross-referenced our predictions with a metabolic reconstruction conducted by Frank *et al.* (under review; supplementary material). We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. The manual reconstruction and predictions inferred with Traitar agreed to a great extent (Table 4.4). Traitar recalled 87.5% (6/7) of the phenotypes inferred via the metabolic reconstruction and also agreed to 81.8% (9/11) on the absent phenotypes. Notable exceptions were that Traitar only found a weak signal for 'D-xylose' utilization. A weak signal means that only a minority of the classifiers in the voting committee assigned these samples to the phenotype-positive class (Methods – Phenotype models). However, the metabolic reconstruction was also inconclusive with respect to xylose fermentation. Furthermore, Traitar only found a weak signal for 'Glucose fermentation' for unFirm_1.

Whilst genomic analysis of unFirm_1 revealed the Embden–Meyerhof–Parnas (EMP) pathway, which would suggest glucose fermentation, gene-centric and metaproteomic analysis of this phylotype indicated that the EMP pathway was probably employed in an anabolic direction (gluconeogenesis); therefore unFirm_1 is also unlikely to ferment D-Mannose. This suggests that unFirm_1 is unlikely to ferment sugars and instead metabolizes acetate (also predicted by Traitar, Table 4) via a syntrophic interaction with hydrogen-utilizing methanogens.
Traitar predicted further phenotypes for both species that were not targeted by the manual reconstruction. One of these predictions was an anaerobic lifestyle, which is likely to be accurate, as the genomes were isolated from an anaerobic bioreactor environment. It also predicted them to be Gram-positive, which is probably correct, as the Gram-positive sortase protein family can be found in both genomes.
This is a Gram-positive biomarker (Paterson et al. 2004). Furthermore, all Firmicutes known so far are Gram-positive (Goodfellow et al. 2012). Additionally, Traitar assigned 'Motile' and 'Spore formation' to unFirm_1, based on the presence of several flagellar proteins (e.g. FliM, MotB, FliS and FliJ) and the sporulation proteins CoatF and YunB.

4.4 Discussion

We have developed Traitar, a software framework for predicting phenotypes from the protein family profiles of bacterial genomes. Traitar provides a quick and fully automated way of assigning 67 different phenotypes to bacteria based on the protein family content of their genomes.
Microbial trait prediction from phyletic patterns has been proposed in previous studies for a limited number of phenotypes (Feldbauer et al. 2015; Kastenmuller et al. 2009; Konietzny, P. B. Pope, et al. 2014; Lingner et al. 2010; MacDonald et al. 2010; Weimann, Trukhina, et al. 2013). To our knowledge, the only currently available software for microbial genotype-phenotype inference is PICA, which is based on learning associations of clusters of orthologous genes (Tatusov et al. 2001) with traits (Feldbauer et al. 2015; MacDonald et al. 2010). Recently, PICA was extended by Feldbauer *et al.* for predicting eleven traits overall, optimized for large datasets and tested on incomplete genomes (Feldbauer et al. 2015; MacDonald et al. 2010). Traitar allows prediction of 67 phenotypes, including 60 entirely novel ones. It furthermore includes

Table 4.4: Phenotype predictions for two novel Clostridiales species with genomes reconstructed from a commercial biogas reactor metagenome. Traitar output (yes, no, weak) was cross-referenced with phenotypes manually reconstructed based on Kyoto Encyclopedia of Genes and Genomes orthology annotation (Frank *et al.* submitted; supplementary material), which are primarily the fermentation phenotypes of various sugars. We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. A weak prediction means that only a minority of the classifiers in the Traitar voting committee assigned this sample to the phenotype-positive class (Traitar phenotype). Table entries colored in red show a difference between the prediction and the reconstruction, whereas green denotes an overlap; yellow is inconclusive.

	unClos_1	unFirm_1
Glucose	yes	weak
Acetate utilization	no	yes
Mannitol	yes	no
Starch hydrolysis	no	no
Xylose	weak	no
L-Arabinose	yes	no
Capnophilic	yes	no
Sucrose	yes	no
D-Mannose	yes	no
Maltose	yes	no
Arginine dihydrolase	no	yes

different prediction modes, one based on phyletic patterns, one additionally including a statistical model of protein family evolution for its predictions. Traitar also suggest associations between phenotypes and protein families. For three traits, we showed that several of these associations are to known key families of establishment of a particular trait, and that furthermore candidate families were suggested, that might serve as targets for experimental studies. Some of the phenotypes annotated in GIDEON are specific for the human habitat (such as 'coagulase production' or 'growth on ordinary blood agar') and the genetic underpinnings learned by Traitar could be interesting to study for infection disease research.

In cross-validation experiments with phenotype data from the GIDEON database, we showed that the Traitar phypat classifier has high accuracy in phenotyping bacterial

samples. Considering ancestral protein family gains and losses in the classification, which is implemented in the Traitar phypat+PGL classifier, improves the accuracy compared to prediction from phyletic patterns only, both for individual phenotypes and overall. Barker *et al.* were first to note the phylogenetic dependence of genomic samples and how this can lead to biased conclusions (Barker et al. 2005). MacDonald *et al.* selected protein families based on correlations with a phenotype and corrected for the taxonomy (MacDonald et al. 2010). Here we accounted for the evolutionary history of the phenotype and the protein families in the classifier training itself to automatically improve phenotype assignment. We additionally demonstrated the reliability of the performance estimates by phenotyping, with a similar accuracy, an independent test dataset with bacteria described in GIDEON, which we did not use in the cross-validation. Traitar also reliably phenotyped a large and heterogenic collection of bacteria that we extracted from Bergey's Manual of Systematic Bacteriology – mostly with only draft genomes available. We didn't observe any bias towards specific taxa in GIDEON, but some of the phenotypes might be realized with different protein families in taxa that are less well represented indicated by the around 15% - 20% less reliable phenotyping results for bacteria described in Bergey's manual of systematic bacteriology. We expect that the accuracy of the phenotype classification models already available in Traitar will further improve the more data will become available and can be incorporated into its training.

We found that Traitar can provide reliable insights into the metabolic capabilities of microbial community members even from partial genomes, which are very common for genomes recovered from single cells or metagenomes. One obvious limitation being for incomplete genomes, the absence of a phenotype prediction may be due to the absence of the relevant protein families from the input genomes. The analysis of both the SAGs and simulated genomes led us to the same conclusions: the phypat classifier is more suitable for exploratory analysis, as it assigned more phenotypes to incomplete genomes, at the price of more false positive predictions. In contrast, the phypat+PGL classifier assigned fewer phenotypes, but also made fewer false assignments. At the moment, genotype–phenotype inference with Traitar only takes into account the presence and absence of protein families of the bacteria analyzed. This information can be readily computed from the genomic and metagenomic data. Future research could focus also on integration of other 'omics' data to allow even more accurate phenotype assignments. Additionally, expert knowledge of the biochemical pathways that are used in manual metabolic reconstructions, for example, could be integrated as prior knowledge into the

model in future studies.

For the phenotyping of novel microbial species, generating a detailed (manual) metabolic reconstruction such as the one by Frank *et al.* (submitted; supplementary material) is time-intensive. Furthermore, such reconstructions are usually focused on specific pathways and are dependent on the research question. This is not an option for studies with 10–50+ genomes, which are becoming more and more common in microbiology (Brown et al. 2015; Hess et al. 2011; MacDonald et al. 2010; Rinke et al. 2013). Traitar thus is likely to be particularly helpful for multi-genome studies. It furthermore may pick up on things outside of the original research focus and could serve as a seed or a first-pass method for a detailed metabolic reconstruction in future studies.

4.5 Methods

4.5.1 The Traitar software

In this section we first describe the Traitar annotation procedure. We proceed with the genome and phenotype data used for the training of Traitar phenotype models; afterwards we explain the training and illustrate how we considered ancestral protein family gains and losses in the models. Finally, we specify the requirements for running the Traitar software.

Annotation

In the case of nucleotide DNA sequence input, Traitar uses Prodigal (Hyatt et al. 2010) for gene prediction prior to Pfam family annotation. The amino acid sequences are then annotated in Traitar with protein families (Pfams) from the Pfam database (version 27.0) (Finn, Bateman, et al. 2014) using the hmmsearch command of HMMER 3.0 (Finn, Clements, et al. 2011).

Each Pfam family has a hand-curated threshold for the bit score, which is set in such a way that no false positive is included (Punta et al. 2012). A fixed threshold of 25 is then applied to the bit score (the log-odds score) and all Pfam domain hits with an E-value above 10^{-2} are discarded. The resulting Pfam family counts (phyletic patterns) are turned into presence or absence values, as we found this representation to yield a favorable classification performance (Weimann, Trukhina, et al. 2013).

Genome and phenotype data

We obtained our phenotype data from the GIDEON database (Berger 2005). In GIDEON a bacterium is labeled either as phenotype-positive, -negative or strain-specific. In the latter case we discarded this phenotype label. The GIDEON traits can be grouped into the categories the use of various substrates as source of carbon and energy for growth, oxygen requirement, morphology, antibiotic susceptibility and enzymatic activity (Table 4.1, Supplementary Table 1). We only considered phenotypes that were available in GIDEON for at least 20 bacteria, with a minimum of 10 bacteria annotated as positive (phenotype presence) for a given phenotype and 10 as negative (phenotype absence) to enable a robust and reliable analysis of the respective phenotypes. Furthermore, to be included in the analysis, we required each bacterial sample to have:

a) at least one annotated phenotype,

b) at least one sequenced strain,

c) a representative in the sTOL.

In total, we extracted 234 species-level bacterial samples with 67 phenotypes with sufficient total, positive and negative labels from GIDEON (GIDEON I). GIDEON associates these bacteria with 9305 individual phenotype labels, 2971 being positive and 6334 negative (Supplementary Table 1, 3). GIDEON species that had at least one sequenced strain available but were not part of the sTOL tree were set aside for a later independent assessment of the classification accuracy. In total, this additional dataset comprised further 42 unique species with 58 corresponding sequenced bacterial strains (GIDEON II, Supplementary Table 1, 4). We obtained 1836 additional phenotype labels for these bacteria, consisting of 574 positive and 1262 negative ones. We searched the Firmicutes and Proteobacteria volumes of Bergey's systematic bacteriology specifically for further bacteria not represented so far in the GIDEON data sets (Goodfellow et al. 2012). In total, we obtained phenotype data from Bergey's for 206 Firmicutes and 90 Proteobacteria with a total of 1152 positive labels and 1376 negative labels (Supplementary Table 1, 5). As in GIDEON, in Bergey's the phenotype information is usually given on the species level.

We downloaded the coding sequences of all complete bacterial genomes that were available via the NCBI FTP server under `ftp://ftp.ncbi.nlm.nih.gov/genomes/` as of

11 May 2014 and genomes from the PATRIC data base as of September 2015 (Wattam et al., 2014). These were annotated with Traitar. For bacteria with more than one sequenced strain available, we chose the union of the Pfam family annotation of the single genomes to represent the pangenome Pfam family annotation, as in (Y. Liu et al. 2006).

Phenotype models

We represented each phenotype from the set of GIDEON phenotypes across all genomes as a vector $\boldsymbol{yp}$, and solved a binary classification problem using the matrix of Pfam phyletic patterns XP across all genomes as input features and $\boldsymbol{yp}$ as the binary target variable (Supplementary Figure 1). For classification, we relied on support vector machines (SVMs), which are a well-established machine learning method (Boser et al. 1992). Specifically, we used a linear L1-regularized L2-loss SVM for classification as implemented in the LIBLINEAR library (Fan et al. 2008). For many datasets, linear SVMs achieve comparable accuracy to SVMs with a non-linear kernel but allow faster training. The weight vector of the separating hyperplane provides a direct link to the Pfam families that are relevant for the classification. L1-regularization enables feature selection, which is useful when applied to highly correlated and high-dimensional datasets, as used in this study (Zou et al. 2005). We used the interface to LIBLINEAR implemented in scikit-learn (Pedregosa et al. 2011). For classification of unseen data points – genomes without available phenotype labels supplied by the user – Traitar uses a voting committee of five SVMs with the best single cross-validation accuracy (Methods – Nested cross-validation). Traitar then assigns each unseen data point to the majority class (phenotype presence or absence class) of the voting committee.

Ancestral protein family and phenotype gains and losses

We constructed an extended classification problem by including ancestral protein family gains and losses, as well as the ancestral phenotype gains and losses in our analysis, as implemented in GLOOME (Cohen and Pupko 2011). Barker *et al.* report that common methods for inferring functional links between genes, that do not take the phylogeny into account, suffer from high rates of false positives (Barker et al. 2005). Here, we jointly derived the classification models from the observable phyletic patterns and phenotype labels, and from phylogenetically unbiased ancestral protein family and phenotype gains and losses, that we inferred via a maximum likelihood approach from the observable

phyletic patterns on a phylogenetic tree, showing the relationships among the samples. (Supplementary Figure 1). Ancestral character state evolution in GLOOME is modeled via a continuous-time Markov process with exponential waiting times. The gain and loss rates are sampled from two independent gamma distributions (Cohen and Pupko, 2010).

GLOOME needs a binary phylogenetic tree with branch lengths as input. The taxonomy of the National Center for Biointechnology Information (NCBI) and other taxonomies are not suitable, because they provide no branch length information. We used the sequenced tree of life (sTOL) (Fang et al. 2013), which is bifurcating and was inferred with a maximum likelihood approach based on unbiased sampling of structural protein domains from whole genomes of all sequenced organisms (Gough et al., 2001). We employed GLOOME with standard settings to infer posterior probabilities for the phenotype and Pfam family gains and losses from the Pfam phyletic patterns of all NCBI bacteria represented in the sTOL and the GIDEON phenotypes. Each GIDEON phenotype p is available for a varying number of bacteria. Therefore, for each phenotype, we pruned the sTOL to those bacteria that were both present in the NCBI database and had a label for the respective phenotype in GIDEON. The posterior probabilities of ancestral Pfam gains and losses were then mapped onto this GIDEON phenotype-specific tree (Gps-sTOL, Supplementary Figure 2).

Let B be the set of all branches in the sTOL and P be the set of all Pfam families. We then denote the posterior probability g_{ij} of an event a for a Pfam family pf to be a gain event on branch b in the sTOL computed with GLOOME as:

$$g_{\mathrm{ij}} = P\left(a = gain | i = b, j = pf\right) \ \forall\ i\ \in B,\ \forall\ j\ \in P,$$

and the posterior probability of a to be a loss event for a Pfam family p on branch b as:

$$l_{\mathrm{ij}} = P\left(a = loss | i = b, j = pf\right) \ \forall\ i\ \in B,\ \forall\ j\ \in\ P.$$

We established a mapping $f : B' \rightarrow B$ between the branches of the sTOL B and the set of branches B' of the Gps-sTOL (Supplementary Figure 2). This was achieved by traversing the tree from the leaves to the root.

There are two different scenarios for a branch b' in B' to map to the branches in B:

a) Branch b' in the Gps-sTOL derives from a single branch b in the sTOL: $f\left(b'\right) = \{b\}$. The posterior probability of a Pfam gain inferred in the Gps-sTOL on branch

b' consequently is the same as that on branch b in the sTOL
$g_{b'j} = g_{\mathrm{bj}} \forall\ j \epsilon P$.

b) Branch b' in the Gps-sTOL derives from m branches $b_1, \ldots, b_m$ in the sTOL: $f(b') = \{b_1, \ldots, b_m\}$ (Supplementary figure 2). In this case, we iteratively calculated the posterior probabilities for at least one Pfam gain g' on branch b' from the posterior probabilities for a gain $g'_{\ b_1 j}$ from the posterior probabilities $g_1, \ldots, g_m$of a gain on branches $b_1, \ldots, b_m$ with the help of h:

$$\begin{aligned} h_1 &= g_{\ b_1 j} \\ h_{n+1} &= (1 - h_n) \cdot g_{\ b_{n+1} j} \\ g'_{\ b_1 j} &= h_m \ \forall j \in P. \end{aligned}$$

Inferring the Gps-sTOL Pfam posterior loss probabilities l'_{ij} from the sTOL posterior Pfam loss probabilities is analogous to deriving the gain probabilities. The posterior probability for a phenotype p to be gained g'_{ip} or lost l'_{ip} can be directly defined for the Gps-sTOL in the same way as for the Pfam probabilities.
For classification, we did not distinguish between phenotype or Pfam gains or losses, assuming that the same set of protein families gained with a phenotype will also be lost with the phenotype. This assumption simplified the classification problem. Specifically, we proceeded in the following way:
1. We computed the joint probability x_{ij} of a Pfam family gain or loss on branch b' and the joint probability y_j of a phenotype gain or loss on branch b':

$$\begin{aligned} x_{\mathrm{ij}} &= g'_{\mathrm{ij}} l'_{\mathrm{ij}} + \left(1 - g'_{\mathrm{ij}}\right) \cdot l'_{\mathrm{ij}} + \left(1 - l'_{\mathrm{ij}}\right) \cdot g'_{\mathrm{ij}} \ \forall\ i\ \in B',\ \forall\ j \in\ P \\ &= g'_{\mathrm{ij}} + \left(1 - g'_{\mathrm{ij}}\right) \cdot l'_{\mathrm{ij}} \end{aligned}$$

$$y_i = g'_{\mathrm{ip}} + \left(1 - g'_{\mathrm{ip}}\right) \cdot\ l'_{\mathrm{ip}} \quad \forall\ i\ \in B'.$$

2. Let $\mathbf{x_i}$ be a vector representing the probabilities x_{ij} for all Pfam families $j \in\ P$ on branch b_i. We discarded any samples ($\mathbf{x}_i$,y_i) that had a probability for a phenotype gain or loss y_i above the reporting threshold of GLOOME but below a threshold t. We set the threshold t to 0.5.
This defines the matrix X and the vector $\mathbf{y}$ as:

$$(X,\ \mathbf{y}) = \{(\mathbf{x_{i,}}\ y_{i,}\ )\ | y_i = 0 \vee y_i \geq t,\ i\ \in B'\}\ .$$

By this means, we avoided presenting the classifier with samples corresponding to uncertain phenotype gain or loss events and used only confident labels in the subsequent classifier training instead.
3. We inferred discrete phenotype labels $\mathbf{y}^{'}$ by applying this threshold t to the joint probability y_i for a phenotype gain or loss to set up a well-defined classification problem with a binary target variable. Whenever the probability for a phenotype to be gained or lost on a specific branch was larger than t, the event was considered to have happened:

$$\mathbf{y}' = \begin{cases} 1, \text{ if } y_i \geq t \\ 0, \qquad \text{otherwise} \end{cases} \quad \forall i \ \in B'.$$

4. Finally, we formulated a joint binary classification problem for each target phenotype yp and the corresponding gain and loss events $\mathbf{y}^{'}$, the phyletic patterns XP, and the Pfam gain and loss events X, which we solved again with a linear L1-regularized L2-loss SVM. We applied this procedure for all GIDEON phenotypes under investigation.

Software Requirements

Traitar can be run on a standard laptop with Linux/Unix. The runtime (wallclock time) for annotating and phenotyping a typical microbial genome with 3 Mbp is 9 minutes (3 min/Mbp) on an Intel(R) Core(TM) i5-2410M dual core processor with 2.30 GHz, requiring only a few megabytes of memory.

4.5.2 Cross-validation

We employed cross-validation to assess the performance of the classifiers individually for each phenotype. For a given phenotype, we divided the bacterial samples that were annotated with that phenotype into ten folds. Each fold was selected once for testing the model, which was trained on the remaining folds. The optimal regularization parameter C needed to be determined independently in each step of the cross-validation; therefore, we employed a further inner cross-validation using the following range of values for the parameter C: 10^{-3}, $10^{-2}\cdot 0.7$, $10^{-2}\cdot 0.5$, $10^{-2}\cdot 0.2$, $10^{-2}\cdot 0.1$, ..., 1. In other words, for each fold kept out for testing in the outer cross-validation, we determined the value of the parameter C that gave the best accuracy in an additional tenfold cross-validation on the remaining folds. This value was then used to train the SVM model in the current outer cross-validation step. Whenever we proceeded to a new cross-validation fold, we re-computed the ancestral character state reconstruction of the phenotype with only the

training samples included (Ancestral protein family and phenotype gains and losses). This procedure is known as nested cross-validation (Ruschhaupt et al. 2004).
The bacterial samples in the training folds imply a Gps-sTOL in each step of the inner and outer cross-validation without the samples in the test fold. We used the same procedure as before to map the Pfam gains and losses inferred previously on the Gps-sTOL onto the tree defined by the current cross-validation training folds. Importantly, the test error is only estimated on the observed phenotype labels rather than on the inferred phenotype gains and losses.

4.5.3 Evaluation metrics

We used evaluation metrics from multi-label classification theory for performance evaluation (Manning et al. 2008). We determined the performance for the individual phenotype-positive and the phenotype-negative classes based on the confusion matrix of true positive (TP), true negative (TN), false negative (FN) and false positive (FP) samples from their binary classification equivalents by averaging over all n phenotypes. We utilized two different accuracy measures for assessing multi-class classification performance (i.e. the accuracy pooled over all classification decisions and the macro-accuracy). Macro-accuracy represents an average over the accuracy of the individual binary classification problems and we computed this from the macro-recall of the phenotype-positive and the phenotype-negative classes as follows:

$$\text{Macro}-\text{recall}_{\text{Pos}} = \frac{\left(\sum_{i=1}^{n} \frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}\right)}{n}$$

$$\text{Macro}-\text{recall}_{\text{Neg}} = \frac{\left(\sum_{i=1}^{n} \frac{\text{TN}_i}{\text{FP}_i+\text{TN}_i}\right)}{n}$$

$$\text{Macro}-\textit{accuracy} = \frac{(\text{Macro}-\text{recall}_{\text{Pos}} + \text{Macro}-\text{recall}_{\text{Neg}})}{2}.$$

However, if there are only few available labels for some phenotypes, the variance of the macro-accuracy will be high and this measure cannot be reliably computed anymore; it cannot be computed at all if no labels are available. The accuracy only assesses the overall classification performance without consideration of the information about specific phenotypes. Large classes dominate small classes (Manning et al. 2008)

$$\text{Recall}_{\text{Pos}} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i}$$

$$\text{Recall}_{\text{Neg}} = \frac{\sum_{i=1}^{n} \text{TN}_i}{\sum_{i=1}^{n} \text{TN}_i + \sum_{i=1}^{n} \text{FP}_i}$$

$$\textit{Accuracy} = \frac{(\text{Recall}_{\text{Pos}} + \text{Recall}_{\text{Neg}})}{2}.$$

4.5.4 Majority feature selection

The weights in linear SVMs can directly be linked to features that are relevant for the classification. We identified the most important protein families used as features from the voting committee of SVMs consisting of the five most accurate models, which were also used for classifying new samples. If the majority, which is at least three predictors, included a positive value for a given protein family, we added this feature to the list of important features. We further ranked these protein families features by their correlation with the phenotype using Pearson's correlation coefficient.

4.5.5 Acknowledgements

We thank Andreas Klötgen, David Lähnemann, Susanne Reimering and Alexander Sczyrba for providing helpful comments on the manuscript; Johannes Dröge and Jens Loers for reviewing the Traitar software and Gary Robertson for helping to set up the Traitar web service. JAF and PBP are supported by a grant from the European Research Council (336355-MicroDE).

4.6 Supplementary material

The supplementary material can be found in the original version of the paper in the appendix and online at `http://dx.doi.org/10.1101/043315`.

CHAPTER 5

Antibiotic resistance prediction from transcriptional and mutational profiles of *Pseudomonas aeruginosa*

Status **in preparation**

Joint work with Monika Schniederjans, Ariane Khaledi, Alice C. McHardy, Susanne Häußler

Wrote the manuscript

Conceived and designed the analyses (with MS, ACM, SH)

Implemented and conducted the analyses

Interpreted the classification results, found biomarkers, etc. (with MS)

5.1 Introduction

Multi-drug resistant bacterial pathogens increasingly threaten public health. They cause many deaths and impose a huge economic burden on the health care sector (ECDC 2009). At the same time the pharmaceutical industry approves less and less novel antibiotics and those brought to the market often involve only minor modification to the original drugs. *Pseudomonas aeruginosa* is a particularly prevalent pathogen responsible for 8% of all hospital acquired infections (Sievert et al. 2013). It is found in diverse natural and human habitats and is particularly dangerous for patients suffering from cystic fibrosis, burn wounds or neutropenia. The pathogen grows persistent biofilms for example in the lung of cystic fibrosis patients, which makes it even harder to eradicate (Costerton 2001).

The large genome and an even larger assessory genome of *P. aeruginosa* provide the genetic basis of this adaptability (Stover et al. 2000). The pathogenicity of *P. aeruginosa* is due to various virulence factors such as toxins, proteases and hemolysins. Many secretions systems and degradative enzymes confer a high natural resistance to many antibiotics (Rumbaugh 2014). Furthermore, the rates of resistance of *P. aeruginosa* to various antibiotic is increasing at an alarming rate (Hirsch et al. 2010).

Antibiotics used in the treatment of *P. aeruginosa* infections include quinolones, β-lactams and aminoglycosides (Meletis et al. 2013). The pathogen realizes different resistance mechanisms against β-lactam antibiotics such as by active efflux, altered outer-membrance permeability or β-lactamase activity including by acquired secondary β-lactamases. Carbapenem-resistant *P. aeruginosa* already accounted for 18% of all reported cases (ECDC 2009). Quinolone resistance is mainly mediated by target mutations, whereas aminoglycosides like gentamicin are mainly inactivated by degradative enzymes (Pechère et al. 1999). (Magiorakos et al. 2012). The polymyxin colistin was already discovered more than 60 years ago. Due to its nephrotoxic effects it was soon after replaced with less toxic drugs. Recently, it has been re-evaluated for treatment of multi-drug resistant *P. aeruginosa* infections, as it often represents the only promising option to combat the pathogen (Michalopoulos et al. 2005).

Quick and cheap genome sequencing of bacterial genomes is about to replace many expensive phenotypic microbiological assays in clinical diagnostics, for instance for the identification of bacterial pathogens. Genome sequencing is also being evaluated for determining phenotypic properties such as virulence or antibiotic resistance. (Didelot et al. 2012). Lieberman *et al.* monitored the evolution of *Burkholderia dolosa*, a highly

drug resistant pathogen found in cystic fibrosis infections, over an extended time period. By genome sequencing they observed parallel adaptive evolution and identified known and novel genes implicated in pathogenesis. (Lieberman et al. 2011). Bradley *et al.* developed Mykrobe predictor, a highly accurate antibiotic resistance profiling tool – comparable to gold-standard phenotypic assays – for *Staphylococcus aureus* and *Mycobacterium tuberculosis*, which assesses known genetic resistance markers (Bradley et al. 2015). To quantify the phenotypic effect of mutations, for example for antibiotic resistance genes, one can also take the gene expression level into account. However, transcriptome sequencing has so far been largely neglected (Palmer et al. 2013). Khaledia *et al.* identified markers of antibiotic resistance in *P.aeruginosa* from the transcriptional and mutational profiles, as determined by RNAseq of 135 clinical *P. aeruginosa* isolates with transcriptome-wide association. Furthermore they explored the potential of machine learning methods to determine global patterns of antibiotic resistance to ciprofloxacin(Khaledi et al. 2016). In this study we also employed machine learning methods to find global sequence variation and gene expression patterns that provide an accurate prediction of resistance to antibiotics of four commonly used drug classes - fluoroquinolones (ciprofloxacin), β-lactams (meropenem and ceftazidim), aminoglycosides (tobramycin) and polymyxins (colistin).

5.2 Results

Antibiotic resistance statistics Determination of the minimum inhibitory concentrations (MIC) of 467 *P. aeruginosa* isolates indicated that over 74% of the isolates were resistant against meropenem. More than 58% of the strains exhibited resistance to ciprofloxacin and ceftazidim. Only tobramycin resistance was less widely distributed, with only 33% resistant isolates, as well as colistin resistance, which was observed for only 5% of the samples. We discarded intermediate resistant isolates in subsequent analysis steps (Table 5.1). Multi-drug resistant isolates represent a large fraction of our data set (Figure 5.1). The resistance profiles of 79 isolates indicated that only the last-resort drug colistin represented a promising treatment option. Alarmingly, five strains out of a total of 20 colistin resistant strains showed resistance to all five antibiotics tested.

Antibiotic resistance prediction with logistic regression Large-scale genomic studies probe thousands or hundreds of thousands of genetic loci in parallel across

Table 5.1: Antibiotic resistance frequency across 467 *P. aeruginosa* clinical isolates as determined by the minimal inhibitory concentration (MIC) [µg/ml]. We determined antibiotic resistance to five antibiotics as defined by a MIC threshold according to the Clinical & Laboratory Standards Institute guidelines. For each antibiotic, there is a MIC that is regarded as intermediate resistant. We also report the number of intermediate resistant samples and the MIC thresholds used.

#isolates	Ciprofloxacin	Tobramycin	Colistin	Ceftazidim	Meropenem
susceptible	163	301	369	162	111
resistant	258	152	20	229	319
intermediate resistant	46	14	78	76	37
MIC [µg/ml] resistant $\geq$	4	8	8	32	16
MIC [µg/ml] susceptible $\leq$	1	4	2	8	4

a small number of samples. Genetic association studies that are typically employed to link genotype and phenotype are limited to finding single markers (P. E. Chen et al. 2015). Machine learning methods can identify groups of co-occuring genetic markers. Here, we employed a logistic regression classifier that scales to large-scale data sets and provides regularization to identify the most relevant genotype-phenotype associations. We trained a classifier to predict resistance or susceptibility to five antibiotics on the transcriptional as well as on single nucleotide polymorphism (SNP) profiles both obtained from RNAseq data. For evaluation of the classifiers, we used the macro accuracy (MACC) and the area under the receiver operating curve (AUC), which we estimated in a five-fold nested cross-validation. The MACC provides a single trade-off between true positive rate and false negative rate, whereas the AUC integrates all possible trade-offs (Material & Methods). Note that due to the limited scope of the project the focus is on presenting the classification performance, and we only provide selected insights into the genetic markers found.

Overall, we obtained very reliable classification results (Figure 5.2). For ciprofloxacin resistance prediction based on SNP profiles, we achieved a MACC of 87.8% and an AUC of 93.5%. The gyrA T83I mutation represented the most important marker, which is well-known to be associated with ciprofloxacin resistance (Yonezawa et al. 1995). Remarkably, we could obtain accurate prediction (MACC 76.1% / AUC 85.6%) using classifier trained on the expression profiles as well, although no expression markers have been associated with ciprofloxacin resistance, which is in line with the results obtained by Khaledi *et al.* (Khaledi et al. 2016).

Despite the fact that tobramycin resistance is mainly mediated by antibiotic resistance genes acquired by horizontal gene transfer (Poole 2005), we obtained highly accurate classification results for this drug as well, both using expression as well as SNP profiles for classifier training. Among the expression markers relevant for tobraymcin resistance prediction, we found *amrB*, which is the RND multi-drug efflux transporter (Westbrock-Wadman et al. 1999), as well as many uncharacterized genes. Interestingly, we found several SNPs in the *amrB* gene to contribute to the prediction as well. However, it is possible that these are an artifact of the SNP inference based on the RNAseq data, as SNPs in genes highly expressed in the tobramycin resistant isolates such as *amrB* might have remained undetected in the susceptible samples, due to low expression counts.

We further report a reliable prediction of meropenem (MACC 81.4%, AUC 92.2%) and ceftazidim (MACC 76.7%, AUC 83.3%), but solely based on gene expression profiles. We found that the over expression of *mexB* and *oprM*, which both encode parts of the well-characterized MexAB-OprM efflux system, are among the most important markers for meropenem resistance. Underexpression of *oprD*, which encodes a porin that confers passage of meropenem into the bacterial cell is also among the most important epxression markers (Meletis et al. 2013). We identified *ampC* as the most important biomarker for resistance to ceftazidim, which encodes a β-lactamase with known degradative activity against ceftazidim (Meletis et al. 2013).

Last, we learned a classifier for predicting colistin resistance. We obtained moderately accurate classification results when using expression (76.7% MACC) or SNPs (66.3% MACC). Interestingly, we observed a rather unreliable prediction of the colistin resistance class vs. a very reliable assignment to the susceptible class. Additionally, we obtained a high AUC of 87.3% based on the SNP profiles and 95.1% based on the expression data, which is the highest value obtained for any of the resistance classifiers. This indicates that we can set-up a highly reliable diagnostic test using an approriate trade-off between the true positive and false negative rate. The overexpression of *pmrB*, part of the two-component signal transduction system PmrAB and known to modulate resistance to colistin, as well as overexpression of genes in close proximity to *pmrB*, were found to be most relevant for the classification (Moskowitz et al. 2004). Interesting, we also found SNPs within this operon contributing to prediction of colistin resistance based on the mutational profiles. As for tobramycin resistance this might be due to low expression counts of these genes in the colistin susceptible group.

Additionally, we found several genes to be relevant for the prediction of antibiotic resistance with no known association with resistance. For instance *gbuA* seems to play

an important role in meropenem resistance and is currently being characterized in the lab of our collaborator Susanne Häußler at the Helmholtz Centre for Infection Research.

5.3 Discussion

The global increase of resistance to various antibiotics poses a serious thread to health care institutions and the economy. There is an imminent need for accurate determination of drug resistances for clinical isolates, to advise the physician to administer the most effective drug and minimize the spread of resistant strains. Currently, antibiotic susceptibility and resistance screening used in clinical practise requires to isolate and culture a bacterial pathogen, which can take several days. Furthermore, a recent study conducted in Spanish hospitals indicated that the determination of antibiotic resistance in clinical practise exhibited low accuracy and high variability across hospitals (Juan et al. 2013). However, determination of molecular markers of antibiotic resistance could lead to cheaper and quicker diagnostics. We have demonstrated in a large-scale study that the transcriptional and mutational profiles of *P. aeruginosa* allow an accurate prediction of antibiotic resistance.

The recent advances in microbial genomics could improve or even replace standard diagnostic microbiology (Reuter et al. 2013). Khaledi *et al.* showed the potential of machine learning methods for predicting antibiotic resistance to ciprofloxacin based on 135 isolates. Here, we show that a machine learning apporach is suitable for a wider range of antibiotics using a much larger data set of 467 samples. The logistic regression classifiers employed detected many genes with known links to antibiotic resistance, but follow-up studies are required to investigate the role of uncharacterized genes, which could provide insights into the antibiotic resistance mechanism.

It is crucial to avoid recommending an antibiotic for treatment, although the pathogen is resistant to the prescribed drug. It is less worse to miss a treatment option and instead recommend to adminster a alternative antibiotic for treatment. Thus, in future research, it will be important to optimize the classifier for prediction of the susceptibility class. Integrating data'omics such as the transcriptional and mutational profile may further improve prediction quality. Additionally, whole genome sequencing could aid in a more reliable unbiased detection of mutational sites implicated in resistance, since it does not depend on the level of expression of certain genes.

Importantly, genome-wide transcriptome and complete genome sequencing seem suitable to detect biomarkers of anitbiotic resistance or susceptibility, but also time-

consuming. However, quicker and cheaper targeted approaches such as PCR-based methods to detect the presence of individual genes, or MassARRY to screen for individual SNPs may be employed to speed up diagnostics, once the biomarkers are identified (Gabriel et al. 2009).
Finally, the framework we have developed is not limited to antibiotic resistance but is applicable to reveal other clinically relevant genotype-phenotype assoications such as pathogenicity. Our results suggest that antibiotic resistance profiling could soon identify molecular markers that accurately distinguish resistant and non-resistant isolates thus providing a cheaper and quicker approach compared to current practise in clinical resistance diagnostics.

5.4 Materials and Methods

Strain collection and antibiotic resistance profiles The clinical isolates included in this study were provided by different hospitals or institutions across Germany and other European countries and sampled from diverse infection sites. The antibiotic resistance profiles for five antibiotics were determined by agar dilution in at least two duplicates. Briefly, cultures were grown in 96-well plates for 4 h at 37 °C in an orbital shaker and adjusted to an $OD600 = 0.08 - 0.1(1 - 2 \cdot 10^8$ cells/ml). After serial dilutions to $2 - 4 \cdot 10^6$ cells/ml, finally 5 μl of the adjusted bacterial suspensions were spotted on cation-adjusted Müller-Hinton agar plates containing different concentrations of the antibiotic. After incubation at 37 °C overnight, the minimal inhibitory concentrations (MICs) were determined. The classification of antibiotic resistance and susceptibility was done according to CLSI (Clinical and Laboratory Standards Institute) guidelines.

RNA sequencing The clinical isolates were cultured in lysogeny broth until $OD600 = 2$ at 37 °C in an orbital shaker before cells were harvested with RNAprotect Bacteria Reagent (Qiagen). The extraction of RNA, sequencing library construction, cDNA sequencing and data analysis was performed according to Krüger *et al.* (Krueger et al. 2016). For logistic regression classification based on gene expression, the logarithmized read counts per gene obtained by mapping to the *Pseudomonas aeruginosa* UCBPP-PA14 reference genome were used. In addition, the read counts were standardized, by removing the mean of each feature and dividing by the standard deviation. The SNP calling was performed using SAMtools (Li et al. 2009). SNPs in coding regions that were covered by at least three reads and had a score of at least 50 were used as binary genotype information, with 1 marking the presence and 0 the

absence of a SNP in an isolate. In cases where the read coverage was not sufficient, an NA was displayed and set to absent for classification.

Logistic regression classification A binary classification was learned for each antibiotic. Antibiotic sensitive and susceptible samples were assigned to two separate target classes and encoded in a binary target variable. Logistic regression classifiers were trained independently on the logarithmized transcript counts and the binary SNP profiles using the LIBLINEAR library (Fan et al. 2008). Specifically we used a L1-regularized L2-penalized logistic regression classifier, which enables feature selection via regularization. We identified the most important genes from an ensemble of the five most accurate logistic regression classifiers, as determined in a five-fold cross-validation for different values of the hyperparameter C, which controls the degree of regularization. If the majority, which is at least three predictors, included a positive value or a negative value for a given protein family, we added this feature to the list of important features.

Classification evaluation The performance of each classifier was evaluated using nested five-fold cross-validation. The isolates were divided in five cross-validation folds. Each fold was once selected for testing the classifier, whereas the other folds were used to train the classifier. The parameter C was optimized in a further inner cross-validation step (Ruschhaupt et al. 2004). Macro-accuracy (MACC) and area under curve (AUC) were used as performance measures to evaluate the classifiers. The macro-accuracy is the average of the recall of the susceptible and the sensitive class (Manning et al. 2008). The AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve is computed for logistic regression models by varying the probability threshold that is required for a samples to be assigned to the resistant class each time recomputing the recall of the susceptible and the sensitive class (Fawcett 2006).

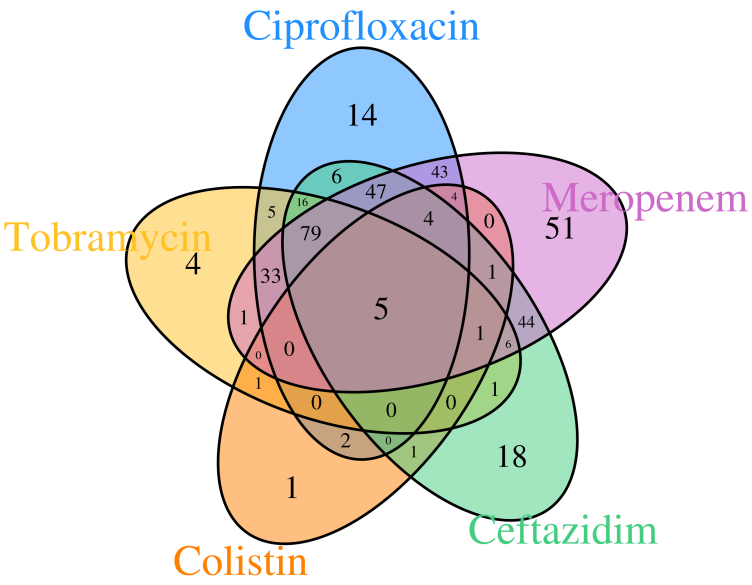
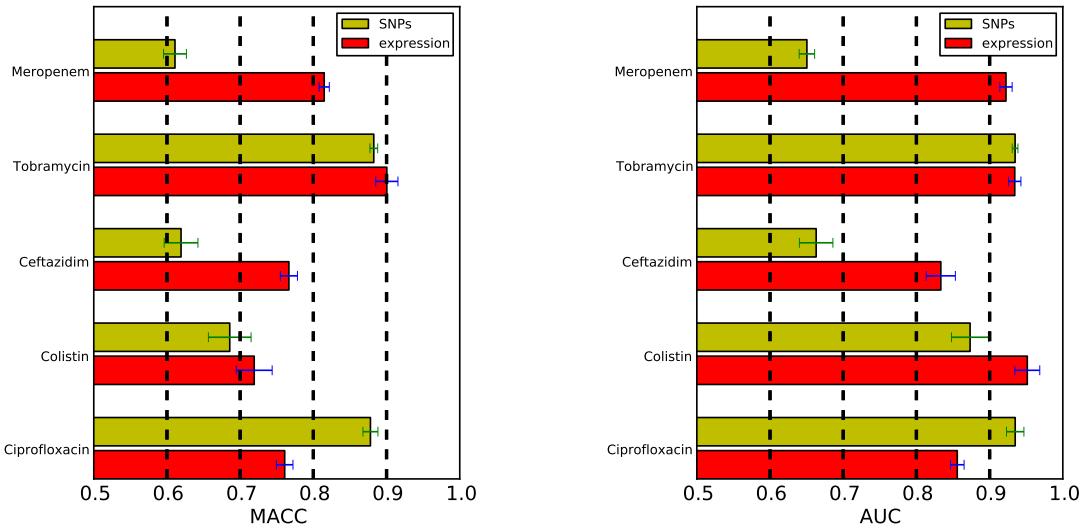


Figure 5.1: Cross-resistance of *P. aeruiginosa* isolates to five different antibiotics. Overlapping shapes represent regions of cross-resistances. The numbers in the intersections denote the number of resistant isolates. Intermediate resistant isolates were discarded for this analysis.



(a) **(b)**

Figure 5.2: We trained a logistic regression classifier on RNAseq data of 467 *P. aeruginosa* isolates to predict resistance to five different antibiotics. In five-fold nested cross-validation we estimated the macro accuracy (MACC) shown in a) and the area under the curve (AUC) show in b) based on expression and single nucleotide polymorphism profiles.

CHAPTER 6

Synopsis

The number of sequenced microbial genomes rapidly grows, creating a need for sophisticated method for genome interpreation to tackle problems in health and infection research. In my PhD project, I have developed three methods to link bacterial genotypes to phenotypes, each with a different focus. I first devised a classifier to predict the genomic components of microbial plant biomass degradation from the genomes of a set of diverse bacterial plant biomass degraders and non-degraders. Subsequently, I developed a software package providing a fully automated prediction of many traits, only requiring as input microbial genomes. Last, I used a large data set of clinical isolates of *P. aeruignosa* to reveal the genetic determinants of antibiotic resistance from transcriptional and mutational profiles.

The sparsity of available phenotype data represents a bottleneck for deriving further genotype-phenotype models (Dutilh et al. 2013). The costs for phenotype assays do not decrease at a similar speed as sequencing costs do. Together with my colleage Sebastian Konietzny, we have manully curated a set of plant biomass degrader and non-degraders from the biomedical literature. Furthermore, I extracted a large dataset of phenotypes from Bergey's systematic bacteriology and GIDEON (Berger 2005; Goodfellow et al. 2012). This exemplifies that existing resources like curated phenotype databases, as

well as the biomedical literature seem promising to provide the training data for learning additional genotype-phenotype associations. In the future, one could apply more advanced text mining approaches like natural lanuage processing to more efficiently and systematic search the microbiological literature.
Another avenue for future research could be to integrate expert knowledge of the biochemical pathways that are used in manual metabolic reconstructions as prior knowledge into the model for learning of the phenotype models. We have shown that antibiotic resistance can be predicted from mutational, as well as transcription profiles. Integrating different "data'omics" such as genomics and transcriptomics, may improve the prediction of phenotypes by joint inference on these data types (Franzosa et al. 2015). Finally, multi-label learning could be used to exploit dependencies across different phenotypes such as cross-resistance of different antibiotics (Zhang et al. 2014). Revealing the genetic components and predicting clinically relevant phenotypes such as drug resistance could, in the future, lead to more effective treatments and provide insights into pathogenesis. Uncovering the genotype-phenotype associations of biotechnologically important traits may deepen the understanding of microbial metabolism and reveal novel enzymes to be used in industrial processes. This thesis provides methods and results that could guide researchers on their path towards these goals.

Glossary

16S component of the 30S small subunit of prokarotic ribosomes.

AUC area under the curve.

C. acidaminovorans Cloacamonas acidaminovorans.

CAZyme Carbohydrate-active enzyme.

CBM carbohydrate binding module.

CE carbohydrate esterase.

CMC carboxymethyl cellulose.

contig contiguous sequence.

DNA desoxynucleic acid.

FN false negative.

FP false positive.

GFM genome from metagenome.

GH glycoside hydrolase.

GIDEON Global Infectious Disease and Epidemiology Network.

Gps-sTOL GIDEON phenotype-specific tree.

GT glycoside transferase.

GWAS genome-wide association study.

HMM hidden Markov model.

KEGG Kyoto Encyclopedia of Genes and Genomes.

MACC macro-accuracy.

MIC minium inhibitory concentration.

NCBI national center of biotechtechnology information.

NCBI-nr NCBI non-redundant database.

nCV nested cross-validation.

NGS Next-generation sequencing.

P. aeruignosa Pseudomonas aeruginosa.

phypat phyletic pattern classifier.

phypat+PGL phyletic pattern and protein gain and loss classifier.

PL polysaccharide lyase.

RNA ribonucleic acid.

RNAseq RNA sequencing.

ROC receiver operating characteristic.

SAG single amplified genome.

SNP short nucleotide polymorphism.

sTOL sequenced tree of life.

SVM support vector machine.

TN true negative.

TP true positive.

WGS whole genome sequencing.

References

Abubucker, S., N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methe, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower (2012). "Metabolic reconstruction for metagenomic data and its application to the human microbiome". *PLoS Computational Biology* 8.6. DOI: 10.1371/journal.pcbi.1002358 (cit. on p. 11).

Adrio, J. L. and A. L. Demain (2014). "Microbial enzymes: tools for biotechnological processes". *Biomolecules* 4.1, pp. 117–139. DOI: 10.3390/biom4010117 (cit. on p. 3).

Alneberg, J., B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince (2014). "Binning metagenomic contigs by coverage and composition". *Nature Methods* 11.11, pp. 1144–1146. DOI: 10.1038/nmeth.3103 (cit. on pp. 7, 48).

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic local alignment search tool". *Journal of Molecular Biology* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2 (cit. on p. 8).

Anderson, I., B. Abt, A. Lykidis, H. P. Klenk, N. Kyrpides, and N. Ivanova (2012). "Genomics of aerobic cellulose utilization systems in actinobacteria".

PLoS One 7.6, e39331. DOI: 10.1371/journal.pone.0039331 (cit. on pp. 26, 38).

Andoh, A., T. Tsujikawa, and Y. Fujiyama (2003). "Role of dietary fiber and short-chain fatty acids in the colon". *Current Pharmaceutical Design* 9.4, pp. 347–358 (cit. on p. 3).

Aravind, L. (2000). "Guilt by association: contextual information in genome analysis". *Genome Research* 10.8, pp. 1074–1077 (cit. on p. 9).

Aspeborg, H., P. M. Coutinho, Y. Wang, H. Brumer, and B. Henrissat (2012). "Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5)". *BMC Evolutionary Biology* 12, p. 186. DOI: 10.1186/1471-2148-12-186 (cit. on p. 27).

Bai, Y., D. B. Müller, G. Srinivas, R. Garrido-Oter, E. Potthoff, M. Rott, N. Dombrowski, P. C. Münch, S. Spaepen, M. Remus-Emsermann, B. Hüttel, A. C. McHardy, J. A. Vorholt, and P. Schulze-Lefert (2015). "Functional overlap of the Arabidopsis leaf and root microbiota". *Nature* 528.7582, pp. 364–369. DOI: 10.1038/nature16192 (cit. on p. 3).

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". *Journal of Computational Biology* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021 (cit. on p. 8).

Bankevich, A. and P. A. Pevzner (2016). "TruSPAdes: barcode assembly of TruSeq synthetic long reads". *Nature Methods* 13.3, pp. 248–250. DOI: 10.1038/nmeth.3737 (cit. on p. 7).

Barker, D. and M. Pagel (2005). "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes". *PLoS Computational Biology* 1.1. DOI: 10.1371/journal.pcbi.0010003 (cit. on pp. 10, 48, 67, 70).

Beerenwinkel, N., M. Dumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter (2003). "Geno2Pheno: estimating phenotypic drug resistance from HIV-1 genotypes." *Nucleic Acids Research* 31.13, pp. 3850–3855 (cit. on p. 24).

Berger, S. A. (2005). "GIDEON: a comprehensive Web-based resource for geographic medicine". *International Journal of Health Geographics* 4.10. DOI: 10.1186/1476-072X-4-10 (cit. on pp. 15, 48, 69, 87).

Boisvert, S., F. Laviolette, and J. Corbeil (2010). "Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies". *Journal of Computational Biology* 17.11, pp. 1519–1533. DOI: 10.1089/cmb.2009.0238 (cit. on p. 8).

Boisvert, S., F. Raymond, É. Godzaridis, F. Laviolette, and J. Corbeil (2012). "Ray Meta: scalable de novo metagenome assembly and profiling". *Genome Biology* 13, R122. DOI: 10.1186/gb-2012-13-12-r122 (cit. on p. 6).

Boraston, A., D. Bolam, H. Gilbert, and G. Davies (2004). "Carbohydrate-binding modules: fine-tuning polysaccharide recognition." *Biochemical Journal* 15, pp. 769–781 (cit. on p. 27).

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). "A training algorithm for optimal margin classifiers". 130401: ACM, pp. 144–152. DOI: 10.1145/130385.130401 (cit. on pp. 12, 24, 41, 70).

Bradley, P., N. C. Gordon, T. M. Walker, L. Dunn, S. Heys, B. Huang, S. Earle, L. J. Pankhurst, L. Anson, M. de Cesare, P. Piazza, A. A. Votintseva, T. Golubchik, D. J. Wilson, D. H. Wyllie, R. Diel, S. Niemann, S. Feuerriegel, T. A. Kohl, N. Ismail, S. V. Omar, E. G. Smith, D. Buck, G. McVean, A. S. Walker, T. E. A. Peto, D. W. Crook, and Z. Iqbal (2015). "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis". *Nature Communications* 6 (cit. on p. 79).

Bremges, A., E. Singer, T. Woyke, and A. Sczyrba (2016). "MeCorS: Metagenome-enabled error correction of single cell sequencing reads". *Bioinformatics*, btw144. DOI: 10.1093/bioinformatics/btw144 (cit. on p. 8).

Brown, C. T., L. A. Hug, B. C. Thomas, I. Sharon, C. J. Castelle, A. Singh, M. J. Wilkins, K. C. Wrighton, K. H. Williams, and J. F. Banfield (2015). "Unusual biology across a group comprising more than 15% of domain Bacteria". *Nature* 523.7559, pp. 208–11. DOI: 10.1038/nature14486 (cit. on pp. 7, 68).

Brulc, J., D. Antonopoulos, M. Berg Miller, M. Wilson, A. Yannarell, E. Dinsdale, R. Edwards, E. Frank, J. Emerson, P. Wacklin, P. Coutinho, B. Henrissat, K. Nelson, and B. White (2009). "Gene-centric metagenomics of the

fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases." *Proceedings of the National Academy of Sciences* 106.6, p. 1948 (cit. on p. 23).

Brumm, P., S. Hermanson, B. Hochstein, J. Boyum, N. Hermersmann, K. Gowda, and D. Mead (2010). "Mining Dictyoglomus turgidum for Enzymatically Active Carbohydrases". *Applied Biochemistry and Biotechnology* DOI 10.1007/s12010-010-9029-6 (cit. on p. 23).

Brumm, P., D. Mead, J. Boyum, C. Drinkwater, K. Gowda, D. Stevenson, and P. Weimer (2010). "Functional Annotation of Fibrobacter succinogenes S85 Carbohydrate Active Enzymes". *Applied Biochemistry and Biotechnology* DOI 10.1007/s12010-010-9070-5 (cit. on p. 23).

Bulgarelli, D., R. Garrido-Oter, P. C. Münch, A. Weiman, J. Dröge, Y. Pan, A. C. McHardy, and P. Schulze-Lefert (2015). "Structure and function of the bacterial root microbiota in wild and domesticated barley". *Cell Host & Microbe* 17.3, pp. 392–403. DOI: 10.1016/j.chom.2015.01.011 (cit. on p. 18).

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat (2009). "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics". *Nucleic Acids Research* 37.Database issue, pp. D233–238. DOI: 10.1093/nar/gkn663 (cit. on p. 39).

Caspi, R., R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, and P. D. Karp (2016). "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases". *Nucleic Acids Research* 44.D1, pp. D471–D480. DOI: 10.1093/nar/gkv1164 (cit. on p. 10).

Chen, P. E. and B. J. Shapiro (2015). "The advent of genome-wide association studies for bacteria". *Current Opinion in Microbiology*. Environmental microbiology - Extremophiles 25, pp. 17–24. DOI: 10.1016/j.mib.2015.03.002 (cit. on pp. 10, 80).

Clark, A. G. et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny". *Nature* 450.7167, pp. 203–18. DOI: 10.1038/nature06341 (cit. on p. 14).

Cleary, B., I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm (2015). "Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning". *Nature Biotechnology* 33.10, pp. 1053–60. DOI: 10.1038/nbt.3329 (cit. on pp. 7, 48).

Cohen, O. and T. Pupko (2011). "Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony–a simulation study". *Genome Biology Evolution* 3, pp. 1265–75. DOI: 10.1093/gbe/evr101 (cit. on p. 70).

Cohen, O., H. Ashkenazy, E. L. Karin, D. Burstein, and T. Pupko (2013). "CoPAP: Coevolution of Presence-Absence Patterns". *Nucleic Acids Research* 41.W1, W232–W237. DOI: 10.1093/nar/gkt471 (cit. on p. 10).

Cortes, C. and V. Vapnik (1995). "Support-vector networks." *Machine Learning* 20.3, pp. 273–297 (cit. on pp. 24, 41).

Costerton, J. W. (2001). "Cystic fibrosis pathogenesis and the role of biofilms in persistent infection". *Trends in Microbiology* 9.2, pp. 50–52 (cit. on p. 78).

DeBoy, R., E. Mongodin, D. Fouts, L. Tailford, H. Khouri, J. Emerson, Y. Mohamoud, K. Watkins, B. Henrissat, H. Gilbert, and K. Nelson (2008). "Insights into plant cell wall degradation from the genome sequence of the soil bacterium Cellvibrio japonicus." *Journal of Bacteriology* 190.15, pp. 5455–5463 (cit. on p. 36).

Didelot, X., R. Bowden, D. J. Wilson, T. E. A. Peto, and D. W. Crook (2012). "Transforming clinical microbiology with bacterial genome sequencing". *Nature Reviews Genetics* 13.9, pp. 601–612. DOI: 10.1038/nrg3226 (cit. on p. 78).

Draper, J., K. Karplus, and K. M. Ottemann (2011). "Identification of a chemoreceptor zinc-binding domain common to cytoplasmic bacterial chemoreceptors". *Journal of Bacteriology* 193.17, pp. 4338–45. DOI: 10.1128/JB.05140-11 (cit. on p. 62).

Dröge, J. and A. C. McHardy (2012). "Taxonomic binning of metagenome samples generated by next-generation sequencing technologies". *Briefings in Bioinformatics* 13.6, pp. 646–655. DOI: 10.1093/bib/bbs031 (cit. on p. 7).

Duan, C. J. and J. X. Feng (2010). "Mining metagenomes for novel cellulase genes". *Biotechnology Letters* 32.12, pp. 1765–1775. DOI: 10.1007/s10529-010-0356-z (cit. on p. 38).

Durot, M., P.-Y. Bourguignon, and V. Schachter (2009). "Genome-scale models of bacterial metabolism: reconstruction and applications". *FEMS Microbiology Reviews* 33.1, pp. 164–190. DOI: 10.1111/j.1574-6976.2008.00146.x (cit. on p. 11).

Dutilh, B. E., L. Backus, R. A. Edwards, M. Wels, J. R. Bayjanov, and S. A. F. T. v. Hijum (2013). "Explaining microbial phenotypes on a genomic scale: GWAS for microbes". *Briefings in Functional Genomics* 12.4, pp. 366–380. DOI: 10.1093/bfgp/elt008 (cit. on pp. 10, 87).

Dutta, R., L. Qin, and M. Inouye (1999). "Histidine kinases: diversity of domain organization". *Molecular Microbiology* 34.4, pp. 633–40 (cit. on p. 62).

ECDC, E. (2009). "The bacterial challenge: time to react". *Stockholm: European Center for Disease Prevention and Control* (cit. on p. 78).

Eddy, S. R. (1998). "Profile hidden Markov models". *Bioinformatics (Oxford, England)* 14.9, pp. 755–763 (cit. on p. 9).

Enright, A. J., I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis (1999). "Protein interaction maps for complete genomes based on gene fusion events". *Nature* 402.6757, pp. 86–90. DOI: 10.1038/47056 (cit. on p. 9).

Everard, A. and P. D. Cani (2013). "Diabetes, obesity and gut microbiota". *Best Practice & Research. Clinical Gastroenterology* 27.1, pp. 73–83. DOI: 10.1016/j.bpg.2013.03.007 (cit. on p. 3).

Fan, R. E., K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin (2008). "LIBLINEAR: A Library for Large Linear Classification". *Journal of Machine Learning Research* 9, pp. 1871–1874 (cit. on pp. 12, 13, 41, 42, 70, 84).

Fang, H., M. E. Oates, R. B. Pethica, J. M. Greenwood, A. J. Sardar, O. J. Rackham, P. C. Donoghue, A. Stamatakis, D. A. de Lima Morais, and J. Gough (2013). "A daily-updated tree of (sequenced) life as a reference for genome research". *Scientific Reports* 3. DOI: 10.1038/srep02015 (cit. on p. 71).

Fawcett, T. (2006). "An Introduction to ROC Analysis". *Pattern Recognition Letters* 27.8, pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010 (cit. on p. 84).

Feldbauer, R., F. Schulz, M. Horn, and T. Rattei (2015). "Prediction of microbial phenotypes based on comparative genomics". *BMC Bioinformatics* 16 Suppl 14. DOI: 10.1186/1471-2105-16-S14-S1 (cit. on pp. 14, 15, 65).

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta (2014). "Pfam: the protein families database". *Nucleic Acids Research* 42.Database issue, pp. D222–30. DOI: 10.1093/nar/gkt1223 (cit. on p. 68).

Finn, R. D., J. Clements, and S. R. Eddy (2011). "HMMER web server: interactive sequence similarity searching". *Nucleic Acids Research* 39.Web Server issue, W29–37. DOI: 10.1093/nar/gkr367 (cit. on pp. 39, 68).

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman (2016). "The Pfam protein families database: towards a more sustainable future". *Nucleic Acids Research* 44.D1, pp. D279–D285. DOI: 10.1093/nar/gkv1344 (cit. on p. 9).

Frank, J. A., Y. Pan, A. Tooming-Klunderud, V. G. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope (2015). "Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data". *bioRxiv*. DOI: 10.1101/026922 (cit. on p. 64).

Franzosa, E. A., T. Hsu, A. Sirota-Madi, A. Shafquat, G. Abu-Ali, X. C. Morgan, and C. Huttenhower (2015). "Sequencing and beyond: integrating molecular 'omics' for microbial community profiling". *Nature Reviews Microbiology* 13.6, pp. 360–372. DOI: 10.1038/nrmicro3451 (cit. on p. 88).

Gabriel, S., L. Ziaugra, and D. Tabbaa (2009). "SNP genotyping using the Sequenom MassARRAY iPLEX platform". *Current protocols in human genetics*, pp. 2–12 (cit. on p. 83).

Gawad, C., W. Koh, and S. R. Quake (2016). "Single-cell genome sequencing: current state of the science". *Nature Reviews Genetics* 17.3, pp. 175–188. DOI: 10.1038/nrg.2015.16 (cit. on p. 8).

Gevers, D., R. Knight, J. F. Petrosino, K. Huang, A. L. McGuire, B. W. Birren, K. E. Nelson, O. White, B. A. Methé, and C. Huttenhower (2012). "The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome". *PLOS Biology* 10.8, e1001377. DOI: 10.1371/journal.pbio.1001377 (cit. on pp. 2, 3).

Gilead, S. and Y. Shoham (1995). "Purification and characterization of alpha-L-arabinofuranosidase from Bacillus stearothermophilus T-6". *Applied and Environmental Microbiology* 61.1, pp. 170–4 (cit. on p. 64).

Goodfellow, M., P. Kämpfer, H.-J. Busse, M. E. Trujillo, K.-i. Suzuki, W. Ludwig, and W. B. Whitman (2012). *Bergey's manual of systematic bacteriology*. Springer New York. ISBN: 0-387-68233-3 (cit. on pp. 2, 15, 47, 49, 65, 69, 87).

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies". *Nature Reviews Genetics* 17.6, pp. 333–351. DOI: 10.1038/nrg.2016.49 (cit. on p. 4).

Gregor, I., J. Droge, M. Schirmer, C. Quince, and A. C. McHardy (2016). "PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes". *PeerJ* 4. DOI: 10.7717/peerj.1603 (cit. on pp. 7, 48).

Grice, E. A. and J. A. Segre (2012). "The human microbiome: our second genome". *Annual Review of Genomics and Human Genetics* 13, pp. 151–170. DOI: 10.1146/annurev-genom-090711-163814 (cit. on p. 3).

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies". *Bioinformatics* 29.8, pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086 (cit. on p. 7).

Hacquard, S., B. Kracher, K. Hiruma, P. C. Münch, R. Garrido-Oter, M. R. Thon, A. Weimann, U. Damm, J.-F. Dallery, M. Hainaut, B. Henrissat, O. Lespinet, S. Sacristán, E. Ver Loren van Themaat, E. Kemen, A. C. McHardy, P. Schulze-Lefert, and R. J. O'Connell (2016). "Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi". *Nature Communications* 7, p. 11362. DOI: 10.1038/ncomms11362 (cit. on p. 17).

Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White (2001). "TIGRFAMs: a protein family resource for the func-

tional identification of proteins". *Nucleic Acids Research* 29.1, pp. 41–3 (cit. on p. 35).

Hardesty, D. (n.d.). *Iron hydroxide precipitate (orange) in a Missouri stream receiving acid drainage from surface coal mining.* URL: https://commons.wikimedia.org/wiki/File:Iron_hydroxide_precipitate_in_stream.jpg (visited on 08/25/2016) (cit. on p. 2).

Harvey, P. H. and M. D. Pagel (1991). *The comparative method in evolutionary biology*. Vol. 239. Oxford University Press Oxford (cit. on p. 48).

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning 2nd edition*. New York: Springer (cit. on pp. 11, 13).

Hervé, C., A. Rogowski, A. Blake, S. Marcus, H. Gilbert, and J. Knox (2010). "Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects." *Proceedings of the National Academy of Sciences* 107.34, pp. 15293–15298 (cit. on p. 36).

Hess, M., A. Sczyrba, R. Egan, T. W. Kim, H. Chokhawala, G. Schroth, S. Luo, D. S. Clark, F. Chen, T. Zhang, R. I. Mackie, L. A. Pennacchio, S. G. Tringe, A. Visel, T. Woyke, Z. Wang, and E. M. Rubin (2011). "Metagenomic discovery of biomass-degrading genes and genomes from cow rumen". *Science* 331.6016, pp. 463–7. DOI: 10.1126/science.1200387 (cit. on pp. 3, 7, 14, 23, 32–34, 47, 68).

Himmel, M. E., S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady, and T. D. Foust (2007). "Biomass recalcitrance: engineering plants and enzymes for biofuels production". *Science* 315.5813, pp. 804–7. DOI: 10.1126/science.1137016 (cit. on p. 22).

Hirsch, E. B. and V. H. Tam (2010). "Impact of multidrug-resistant Pseudomonas aeruginosa infection on patient outcomes". *Expert Review of Pharmacoeconomics & Outcomes Research* 10.4, pp. 441–451. DOI: 10.1586/erp.10.49 (cit. on p. 78).

Hosking, E. R., C. Vogt, E. P. Bakker, and M. D. Manson (2006). "The Escherichia coli MotAB proton channel unplugged". *Journal of Molecular Biology* 364.5, pp. 921–37. DOI: 10.1016/j.jmb.2006.09.035 (cit. on p. 62).

Howe, A. C., J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown (2014). "Tackling soil diversity with the assembly of large, com-

plex metagenomes". *Proceedings of the National Academy of Sciences* 111.13, pp. 4904–4909. DOI: 10.1073/pnas.1402564111 (cit. on p. 6).

Howe, A. and P. S. G. Chain (2015). "Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial)". *Frontiers in Microbiology* 6. DOI: 10.3389/fmicb.2015.00678 (cit. on pp. 5, 6).

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork (2016). "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences". *Nucleic Acids Research* 44.D1, pp. D286–D293. DOI: 10.1093/nar/gkv1248 (cit. on p. 9).

Hugenholtz, P., B. M. Goebel, and N. R. Pace (1998). "Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity". *Journal of Bacteriology* 180.24, p. 6793 (cit. on p. 5).

Hyatt, D., G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". *BMC Bioinformatics* 11. DOI: 10.1186/1471-2105-11-119 (cit. on pp. 8, 68).

Imelfort, M., D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson (2014). "GroopM: an automated tool for the recovery of population genomes from related metagenomes". *PeerJ* 2. DOI: 10.7717/peerj.603 (cit. on pp. 7, 48).

Ivanova, N., J. Sikorski, O. Chertkov, M. Nolan, S. Lucas, N. Hammon, S. Deshpande, J.-F. Cheng, R. Tapia, C. Han, L. Goodwin, S. Pitluck, M. Huntemann, K. Liolios, I. Pagani, K. Mavromatis, G. Ovchinikova, A. Pati, A. Chen, K. Palaniappan, M. Land, L. Hauser, E.-M. Brambilla, K. P. Kannan, M. Rohde, B. J. Tindall, M. Göker, J. C. Detter, T. Woyke, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, N. C. Kyrpides, H.-P. Klenk, and A. Lapidus (2011). "Complete genome sequence of the extremely halophilic Halanaerobium praevalens type strain (GSL)". *Standards in Genomic Sciences* 4.3, pp. 312–321. DOI: 10.4056/sigs.1824509 (cit. on pp. 26, 38).

Jaroszewski, L., Z. Li, S. S. Krishna, C. Bakolitsa, J. Wooley, A. M. Deacon, I. A. Wilson, and A. Godzik (2009). "Exploration of uncharted regions of the protein universe". *PLoS Biology* 7.9, e1000205. DOI: 10.1371/journal.pbio.1000205 (cit. on p. 8).

Josenhans, C. and S. Suerbaum (2002). "The role of motility as a virulence factor in bacteria". *International Journal of Medical Microbiology* 291.8, pp. 605–14. DOI: 10.1078/1438-4221-00173 (cit. on p. 61).

Juan, C., M. C. Conejo, N. Tormo, C. Gimeno, Á. Pascual, and A. Oliver (2013). "Challenges for accurate susceptibility testing, detection and interpretation of β-lactam resistance phenotypes in Pseudomonas aeruginosa: results from a Spanish multicentre study". *The Journal of Antimicrobial Chemotherapy* 68.3, pp. 619–630. DOI: 10.1093/jac/dks439 (cit. on p. 82).

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe (2015). "KEGG as a reference resource for gene and protein annotation". *Nucleic Acids Research*, gkv1070. DOI: 10.1093/nar/gkv1070 (cit. on p. 10).

Kang, D. D., J. Froula, R. Egan, and Z. Wang (2015). "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities". *PeerJ* 3. DOI: 10.7717/peerj.1165 (cit. on pp. 7, 48).

Kastenmuller, G., M. E. Schenk, J. Gasteiger, and H. W. Mewes (2009). "Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes". *Genome Biology* 10.3. DOI: 10.1186/gb-2009-10-3-r28 (cit. on pp. 14, 65).

Kaylen, M., D. Van Dyne, Y. Choi, and M. Blasé (2000). "Economic feasibility of producing ethanol from lignocellulosic feedstocks." *Bioresource Technology* 72, pp. 19–32 (cit. on p. 22).

Kensche, P. R., V. van Noort, B. E. Dutilh, and M. A. Huynen (2008). "Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution". *Journal of the Royal Society, Interface / the Royal Society* 5.19, pp. 151–170. DOI: 10.1098/rsif.2007.1047 (cit. on p. 10).

Khaledi, A., M. Schniederjans, S. Pohl, R. Rainer, U. Bodenhofer, B. Xia, F. Klawonn, S. Bruchmann, M. Preusse, D. Eckweiler, A. Dötsch, and S. Häussler (2016). "Transcriptome profiling of antimicrobial resistance in Pseudomonas aeruginosa". *Antimicrobial Agents and Chemotherapy*, AAC.00075–16. DOI: 10.1128/AAC.00075-16 (cit. on pp. 10, 79, 80).

Konietzny, S. G. A., P. B. Pope, A. Weimann, and A. C. McHardy (2014). "Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders". *Biotechnology for Biofuels* 7.124. DOI: 10.1186/s13068-014-0124-8 (cit. on pp. 14, 18, 48, 65).

Konietzny, S. G. A., L. Dietz, and A. C. McHardy (2011). "Inferring functional modules of protein families with probabilistic topic models". *BMC Bioinformatics* 12, p. 141. DOI: 10.1186/1471-2105-12-141 (cit. on p. 10).

Kostic, A. D., R. J. Xavier, and D. Gevers (2014). "The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead". *Gastroenterology*. The Gut Microbiome in Health and Disease 146.6, pp. 1489–1499. DOI: 10.1053/j.gastro.2014.02.009 (cit. on p. 3).

Krueger, J., S. Pohl, M. Preusse, A. Kordes, N. Rugen, M. Schniederjans, A. Pich, and S. Häussler (2016). "Unravelling post-transcriptional PrmC-dependent regulatory mechanisms in Pseudomonas aeruginosa". *Environmental Microbiology*, n/a–n/a. DOI: 10.1111/1462-2920.13435 (cit. on p. 83).

Laehnemann, D., A. Borkhardt, and A. C. McHardy (2016). "Denoising DNA deep sequencing data - high-throughput sequencing errors and their correction". *Briefings in Bioinformatics* 17.1, pp. 154–179. DOI: 10.1093/bib/bbv029 (cit. on p. 4).

Lam, W. W., E. J. Woo, M. Kotaka, W. K. Tam, Y. C. Leung, T. K. Ling, and S. W. Au (2010). "Molecular interaction of flagellar export chaperone FliS and cochaperone HP1076 in Helicobacter pylori". *Federation of American Societies For Experimental Biology Journal* 24.10, pp. 4020–32. DOI: 10.1096/fj.10-155242 (cit. on p. 61).

Land, M., L. Hauser, S.-R. Jun, I. Nookaew, M. R. Leuze, T.-H. Ahn, T. Karpinets, O. Lund, G. Kora, T. Wassenaar, S. Poudel, and D. W. Ussery (2015). "Insights from 20 years of bacterial genome sequencing". *Functional & Integrative Genomics* 15.2, pp. 141–161. DOI: 10.1007/s10142-015-0433-4 (cit. on p. 4).

Lasken, R. S. and J. S. McLean (2014). "Recent advances in genomic DNA sequencing of microbial species from single cells". *Nature Reviews Genetics* 15.9, pp. 577–84. DOI: 10.1038/nrg3785 (cit. on pp. 8, 48).

Lee, D., O. Redfern, and C. Orengo (2007). "Predicting protein function from sequence and structure". *Nature Reviews. Molecular Cell Biology* 8.12, pp. 995–1005. DOI: 10.1038/nrm2281 (cit. on p. 8).

Lee, J. (1997). "Biological conversion of lignocellulosic biomass to ethanol." *Journal of Biotechnology* 56, pp. 1–24 (cit. on p. 22).

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1. G. P. D. P. Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352 (cit. on p. 83).

Lieberman, T. D., J.-B. Michel, M. Aingaran, G. Potter-Bynoe, D. Roux, M. R. Davis, D. Skurnik, N. Leiby, J. J. LiPuma, J. B. Goldberg, A. J. McAdam, G. P. Priebe, and R. Kishony (2011). "Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes". *Nature Genetics* 43.12, pp. 1275–1280. DOI: 10.1038/ng.997 (cit. on p. 79).

Limam, R. D., R. Chouari, L. Mazeas, T. D. Wu, T. Li, J. Grossin-Debattista, J. L. Guerquin-Kern, M. Saidi, A. Landoulsi, A. Sghir, and T. Bouchez (2014). "Members of the uncultured bacterial candidate division WWE1 are implicated in anaerobic digestion of cellulose". *Microbiologyopen* 3.2, pp. 157–67. DOI: 10.1002/mbo3.144 (cit. on p. 58).

Lingner, T., S. Muhlhausen, T. Gabaldon, C. Notredame, and P. Meinicke (2010). "Predicting phenotypic traits of prokaryotes from protein domain frequencies". *BMC Bioinformatics* 11.481. DOI: 10.1186/1471-2105-11-481 (cit. on pp. 10, 14, 65).

Liu, B. and M. Pop (2011). "MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets". *BMC Proceedings* 5.2, pp. 1–12. DOI: 10.1186/1753-6561-5-S2-S9 (cit. on p. 11).

Liu, R. and H. Ochman (2007). "Stepwise formation of the bacterial flagellar system". *Proceedings of the National Academy of Sciences of the United States of America* 104.17, pp. 7116–21. DOI: 10.1073/pnas.0700266104 (cit. on p. 61).

Liu, Y., J. Li, L. Sam, C. S. Goh, M. Gerstein, and Y. A. Lussier (2006). "An integrative genomic approach to uncover molecular mechanisms of prokaryotic

traits". *PLoS Computational Biology* 2.11. DOI: 10.1371/journal.pcbi.0020159 (cit. on pp. 10, 70).

Lynd, L. R., P. J. Weimer, W. H. van Zyl, and I. S. Pretorius (2002). "Microbial cellulose utilization: fundamentals and biotechnology". *Microbiology and olecular Biology Reviews* 66.3, pp. 506–77 (cit. on pp. 38, 39).

MacDonald, N. J. and R. G. Beiko (2010). "Efficient learning of microbial genotype-phenotype association rules". *Bioinformatics* 26.15, pp. 1834–40. DOI: 10.1093/bioinformatics/btq305 (cit. on pp. 10, 15, 65, 67, 68).

Macnab, R. M. (2003). "How bacteria assemble flagella". *Annual Review Microbiology* 57, pp. 77–100. DOI: 10.1146/annurev.micro.57.030502.090832 (cit. on p. 48).

Magiorakos, A.-P., A. Srinivasan, R. B. Carey, Y. Carmeli, M. E. Falagas, C. G. Giske, S. Harbarth, J. F. Hindler, G. Kahlmeter, B. Olsson-Liljequist, D. L. Paterson, L. B. Rice, J. Stelling, M. J. Struelens, A. Vatopoulos, J. T. Weber, and D. L. Monnet (2012). "Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance". *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 18.3, pp. 268–281. DOI: 10.1111/j.1469-0691.2011.03570.x (cit. on p. 78).

Manichanh, C., L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca, and J. Dore (2006). "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach". *Gut* 55.2, pp. 205–211. DOI: 10.1136/gut.2005.073817 (cit. on p. 3).

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge, UK (cit. on pp. 74, 84).

Markowitz, V. M., I. M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides (2012). "IMG: the Integrated Microbial Genomes database and comparative analysis system". *Nucleic Acids*

Research 40.Database issue, pp. D115–22. DOI: 10.1093/nar/gkr1044 (cit. on p. 39).

Markowitz, V. M., N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. M. A. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugenholtz, and N. C. Kyrpides (2008). "IMG/M: a data management and analysis system for metagenomes". *Nucleic Acids Research* 36.suppl_1, pp. D534–538. DOI: 10.1093/nar/gkm869 (cit. on p. 39).

Martinez, D., R. M. Berka, B. Henrissat, M. Saloheimo, M. Arvas, S. E. Baker, J. Chapman, O. Chertkov, P. M. Coutinho, D. Cullen, E. G. Danchin, I. V. Grigoriev, P. Harris, M. Jackson, C. P. Kubicek, C. S. Han, I. Ho, L. F. Larrondo, A. L. de Leon, J. K. Magnuson, S. Merino, M. Misra, B. Nelson, N. Putnam, B. Robbertse, A. A. Salamov, M. Schmoll, A. Terry, N. Thayer, A. Westerholm-Parvinen, C. L. Schoch, J. Yao, R. Barabote, M. A. Nelson, C. Detter, D. Bruce, C. R. Kuske, G. Xie, P. Richardson, D. S. Rokhsar, S. M. Lucas, E. M. Rubin, N. Dunn-Coleman, M. Ward, and T. S. Brettin (2008). "Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina)". *Nature Biotechnology* 26.5, pp. 553–60. DOI: 10.1038/nbt1403 (cit. on p. 64).

Martinez, J. L. (2008). "Antibiotics and antibiotic resistance genes in natural environments". *Science* 321.5887, pp. 365–7. DOI: 10.1126/science.1159483 (cit. on pp. 2, 48).

Martiny, J. B., S. E. Jones, J. T. Lennon, and A. C. Martiny (2015). "Microbiomes in light of traits: A phylogenetic perspective". *Science* 350.6261. DOI: 10.1126/science.aac9323 (cit. on pp. 47, 48).

Meletis, G. and M. Bagkeri (2013). "Pseudomonas aeruginosa: Multi-Drug-Resistance Development and Treatment Options". *Infection Control*, p. 33 (cit. on pp. 78, 81).

Michalopoulos, A. S., S. Tsiodras, K. Rellos, S. Mentzelopoulos, and M. E. Falagas (2005). "Colistin treatment in patients with ICU-acquired infections caused by multiresistant Gram-negative bacteria: the renaissance of an old antibiotic". *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 11.2, pp. 115–121. DOI: 10.1111/j.1469-0691.2004.01043.x (cit. on p. 78).

Mikheenko, A., V. Saveliev, and A. Gurevich (2015). "MetaQUAST: evaluation of metagenome assemblies". *Bioinformatics*, btv697. DOI: 10.1093/bioinformatics/btv697 (cit. on p. 7).

Mitchell, W. (1998). "Physiology of carbohydrate to solvent conversion by clostridia." *Advances in Microbial Physiology* 39, pp. 31–130 (cit. on p. 22).

Moreno-Vivian, C., P. Cabello, M. Martinez-Luque, R. Blasco, and F. Castillo (1999). "Prokaryotic nitrate reduction: molecular properties and functional distinction among bacterial nitrate reductases". *Journal of Bacteriology* 181.21, pp. 6573–84 (cit. on pp. 62, 63).

Morrison, M., P. Pope, S. Denman, and C. McSweeney (2009). "Plant biomass degradation by gut microbiomes: more of the same or something new?" *Current Opinion in Biotechnology* 20, pp. 358–363 (cit. on p. 23).

Moskowitz, S. M., R. K. Ernst, and S. I. Miller (2004). "PmrAB, a Two-Component Regulatory System of Pseudomonas aeruginosa That Modulates Resistance to Cationic Antimicrobial Peptides and Addition of Aminoarabinose to Lipid A". *Journal of Bacteriology* 186.2, pp. 575–579. DOI: 10.1128/JB.186.2.575-579.2004 (cit. on p. 81).

Naas, A. E., A. K. Mackenzie, J. Mravec, J. Schückel, W. G. T. Willats, V. G. H. Eijsink, and P. B. Pope (2014). "Do Rumen Bacteroidetes Utilize an Alternative Mechanism for Cellulose Degradation?" *mBio* 5.4, e01401–14. DOI: 10.1128/mBio.01401-14 (cit. on p. 14).

Narihiro, T. and Y. Sekiguchi (2007). "Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update". *Current Opinion in Biotechnology* 18.3, pp. 273–278. DOI: 10.1016/j.copbio.2007.04.003 (cit. on p. 3).

NASA (n.d.). *The ice wedge in permafrost.* URL: http://science.nasa.gov/newhome/headlines/ast27jul99_1.htm (visited on 08/25/2016) (cit. on p. 2).

Nielsen, H. B., M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J. M. Batto, M. B. Quintanilha Dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Dore, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas,

S. Kennedy, K. Kristiansen, J. R. Kultima, P. Leonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sorensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, H. I. T. C. Meta, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, S. D. Ehrlich, and H. I. T. C. Meta (2014). "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes". *Nature Biotechnology* 32.8, pp. 822–8. DOI: 10.1038/nbt.2939 (cit. on pp. 7, 48).

Sa-Nogueira, I., T. V. Nogueira, S. Soares, and H. de Lencastre (1997). "The Bacillus subtilis L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression". *Microbiology* 143 (Pt 3), pp. 957–69 (cit. on p. 64).

Ochman, H., J. G. Lawrence, and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation". *Nature* 405.6784, pp. 299–304. DOI: 10.1038/35012500 (cit. on p. 48).

Olapade, O. A. and A. J. Ronk (2015). "Isolation, characterization and community diversity of indigenous putative toluene-degrading bacterial populations with catechol-2,3-dioxygenase genes in contaminated soils". *Microbial Ecology* 69.1, pp. 59–65. DOI: 10.1007/s00248-014-0466-6 (cit. on p. 3).

Olson, D., S. Tripathi, R. Giannone, J. Lo, N. Caiazza, D. Hogsett, R. Hettich, A. Guss, G. Dubrovsky, and L. Lynd (2010). "Deletion of the Cel48S cellulase from Clostridium thermocellum". *Proceedings of the National Academy of Sciences* doi: 10.1073/pnas.1003584107 (cit. on p. 36).

Pal, C., B. Papp, and M. J. Lercher (2005). "Adaptive evolution of bacterial metabolic networks by horizontal gene transfer". *Nature Genet* 37.12, pp. 1372–5. DOI: 10.1038/ng1686 (cit. on p. 48).

Palmer, A. C. and R. Kishony (2013). "Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance". *Nature Reviews Genetics* 14.4, pp. 243–248. DOI: 10.1038/nrg3351 (cit. on p. 79).

Pantel, I., P. E. Lindgren, H. Neubauer, and F. Gotz (1998). "Identification and characterization of the Staphylococcus carnosus nitrate reductase operon". *Molecular Genetics and Genomics* 259.1, pp. 105–14 (cit. on pp. 62, 63).

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson (2015). "CheckM: assessing the quality of microbial genomes recovered from

isolates, single cells, and metagenomes". *Genome Research* 25.7, pp. 1043–55. DOI: 10.1101/gr.186072.114 (cit. on pp. 7, 61).

Paterson, G. K. and T. J. Mitchell (2004). "The biology of Gram-positive sortase enzymes". *Trends in Microbiology* 12.2, pp. 89–95. DOI: 10.1016/j.tim.2003.12.007 (cit. on p. 65).

Pechère, J.-C. and T. Köhler (1999). "Patterns and modes of beta-lactam resistance in Pseudomonas aeruginosa". *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 5 Suppl 1, S15–S18 (cit. on p. 78).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg (2011). "Scikit-learn: Machine learning in Python". *Journal of Machine Learning* 12, pp. 2825–30 (cit. on p. 70).

Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin (2011). "Meta-IDBA: a de Novo assembler for metagenomic data". *Bioinformatics* 27.13, pp. i94–i101. DOI: 10.1093/bioinformatics/btr216 (cit. on p. 6).

Platt, J. C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, pp. 61–74 (cit. on p. 13).

Poole, K. (2005). "Aminoglycoside Resistance in Pseudomonas aeruginosa". *Antimicrobial Agents and Chemotherapy* 49.2, pp. 479–487. DOI: 10.1128/AAC.49.2.479-487.2005 (cit. on p. 81).

Pope, P. B., A. K. Mackenzie, I. Gregor, W. Smith, M. A. Sundset, A. C. McHardy, M. Morrison, and V. G. Eijsink (2012). "Metagenomics of the Svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci". *PLoS One* 7.6, e38571. DOI: 10.1371/journal.pone.0038571 (cit. on p. 23).

Pope, P. B., W. Smith, S. E. Denman, S. G. Tringe, K. Barry, P. Hugenholtz, C. S. McSweeney, A. C. McHardy, and M. Morrison (2011). "Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies". *Science* 333.6042, pp. 646–648. DOI: 10.1126/science.1205760 (cit. on p. 3).

Pope, P., S. Denman, M. Jones, S. Tringe, K. Barry, S. Malfatti, A. McHardy, J.-F. Cheng, P. Hugenholtz, C. McSweeney, and M. Morrison (2010). "Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different to other herbivores". *Proc. Natl. Acad. Sci. USA* 107, pp. 14793–14798 (cit. on pp. 23, 48).

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn (2012). "The Pfam protein families database". *Nucleic Acids Research* 40.Database issue, pp. D290–301. DOI: 10.1093/nar/gkr1065 (cit. on pp. 35, 39, 68).

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Antolin, F. Artiguenave, H. Blottiere, N. Borruel, T. Bruls, F. Casellas, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, M. Forte, C. Friss, M. v. d. Guchte, E. Guedon, F. Haimet, A. Jamet, C. Juste, G. Kaci, M. Kleerebezem, J. Knol, M. Kristensen, S. Layec, K. L. Roux, M. Leclerc, E. Maguin, R. M. Minardi, R. Oozeer, M. Rescigno, N. Sanchez, S. Tims, T. Torrejon, E. Varela, W. d. Vos, Y. Winogradsky, E. Zoetendal, P. Bork, S. D. Ehrlich, and J. Wang (2010). "A human gut microbial gene catalogue established by metagenomic sequencing". *Nature* 464.7285, pp. 59–65. DOI: 10.1038/nature08821 (cit. on p. 3).

Quick, J., P. Ashton, S. Calus, C. Chatt, S. Gossain, J. Hawker, S. Nair, K. Neal, K. Nye, T. Peters, E. De Pinna, E. Robinson, K. Struthers, M. Webber, A. Catto, T. J. Dallman, P. Hawkey, and N. J. Loman (2015). "Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella". *Genome Biology* 16, p. 114. DOI: 10.1186/s13059-015-0677-2 (cit. on p. 4).

Reuter, S., M. J. Ellington, E. J. P. Cartwright, C. U. Köser, M. E. Török, T. Gouliouris, S. R. Harris, N. M. Brown, M. T. G. Holden, M. Quail, J. Parkhill, G. P. Smith, S. D. Bentley, and S. J. Peacock (2013). "Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology". *JAMA Internal Medicine* 173.15, pp. 1397–1404. DOI: 10.1001/jamainternmed.2013.7734 (cit. on p. 82).

Rho, M., H. Tang, and Y. Ye (2010). "FragGeneScan: predicting genes in short and error-prone reads". *Nucleic Acids Research* 38.20, e191–e191. DOI: 10.1093/nar/gkq747 (cit. on p. 8).

Rinke, C., P. Schwientek, A. Sczyrba, N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke (2013). "Insights into the phylogeny and coding potential of microbial dark matter". *Nature* 499.7459, pp. 431–7. DOI: 10.1038/nature12352 (cit. on pp. 8, 48, 58, 68).

Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light, D. Ekman, and D. A. Liberles (2007). "Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms". *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 308.1, pp. 58–73. DOI: 10.1002/jez.b.21124 (cit. on p. 9).

Rubin, E. M. (2008). "Genomics of cellulosic biofuels". *Nature* 454.7206, pp. 841–5. DOI: 10.1038/nature07190 (cit. on pp. 13, 22).

Rumbaugh, K. P. (2014). "Genomic complexity and plasticity ensure Pseudomonas success". *FEMS Microbiology Letters* 356.2, pp. 141–143. DOI: 10.1111/1574-6968.12517 (cit. on pp. 15, 78).

Ruschhaupt, M., W. Huber, A. Poustka, and U. Mansmann (2004). "A compendium to ensure computational reproducibility in high-dimensional classification tasks". *Statistical Applications in Genetics and Molecular Biology* 3, Article37. DOI: 10.2202/1544-6115.1078 (cit. on pp. 41, 74, 84).

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". *Journal of Molecular Biology* 94.3, pp. 441–448 (cit. on p. 4).

Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye (2012). "Database resources of the National Center for Biotechnology Information". *Nucleic Acids Research* 40.Database issue, pp. D13–25. DOI: 10.1093/nar/gkr1184 (cit. on p. 24).

Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork (2000). "SMART: a web-based tool for the study of genetically mobile domains". *Nucleic Acids Research* 28.1, pp. 231–4 (cit. on p. 35).

Sievert, D. M., P. Ricks, J. R. Edwards, A. Schneider, J. Patel, A. Srinivasan, A. Kallen, B. Limbago, S. Fridkin, and National Healthcare Safety Network (NHSN) Team and Participating NHSN Facilities (2013). "Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the National Healthcare Safety Network at the Centers for Disease Control and Prevention, 2009-2010". *Infection Control and Hospital Epidemiology* 34.1, pp. 1–14. DOI: 10.1086/668770 (cit. on p. 78).

Someya, S., M. Kakuta, M. Morita, K. Sumikoshi, W. Cao, Z. Ge, O. Hirose, S. Nakamura, T. Terada, and K. Shimizu (2010). "Prediction of carbohydrate-binding proteins from sequences using support vector machines". *Advances Bioinformatics*. DOI: 10.1155/2010/289301 (cit. on p. 24).

Sommer, M. O. A. and G. Dantas (2011). "Antibiotics and the resistant microbiome". *Current Opinion in Microbiology* 14.5, pp. 556–563. DOI: 10.1016/j.mib.2011.07.005 (cit. on p. 2).

Sonnhammer, E. L. L. and E. V. Koonin (2002). "Orthology, paralogy and proposed classification for paralog subtypes". *Trends in Genetics* 18.12, pp. 619–620 (cit. on p. 9).

Staley, J. T. and a. A. Konopka (1985). "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats". *An-*

nual Review of Microbiology 39.1, pp. 321–346. DOI: 10.1146/annurev.mi.39.100185.001541 (cit. on p. 5).

Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. L. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K.-S. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. W. Hancock, S. Lory, and M. V. Olson (2000). "Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen". *Nature* 406.6799, pp. 959–964. DOI: 10.1038/35023079 (cit. on p. 78).

Suen, G., P. J. Weimer, D. M. Stevenson, F. O. Aylward, J. Boyum, J. Deneke, C. Drinkwater, N. N. Ivanova, N. Mikhailova, O. Chertkov, L. A. Goodwin, C. R. Currie, D. Mead, and P. J. Brumm (2011). "The complete genome sequence of Fibrobacter succinogenes S85 reveals a cellulolytic and metabolic specialist". *PLoS ONE* 6.4, e18814. DOI: 10.1371/journal.pone.0018814 (cit. on p. 35).

Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin (2001). "The COG database: new developments in phylogenetic classification of proteins from complete genomes". *Nucleic Acids Research* 29.1, pp. 22–8. DOI: 10.1093/nar/29.1.22 (cit. on pp. 9, 65).

Taylor, L., B. Henrissat, P. Coutinho, N. Ekborg, S. Hutcheson, and R. Weiner (2006). "Complete cellulase system in the marine bacterium Saccharophagus degradans strain 2-40T." *Journal of Bacteriology* 188.11, pp. 3849–3861 (cit. on p. 36).

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest". *Nature* 444.7122, pp. 1027–131. DOI: 10.1038/nature05414 (cit. on p. 3).

U.S. National Oceanic and Atmospheric Administration (2004). *White flocculent mats in and around the extremely gassy, high-temperature (100°C, 212°F) white smokers at Champagne Vent.* URL: http://oceanexplorer.noaa.gov/

explorations/04fire/logs/hirez/champagne_vent_hirez.jpg (visited on 08/25/2016) (cit. on p. 2).

Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter (2007). "Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite". *Nature* 450.7169, pp. 560–5. DOI: 10.1038/nature06269 (cit. on p. 23).

Weimann, A., Y. Trukhina, P. B. Pope, S. G. Konietzny, and A. C. McHardy (2013). "De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes". *Biotechnology for Biofuels* 6.24. DOI: 10.1186/1754-6834-6-24 (cit. on pp. 15, 17, 48, 65, 68).

Weimann, A., K. Mooren, J. Frank, P. B. Pope, A. Bremges, and A. C. McHardy (2016a). "From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer". *mSystems* 1.6. Ed. by N. Segata. DOI: 10.1128/mSystems.00101-16 (cit. on pp. 17, 46).

— (2016b). "From genomes to phenotypes: Traitar, the microbial trait analyzer". *bioRxiv*, p. 043315. DOI: 10.1101/043315 (cit. on p. 17).

Weinberg, Z. (2012). *SVM separating hyperplanes*. URL: https://commons.wikimedia.org/wiki/File:Svm_separating_hyperplanes_(SVG).svg (visited on 08/25/2016) (cit. on p. 12).

Westbrock-Wadman, S., D. R. Sherman, M. J. Hickey, S. N. Coulter, Y. Q. Zhu, P. Warrener, L. Y. Nguyen, R. M. Shawar, K. R. Folger, and C. K. Stover (1999). "Characterization of a Pseudomonas aeruginosa efflux pump contributing to aminoglycoside impermeability". *Antimicrobial Agents and Chemotherapy* 43.12, pp. 2975–2983 (cit. on p. 81).

Wheals, A., L. Basso, D. Alves, and H. Amorim (1999). "Fuel ethanol after 25 years." *TIBTECH* 17, pp. 482–487 (cit. on p. 22).

Wilson, D. B. (2009). "Evidence for a novel mechanism of microbial cellulose degradation". *Cellulose* 16.4, pp. 723–727. DOI: 10.1007/s10570-009-9326-9 (cit. on pp. 36, 38, 39).

Wooley, J. C., A. Godzik, and I. Friedberg (2010). "A Primer on Metagenomics". *PLOS Computational Biology* 6.2, e1000667. DOI: 10.1371/journal.pcbi.1000667 (cit. on p. 6).

Wright, G. D. (2012). "Antibiotics: a new hope". *Chemistry & Biology* 19.1, pp. 3–10. DOI: 10.1016/j.chembiol.2011.10.019 (cit. on pp. 3, 15).

Xie, G., D. C. Bruce, J. F. Challacombe, O. Chertkov, J. C. Detter, P. Gilna, C. S. Han, S. Lucas, M. Misra, G. L. Myers, P. Richardson, R. Tapia, N. Thayer, L. S. Thompson, T. S. Brettin, B. Henrissat, D. B. Wilson, and M. J. McBride (2007). "Genome Sequence of the Cellulolytic Gliding Bacterium Cytophaga hutchinsonii". *Applied and Environmental Microbiology* 73.11, pp. 3536–3546. DOI: 10.1128/aem.00225-07 (cit. on p. 23).

Yamada, T., M. Kanehisa, and S. Goto (2006). "Extraction of phylogenetic network modules from the metabolic network". *BMC Bioinformatics* 7, p. 130. DOI: 10.1186/1471-2105-7-130 (cit. on p. 10).

Yaun, G.-X., K.-W. Chang, C.-J. Hsieh, and C.-J. Lin (2010). "A Comparison of Optimization methods for Large-scale L1-regularized Linear Classification". *Journal of Machine Learning Research* 11, pp. 3183–3234 (cit. on p. 41).

Yin, Y., X. Mao, J. Yang, X. Chen, F. Mao, and Y. Xu (2012). "dbCAN: a web resource for automated carbohydrate-active enzyme annotation". *Nucleic Acids Research* doi:10.1093/nar/gks479. DOI: 10.1093/nar/GKS479 (cit. on pp. 9, 40).

Yonezawa, M., M. Takahata, N. Matsubara, Y. Watanabe, and H. Narita (1995). "DNA gyrase gyrA mutations in quinolone-resistant clinical isolates of Pseudomonas aeruginosa." *Antimicrobial Agents and Chemotherapy* 39.9, pp. 1970–1972 (cit. on p. 80).

Yosef, N., J. Gramm, Q.-F. Wang, W. Noble, R. Karp, and R. Sharan (2010). "Prediction of phenotype information from genotype data." *Communications in information and systems* 10.2, pp. 99–114 (cit. on p. 24).

Zhang, M. L. and Z. H. Zhou (2014). "A Review on Multi-Label Learning Algorithms". *IEEE Transactions on Knowledge and Data Engineering* 26.8, pp. 1819–1837. DOI: 10.1109/TKDE.2013.39 (cit. on p. 88).

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series A* 67.2, pp. 301–320 (cit. on p. 70).

APPENDIX A

Journal versions of the published articles

Weimann et al. Biotechnology for Biofuels 2013, 6:24
http://www.biotechnologyforbiofuels.com/content/6/1/24

Biotechnology for Biofuels

RESEARCH Open Access

De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes

Aaron Weimann[1,3†], Yulia Trukhina[1,3†], Phillip B Pope[2], Sebastian GA Konietzny[1,3] and Alice C McHardy[1,3*]

Abstract

Background: Understanding the biological mechanisms used by microorganisms for plant biomass degradation is of considerable biotechnological interest. Despite of the growing number of sequenced (meta)genomes of plant biomass-degrading microbes, there is currently no technique for the systematic determination of the genomic components of this process from these data.

Results: We describe a computational method for the discovery of the protein domains and CAZy families involved in microbial plant biomass degradation. Our method furthermore accurately predicts the capability to degrade plant biomass for microbial species from their genome sequences. Application to a large, manually curated data set of microbial degraders and non-degraders identified gene families of enzymes known by physiological and biochemical tests to be implicated in cellulose degradation, such as GH5 and GH6. Additionally, genes of enzymes that degrade other plant polysaccharides, such as hemicellulose, pectins and oligosaccharides, were found, as well as gene families which have not previously been related to the process. For draft genomes reconstructed from a cow rumen metagenome our method predicted Bacteroidetes-affiliated species and a relative to a known plant biomass degrader to be plant biomass degraders. This was supported by the presence of genes encoding enzymatically active glycoside hydrolases in these genomes.

Conclusions: Our results show the potential of the method for generating novel insights into microbial plant biomass degradation from (meta-)genome data, where there is an increasing production of genome assemblages for uncultured microbes.

Background

Lignocellulosic biomass is the primary component of all plants and one of the most abundant organic compounds on earth. It is a renewable, geographically distributed and a source of sugars, which can subsequently be converted into biofuels with low greenhouse gas emissions, such as ethanol. Chemically, it primarily consists of cellulose, hemicellulose and lignin. Saccharification - the process of degrading lignocellulose into the individual component sugars - is of considerable biotechnological interest. Several mechanical and chemical procedures for saccharification have been established; however, all are relatively expensive, slow and inefficient [1]. An alternative approach is realized in nature by various microorganisms, which use enzyme-driven lignocellulose degradation to generate sugars as sources of carbon and energy. The search for novel enzymes allowing an efficient breakdown of plant biomass has therefore attracted considerable interest [2-5]. In particular, the discovery of novel cellulases for saccharification is considered crucial in this context [6]. However, the complexity of the underlying biological mechanisms and the lack of robust enzymes that can be economically produced in larger quantities currently still prevent industrial application.

For some lignocellulose-degrading species, carbohydrate-active enzymes (CAZymes) and protein domains implicated in lignocellulose degradation are well known. Many of

* Correspondence: alice.mchardy@uni-duesseldorf.de
†Equal contributors
[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany
[3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
Full list of author information is available at the end of the article

these have been recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as the enzymes of the glycoside hydrolase (GH) families GH6 and GH9 and the endoglucanase-containing family GH5. Two well-studied paradigms are currently known for microbial cellulose degradation: The 'free-enzyme system' is realized in most aerobic microbes and entails secretion of a set of cellulases to the outside of the cell. In anaerobic microorganisms large multi-enzyme complexes, known as cellulosomes, are assembled on the cell surface and catalyze degradation. In both cases, the complete hydrolysis of cellulose requires endoglucanases (GH5 and GH9), which are believed to target non-crystalline regions, and exo-acting cellobiohydrolases, which attack crystalline structures from either the reducing (GH7 and GH48) or non-reducing (GH6) end of the beta-glucan chain. However, in the genomes of some plant biomass-degrading species, homologs of such enzymes have not been found. Recent genome analyses of the lignocellulose-degrading microorganisms, such as the aerobe *Cytophaga hutchinsonii* [7], the anaerobe *Fibrobacter succinogenes* [8,9] and the extreme thermophile anaerobe *Dictyoglomus turgidum* [10] have revealed only GH5 and GH9 endoglucanases. Genes encoding exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures (dockerins and cohesins) are absent.

Metagenomics offers the possibility of studying the genetic material of difficult-to-culture (i.e. uncultured) species within microbial communities with the capability to degrade plant biomass. Recent metagenome studies of the gut microbiomes of the wood-degrading higher termites (*Nasutitermes*), the Australian Tammar wallaby (*Macropus eugenii*) [11,12] and two studies of the cow rumen metagenome [13,14] have revealed new insights into the mechanisms of cellulose degradation in uncultured organisms and microbial communities. Microbial communities of different herbivores have been shown to be dominated by lineages affiliated to the Bacteroidetes and Firmicutes, of which different Bacteroidetes lineages exhibited endoglucanse activity [11,15]. Notably, exo-acting families and cellulosomal structures have a low representation or are entirely absent from gut metagenomes sequenced to date. Thus, current knowledge about genes and pathways involved in plant biomass degradation in different species, particularly uncultured microbial ones, is still incomplete.

We describe a method for the *de novo* discovery of protein domains and CAZy families associated with microbial plant biomass degradation from genome and metagenome sequences. It uses protein domain and gene family annotations as input and identifies those domains or gene families, which in concert are most distinctive for the lignocellulose degraders. Among the gene and protein domains identified with our method were known key genes of plant biomass degradation. Additionally, it identified several novel protein domains and gene families as being relevant for the process. These might represent novel leads towards elucidating the mechanisms of plant biomass degradation for the currently less well understood microbial species. Our method furthermore can be used to identify plant biomass-degrading species from the genomes of cultured or uncultured microbes. Application to draft genomes assembled from the metagenome of a switchgrass-adherent microbial community in cow rumen predicted genomes from several Bacteroidales lineages which encode active glycoside hydrolases and a relative to a known plant biomass degrader to represent lignocellulose degraders.

In technical terms, our method selects the most informative features from an ensemble of L1-regularized L2-loss linear Support Vector Machine (SVM) classifiers, trained to distinguish genomes of cellulose-degrading species from non-degrading species based on protein family content. Protein domain annotations are available in public databases and new protein sequences can be rapidly annotated with Hidden Markov Models (HMMs) or - somewhat slower - with BLAST searches of one protein versus the NCBI-nr database [16]. Co-occurrence of protein families in the biomass-degrading fraction of samples and an absence of these families within the non-degrading fraction allows the classifier to link these proteins to biomass degradation *without* requiring sequence homology to known proteins involved in lignocellulose degradation. Classification with SVMs has been previously used successfully for phenotype prediction from genetic variations in genomic data. In Beerenwinkel *et al.* [17], support vector regression models were used for predicting phenotypic drug resistance from genotypes. SVM classification was used by Yosef *et al.* [18] for predicting plasma lipid levels in baboons based on single nucleotide polymorphism data. In Someya *et al.* [19], SVMs were used to predict carbohydrate-binding proteins from amino acid sequences. The SVM [20,21] is a discriminative learning method that infers, in a supervised fashion, the relationship between input features (such as the distribution of conserved gene clusters or single nucleotide polymorphisms across a set of sequence samples) and a target variable, such as a certain phenotype, from labeled training data. The inferred function is subsequently used to predict the value of this target variable for new data points. This type of method makes no *a priori* assumptions about the problem domain. SVMs can be applied to datasets with millions of input features and have good generalization abilities, in that models inferred from small amounts of training data show good predictive accuracy on novel data. The use of models that include an L1-regularization term favors solutions in which few

features are required for accurate prediction. There are several reasons why sparseness is desirable: the high dimensionality of many real datasets results in great challenges for processing. Many features in these datasets are usually non-informative or noisy, and a sparse classifier can lead to a faster prediction. In some applications, like ours, a small set of relevant features is desirable because it allows direct interpretation of the results.

Results

We trained an ensemble of SVM classifiers to distinguish between plant biomass-degrading and non-degrading microorganisms based on either Pfam domain or CAZY gene family annotations (see Methods section for the training and evaluation of the SVM classification ensemble). We used a manually curated data set of 104 microbial (meta-)genome sequence samples for this purpose, which included 19 genomes and 3 metagenomes of lignocellulose degraders and 82 genomes of non-degraders (Figure 1, Figure 2, Additional file 1: Table S1). Fungi are known to use several enzymes for plant biomass degradation for which the corresponding genes are not found in prokaryotic genomes and vice versa, while other genes are shared by prokaryotic and eukaryotic degraders. To investigate similarities and differences detectable with our method, we included the genome of lignocellulose degrading fungus *Postia placenta* into our analysis. After training, we identified the most distinctive protein domains and CAZy families of plant biomass degraders from the resulting models. We compared these protein domains and gene families with known plant biomass degradation genes. We furthermore applied our method to identify plant biomass degraders among 15 draft genomes from the metagenome of a microbial community adherent to switch grass in cow rumen.

Distinctive Pfam domains of microbial plant biomass degraders

For the training of a classifier which distinguishes between plant biomass-degrading and non-degrading microorganisms we used Pfam annotations of 101 microbial genomes and two metagenomes. This included metagenomes of microbial communities from the gut of a wood-degrading higher termite and from the foregut of the Australian Tammar Wallaby as examples for plant biomass-degrading communities. Furthermore, 19 genomes of microbial lignocellulose degraders were included - of the phyla Firmicutes (7 isolate genome sequences), Actinobacteria (5), Proteobacteria (3), Bacteroidetes (1), Fibrobacteres (1), Dictyoglomi (1) and Basidiomycota (1). Eighty-two microbial genomes annotated to not possess the capability to degrade lignocellulose were used as examples of non-lignocellulose-degrading microbial species (Additional file 1: Table S1).

We assessed the value of information about the presence or absence of protein domains for distinguishing lignocellulose degraders from non-degraders. With the respective classifier, $eSVM_{bPFAM}$, each microbial (meta-) genome sequence was represented by a feature vector with the features indicating the presence or absence of Pfam domains (see Methods). The nested cross-validation macro-accuracy of $eSVM_{bPFAM}$ in distinguishing plant biomass-degrading from non-degrading microorganisms was 0.91. This corresponds to 94% (97 of 103) of the (meta-)genome sequences being classified correctly. Only three of the 21 cellulose-degrading samples and three of the non-degraders were misclassified (Table 1, Table 2). Among these were four Actinobacteria and one genome affiliated with the Basidiomycota and Theromotogae each.

We identified the Pfam domains with the greatest importance for assignment to the lignocellulose-degrading class by $eSVM_{bPFAM}$ (Figure 1; see Methods for the feature selection algorithm). Among these are several protein domains known to be relevant for plant biomass degradation. One of them is the GH5 family, which is present in all of the plant biomass-degrading samples. Almost all activities determined within this family are relevant to plant biomass degradation. Because of its functional diversity, a subfamily classification of the GH5 family was recently proposed [24]. The carbohydrate-binding modules CBM_6 and CBM_4_9 were also selected. Both families are Type B carbohydrate-binding modules (CBMs), which exhibit a wide range of specificities, recognizing single glycan chains comprising hemicellulose (xylans, mannans, galactans and glucans of mixed linkages) and/or non-crystalline cellulose [25]. Type A CBMs (e.g. CBM2 and CBM3), which are more commonly associated with binding to insoluble, highly crystalline cellulose, were not identified as relevant by $eSVM_{bPFAM}$. Furthermore, numerous enzymes that degrade non-cellulosic plant structural polysaccharides were identified, including those that attack the backbone and side chains of hemicellulosic polysaccharides. Examples include the GH10 xylanases and GH26 mannanases. Additionally, enzymes that generally display specificity for oligosaccharides were selected, including GH39 β-xylosidases and GH3 enzymes.

We subsequently trained a classifier - $eSVM_{fPFAM}$ - with a weighted representation of Pfam domain frequencies for the same data set. The macro-accuracy of $eSVM_{fPFAM}$ was 0.84 (Table 2); lower than that of the $eSVM_{bPFAM}$; with nine misclassified samples (4 Actinobacteria, 2 Bacteroidetes, 1 Basidiomycota, 1 Thermotogae phyla and the Tammar Wallaby metagenome). Again, we determined the most relevant protein domains for identifying a plant biomass-degrading sequence sample from the models by

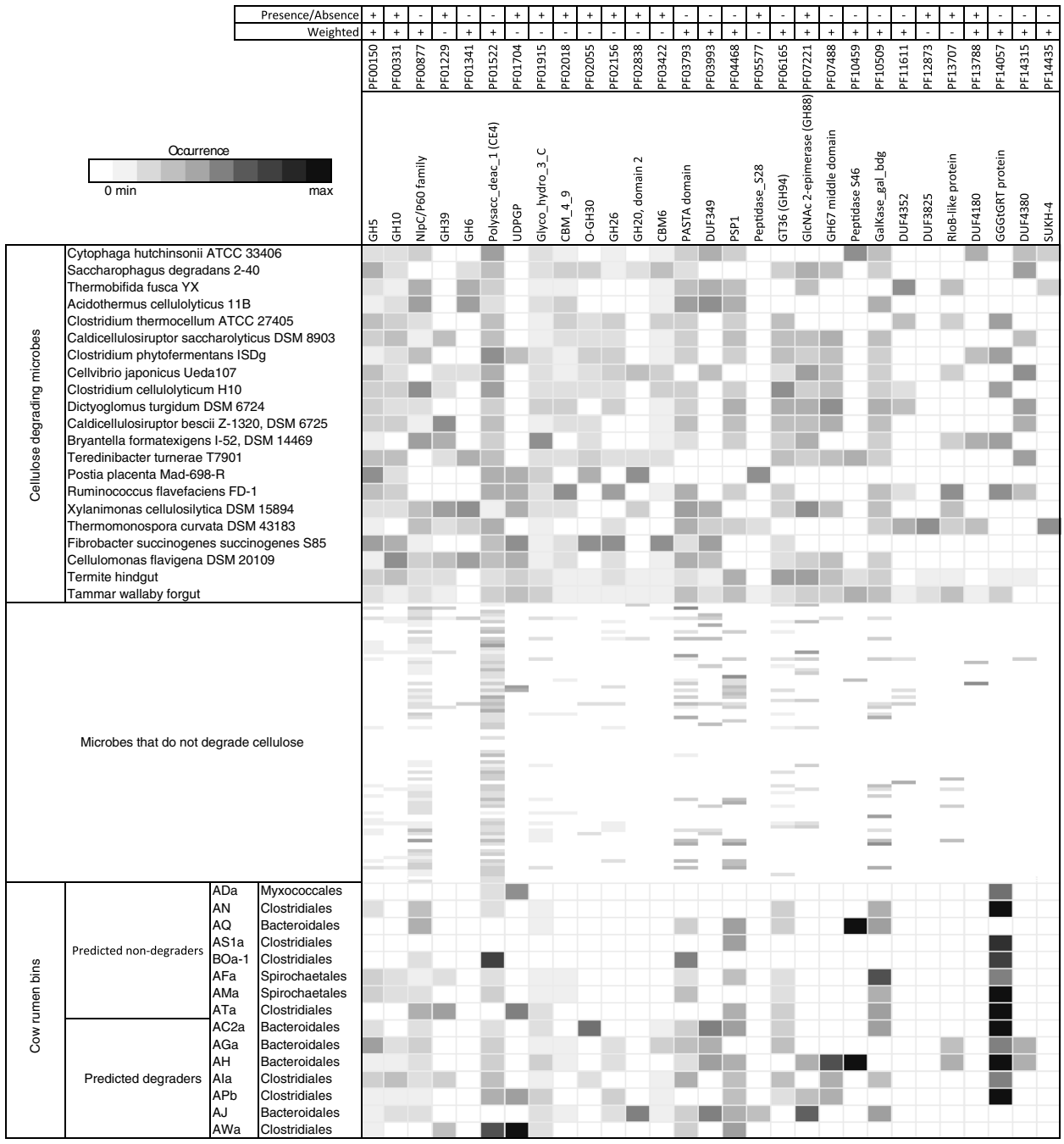


Figure 1 Frequencies of the selected Pfam families in the individual genomes and metagenomes. The data for each entry are rescaled by the total number of Pfam domains annotated to the microbial genome or metagenome. The color scale from grey to black indicates domain families that are present in low to high amounts, respectively. White indicates absent protein domains. The signs "+" and "-" indicate whether a protein domain was chosen in the respective experiment.

feature selection. Among the most important protein families were, as before, GH5, GH10 and GH88 (PF07221: N-acylglucosamine 2-epimerase) (Figure 1). GH6, GH67 and CE4 acetyl xylan esterases ("accessory enzymes" that contribute towards complete hydrolysis of xylan) were only relevant for prediction with the $eSVM_{fPFAM}$ classifier. Additionally, both models specified protein domains not commonly associated with plant biomass degradation as being relevant for assignment, such as the lipoproteins DUF4352 and PF00877 (NlpC/P60 family) and binding domains PF10509 (galactose-binding signature domain) and PF03793 (PASTA domain) (Figure 1).

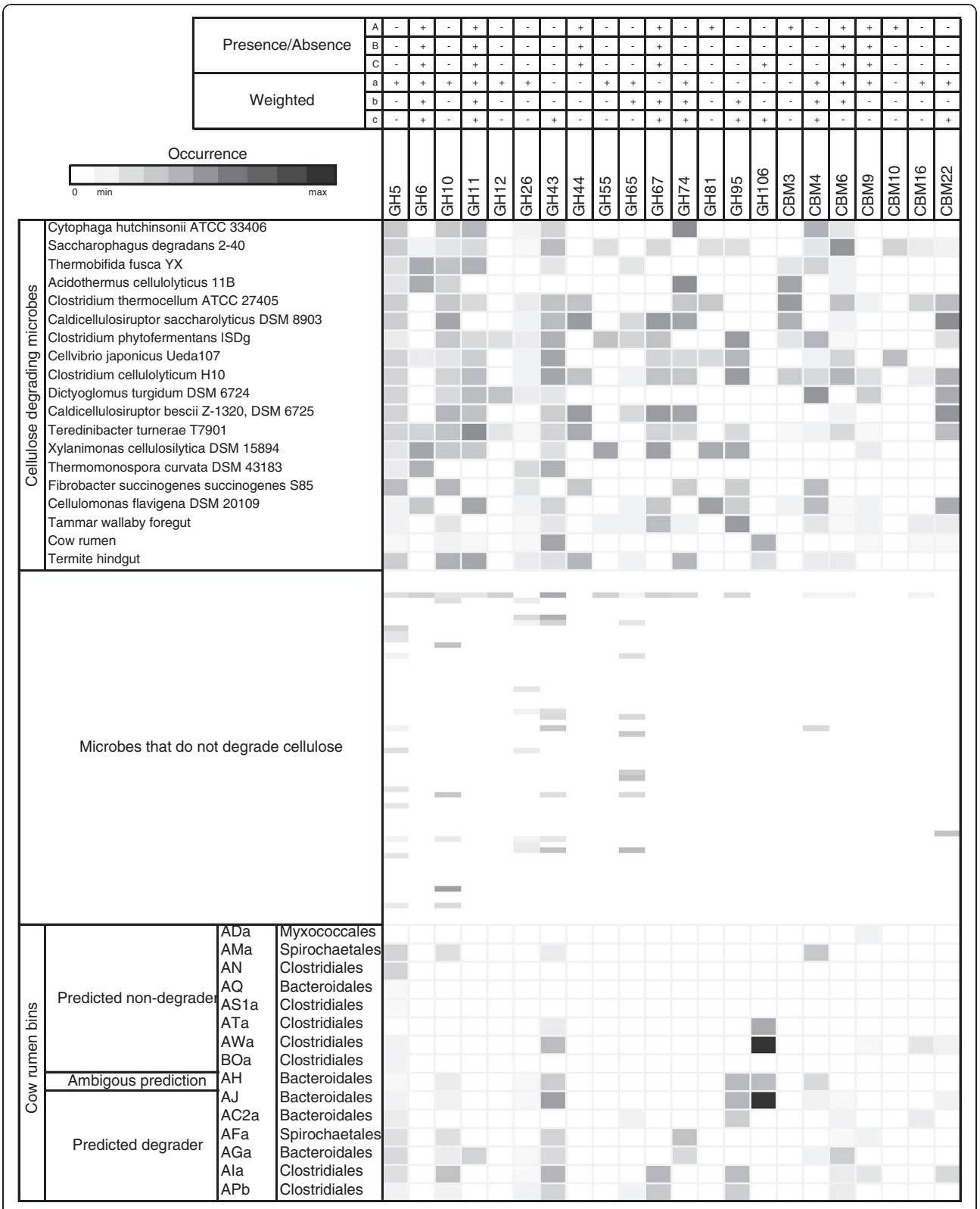


Figure 2 Frequencies of selected glycoside hydrolase (GH) families and carbohydrate binding modules (CBMs) in the (meta-) genome sequences. The data for each entry are rescaled by the total number of GH and CBM domains annotated to the microbial genome or metagenome. The coloring from black to grey indicates domains that are present in high to low amounts, respectively. White indicates absent domain families ("A", "a", "B", "b", "C", "c" as described in Table 1).

Table 1 Misclassified species in the SVM analyses

	$eSVM_{bPFAM}$	$eSVM_{CAZY\_B}$
False negatives	*Postia placenta Mad-698-R*	*Thermomonospora curvata DSM 43183*
	Xylanimonas cellulosilytica DSM 15894	
	Thermomonospora curvata DSM 43183	
False positives	*Actinosynnema mirum 101*	*Actinosynnema mirum 101*
	Arthrobacter aurescens TC1	
	Thermotoga lettingae TMO	

Shown are species which were misclassified with the $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers. Contrary to previous beliefs [22], recent literature indicates in agreement with our predictions that *T. curvata* is a non-degrader. Furthermore, recent evidence supports that *A. mirum* is a lignocellulose degrader, which has not been previously described [23].

Distinctive CAZy families of microbial plant biomass degraders

We searched for distinctive CAZy families of microbial plant biomass degraders with our method. CAZy families include glycoside hydrolases (GH), carbohydrate-binding modules (CBM), glycosyltransferases (GT), polysaccharide lyases (PL) and carbohydrate esterases (CE). The annotations from the CAZy database comprised 64 genomes of non-lignocellulose-degrading species and 16 genomes of lignocellulose-degraders. There were no CAZy annotations available for the remaining genomes. In addition, we included the metagenomes of the gut microbiomes of the Tammar wallaby (TW), the wood-degrading higher termite and of the cow rumen microbiome (Additional file 1: Table S1). We evaluated the value of information about the presence or absence of CAZy domains, or of their relative frequencies for identification of lignocellulose-degrading microbial (meta-)genomes in the following experiments:

1) By training of the classifiers $eSVM_{CAZY\_A}$ (presence/absence) and $eSVM_{CAZY\_a}$ (counts), based on genome annotations with all CAZy families.
2) By training of the classifiers $eSVM_{CAZY\_B}$ (presence/absence) and $eSVM_{CAZY\_b}$ (counts), based on the annotations of the genomes and the TW sample with all CAZy families, except for the GT family members, which were not annotated for the TW sample.
3) By training of the classifiers $eSVM_{CAZY\_C}$ (presence/absence) and $eSVM_{CAZY\_c}$ (counts) with the entire data set based on GH family and CBM annotations, as these were the only ones available for the three metagenomes.

The macro-accuracy of these classifiers ranged from 0.87 to 0.96, similar to the Pfam-domain-based models (Table 2). Notably, almost exclusively Actinobacteria were misclassified by the $eSVM_{CAZY}$ classifiers, except for the Firmicute *Caldicellulosiruptor saccharolyticus*. The best classification results were obtained with the presence-absence information for all CAZy families except for the GT families of the microbial genomes and the TW sample. In this setting ($eSVM_{CAZY\_B}$) only two species (*Thermomonospora curvata* and *Actinosynnema mirum*) were misclassified (Table 1). These species remained misclassified with all six classifiers.

Using feature selection, we determined the CAZy families from the six $eSVM_{CAZy}$ classifiers that are most relevant for identifying microbial cellulose-degraders. Many of these GH families and CBMs are present in all (meta-) genomes (Figure 2). This analysis identified further gene families known to be relevant for plant biomass degradation. Among them are cellulase-containing families (GH5, GH6, GH12, GH44, GH74), hemicellulase-containing families (GH10, GH11, GH26, GH55, GH81, GH115), families with known oligosaccharide/side-chain-degrading activities (GH43, GH65, GH67, GH95) and several CBMs (CBM3, -4, -6, -9, -10, -16, -22, -56). Several of these (GH6, GH11, GH44, GH67, GH74, CBM4, CBM6, CBM9) were consistently identified by at least half of the six classifiers as distinctive for plant biomass degraders. These might be considered signature genes of the plant biomass-degrading microorganisms we analyzed. Additionally, several GT, PL and CE domains were identified as relevant ($eSVM_{CAZY\_A}$: PL1, PL11 and CE5, "$eSVM_{CAZY\_B}$: CE5; $eSVM_{CAZY\_a}$: GT39, PL1 and CE2, $eSVM_{CAZY\_b}$: none). These CAZy families, as well as GH115 and CBM56, are not included in Figure 2, as they are not annotated for all sequences.

Identification of plant biomass degraders from a cow rumen metagenome

We used our method to predict the plant biomass-degrading capabilities for 15 draft genomes of uncultured microbes reconstructed from the metagenome of a microbial community adherent to switchgrass in cow rumen [14] (see Methods for the classification with an ensemble of SVM classifiers). The draft genomes represent genomes with more than 50% of the sequence reconstructed by taxonomic binning of the metagenome

Table 2 Accuracy of classifying microbes as lignocellulose-degraders or non-degraders

	Presence/absence of Pfam domains	Weighted Pfam domain representation	Presence/absence CAZy family representation			Weighted CAZy family representation		
			A	B	C	a	b	c
nCV macro-accuracy	0.91	0.84	0.90	0.96	0.94	0.91	0.93	0.87
nCV recall	0.86	0.73	0.81	0.94	0.90	0.88	0.88	0.79
nCV true negative rate	0.96	0.96	0.98	0.98	0.98	0.95	0.98	0.95

L1-regularized SVMs were trained with Pfam domain or CAZY family (meta-)genome annotations. Capital letters denote classifiers trained based on the presence or absence of CAZy families and small letters indicate classifiers trained based on the relative abundances of CAZy families in annotations. Abbreviations "A", "a"," B", "b", "C", "c" denote the following: Classifiers "A","a" were trained with annotations of all CAZy families for 16 microbial genomes; Classifiers "B","b" were trained with annotations for all CAZy families, except for the GT family members (which were not annotated for the Tammar Wallaby metagenome), for 16 genomes and the TW metagenome of plant biomass degraders; Classifiers "C","c" were trained with annotations for the GH families and CBMs for the 16 microbial genomes and three metagenomes of plant biomass degraders, as only these were annotated for the metagenomes. All CAZy-based classifiers were trained with available annotations for 64 genomes of non-biomass degraders. The Pfam-based classifiers were trained with 21 (meta-)genomes of biomass-degraders and 82 microbial genomes of non-degraders. For more details on the experimental set-up and the evaluation measures shown see the Methods section on performance evaluation.

sample. The microbial community adherent to switchgrass is likely to be enriched with plant biomass degraders, as it was found to differ from the rumen fluid community in its taxonomic composition and degradation of switch grass after incubation in cow rumen had occurred. For identification of plant biomass-degrading microbes, we classified each draft genome individually with the $eSVM_{bPFAM}$ and $eSVM_{CAZY\_B}$ models, which had the highest macro-accuracy based on Pfam domain or CAZy family annotations, respectively. The $eSVM_{bPFAM}$ classifier assigned seven of the draft genomes to plant biomass degraders (Table 3). One of these, genome *APb*, was found by 16S rRNA analysis to be related to the fibrolytic species *Butyrivibrio fibrisolvens*. Four others (*AC2a*, *AGa*, *AJ* and *AH*) are of the order of Bacteroidales, and include all but one draft genomes affiliated to the Bacteroidales. The 6th and 7th predicted degrader, represented by genome *AIa* and *AWa*, belong to the Clostridiales, like genome *APb*. The $eSVM_{CAZY\_B}$ classifier also assigned five of these genomes to the plant biomass degraders. Additionally it classified genome *AH* as plant biomass-degrading, while being ambiguous in the assignment of *AFa* (Table 3). To validate these predictions, we searched the draft genomes for genes encoding 51 enzymatically active glycoside hydrolases characterized from the same rumen dataset (for the results of these experiments see Figure three in Hess *et al.* [14]). Genomes *AGa*, *AC2a*, *AJ* and *AIa* were all linked to different enzymes of varying specificities (Table 3). *AC2a* was linked to cellulose degradation, specifically to a carboxymethyl cellulose (CMC)-degrading GH5 endoglucanase as well as GH9 enzyme capable of degrading insoluble cellulosic substrates such as Avicel®. *AIa* demonstrated capabilities towards xylan and soluble cellulosic substrates with affiliations to four GH10 xylanases. Both *AGa* and *AJ* demonstrated broader substrate versatility and were linked to enzymes with capabilities towards cellulosic substrates CMC and Avicel® (GH5, GH9 and GH26), hemicellulosic substrates lichenan (β-1,3, β-1,4 β-glucan) and xylan (GH5, GH9 and GH10), as well as the natural feedstocks miscanthus and switchgrass (GH5 and GH9). Importantly, no carbohydrate-active enzymes were affiliated to draft genomes that were predicted to not possess plant biomass-degrading capabilities (Table 3). Overall, assignments were largely consistent between the two classifiers and supporting evidence for the capability to degrade plant biomass was found for five of the predicted degraders.

Timing experiments

Our method uses annotations with Pfam domains or CAZy families as input. Generating these by similarity-searches with profile HMMs rather than with BLAST provides a better scalability for next-generation sequencing data sets. HMM databases such as dbCAN contain a representation of entire protein families rather than of individual gene family members, which largely decreases the number of entries one has to compare against. For example, searching the ORFs of the *Fibrobacter succinogenes* genome [26] for similarities to CAZy families with the dbCAN HMM models took 23 seconds on an Intel® Xeon® 1.6 GHz CPU. In comparison, searching for similarities to CAZy families by BLASTing the same set of ORFs against all sequences with CAZy family annotation of the NCBI non-redundant protein database (downloaded from http://csbl.bmb.uga.edu/dbCAN/ on April 19th 2011) on the same machine required approximately 1 hour and 55 minutes, a difference of two orders of magnitude. Because of their better scalability and also because they are well-established for identifying protein domains or gene families [27-29], we recommend the use of HMM-based similarities and annotations as input to our method.

Discussion

We investigated the value of information about the presence-or-absence of CAZy families and Pfam protein domains, as well as information about their relative abundances, for the identification of lignocellulose degraders. Classifiers trained with CAZy family or Pfam

Table 3 Prediction of the plant biomass degradation capabilities for 15 draft genomes

	AC2a	AGa	AIa-2	AJ	APb	AFa	AH	AWa	ADa	AMa	AN	AQ	AS1	ATa	BOa
$eSVM_{CAZY\_B}$	++	++	++	+	++	++	0	- -	- -	- -	- -	- -	- -	- -	- -
$eSVM_{bPFAM}$	++	++	++	++	++	-	++	+	- -	-	- -	- -	- -	-	- -
CMC	GH5 (TW-33)	GH5 (TW-40) GH5 (MH-2)	GH10 (TW-34)	GH5 (TW-39) GH26 (TW-10) GH10 (TW-8)											
XYL		GH10 (TW-25)	GH10 (TW-30) GH10 (TW-31) GH10 (TW-37)	GH10 (TW-8)											
SWG		GH5 (TW-40) GH5 (MH-2)													
MIS	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
AVI	GH9 (TW-64)	GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											
LIC		GH5 (TW-40) GH5 (MH-2) GH9 (TW-50)		GH5 (TW-39)											

Genome reconstructions from the metagenome of a microbial community adherent to switchgrass in the cow rumen were obtained by taxonomic binning of assembled sequences in the original study. Symbols depict the prediction outcome of a voting committee of the 5 $eSVM_{CAZY\_B}$ and the $eSVM_{bPFAM}$ classifiers with the best macro-accuracy (see text for the description of the classifiers). ++: genome classified as plant biomass degrader by all classifiers; +: genome classified as plant biomass degrader by 4 out of 5 classifiers; 0: ambiguous prediction; -: genome classified as not plant biomass degrader by 4 out of 5 classifiers; --: genome classified as not plant biomass degrader by all classifiers. For every draft genome, the presence of genes encoding glycoside hydrolases with verified enzymatic activity for different substrates in this study [14] is indicated. The genome and substrate names correspond to those of Figure 3 and Table S6 of the study.
Hydrolytic activity detected on:
(CMC) 1% (w/v) carboxymethyl cellulose agar.
(XYL) 1% (w/v) Xylan.
(SWG) 1% (w/v) IL-Switchgrass.
(MIS) 1% (w/v) IL-Miscanthus.
(AVI) 1% (w/v) IL-Avicel.

domain annotations allowed an accurate identification of plant biomass degraders and determined similar domains and CAZy families as being most distinctive. Many of these are recognized by physiological and biochemical tests as being relevant for the biochemical process of cellulose degradation itself, such as GH6, members of the GH5 family and to a lesser extent GH44 and GH74. In contrast to widely accepted paradigms for microbial cellulose degradation, recent genome analysis of cellulolytic bacteria has identified examples (i.e. *Fibrobacter*) where there is an absence of genes encoding exo-acting cellobiohydrolases (GH6 and GH48) and cellulosome structures [30]. In addition, these exo-acting families and cellulosomal structures have had a low representation or are entirely absent from sequenced gut metagenomes. Our method also finds the exo-acting cellobiohydrolases GH7 and GH48 to be less important. GH7 represents fungal enzymes, so its absence makes sense; however, the lower importance assigned to GH48 is interesting. The role of GH48 is believed to be of high importance, although recent research has raised questions. Olson *et al.* [31] have found that a complete solubilization of crystalline cellulose can occur in *Clostridium thermocellum* without the expression of GH48, albeit at significantly lower rates. Furthermore, genome analysis of cellulose-degrading microbes *Cellvibrio japonicus* [32] and *Saccharophagus degradans* [33] have determined the presence of only non-reducing end enzymes (GH6) and an absence of a reducing end cellobiohydrolase (GH48), suggesting that the latter are not essential for all cellulolytic enzyme systems.

While we have focused on cellulose degradation, our method has also identified enzymes that degrade other plant polysaccharides as being relevant, such as hemicellulose (GH10, GH11, GH12, GH26, GH55, GH81, CE4), pectins (PL1, GH88 and GH43), oligosaccharides (GH3, GH30, GH39, GH43, GH65, GH95) and the side-chains attached to noncellulosic polysaccharides (GH67, GH88, GH106). This was expected, since many cellulose-degrading microbes produce a repertoire of different glycoside hydrolases, lyases and esterases (see, for example, [32,33]) that target the numerous linkages that are present within different plant polysaccharides, which often exist in tight cross-linked forms within the plant cell wall. The results from our method add further weight to this. The observation of numerous CBMs being relevant in the CAZy analysis also agrees with previous findings that many different CBM-GH combinations are possible in bacteria. Moreover, recent reports have demonstrated that the targeting actions of CBMs have strong proximity effects within cell wall structures, i.e. CBMs directed to a cell wall polysaccharide (e.g. cellulose) other than the target substrate of their appended glycoside hydrolase (e.g. xylanase) can promote enzyme action against the target substrate (e.g. xylan) within the cell wall [34]. This provides explanations as to why cellulose-directed CBMs are appended to many non-cellulase cell wall hydrolases.

Several Pfam domains of unknown function (DUFs) or protein domains which have not previously been associated with cellulose degradation are predicted as being relevant. These include transferases (PF01704) and several putative lipoproteins (DUF4352), some of which have predicted binding properties (NlpC/P60 family: PF00877, PASTA domain: PF03793). The functions of these domains in relation to cellulose degradation are not known, but possibilities include binding to cellulose, binding to other components of the cellulolytic machinery or interaction with the cell surface.

Another result of our study are the classifiers for identifying microbial lignocellulose-degraders from genomes of cultured and uncultured microbial species reconstructed from metagenomes. Classification of draft genomes reconstructed from switchgrass-adherent microbes from cow rumen with the most accurate classifiers predicted six or seven of these to represent plant biomass-degrading microbes, including a close relative to the fibrolytic species *Butyrivibrio fibrisolvens*. Cross-referencing of all draft genomes against a catalogue of enzymatically active glycoside hydrolases provided a degree of method validation and was in majority agreement with our predictions. Four genomes (*AGa*, *AC2a*, *AJ* and *AIa*) predicted positive were linked to cellulolytic and/or hemicellulolytic enzymes, and importantly no genomes that were predicted negative were linked to carbohydrate-active enzymes from that catalogue of enzymatically active enzymes. Also, no connections to carbohydrate-active enzymes from that catalogue were observed for the three genomes (*AFa*,*AH* and *AWa*) where ambiguous predictions were made. As both draft genomes as well as the catalogue of carbohydrate active enzymes in cow rumen are incomplete, in addition to our training data not covering all plant-biomass-degrading taxa, such ambiguous assignments might be better resolvable with more information in the future.

We trained a previous version of our classifier with the genome of *Methanosarcina barkeri fusaro* incorrectly labeled as a plant biomass degrader, according to information provided by IMG. In cross-validation experiments, our method correctly assigned *M. barkeri* to be a non-plant biomass-degrading species. We labeled *Thermonospora curvata* as a plant biomass degrader and *Actinosynnema mirum* as non-degrader according to information from the literature (see Additional file 1: Table S1). Both were misassigned by all classifiers in the cross-validation experiments. However, in a recent work by Anderson *et al.* [23] it was shown that in cellulose activity assays *A. mirum* could degrade various cellulose substrates. In the same study, *T. curvata* did not show cellulolytic activity against

any of these substrates, contrary to previous beliefs [22]. The authors found out that the cellulolytic *T. curvata* strain was in fact a *T. fusca* strain. Thus, our method could correctly assign both strains despite of the incorrect phenotypic labeling. The genome of *Postia placenta*, the only fungal plant biomass degrader of our data set was misassigned in the Pfam-based SVM analyses. Fungi possess cellulases not found in prokaryotic species [35] and might employ a different mechanism for plant biomass degradation [36,37]. Indeed, in our data set, *Postia placenta* is annotated with the cellulase-containing GH5 family and xylanase GH10, but the hemicellulase family GH26 does not occur. Furthermore, the (hemi-)cellulose binding CBM domains CBM6 and CBM_4_9, which were identified as being relevant for assignment to lignocellulose degraders with the $eSVM_{bPFAM}$ classifier, are absent. All of the latter ones, GH26, CBM6 and especially CBM4 and CBM9, occur very rarely in eukaryotic genome annotations, according to the CAZy database.

Conclusions

We have developed a computational technique for the identification of Pfam protein domains and CAZy families that are distinctive for microbial plant biomass degradation from (meta-)genome sequences and for predicting whether a (draft) genome of cultured or uncultured microorganisms encodes a plant biomass-degrading organism. Our method is based on feature selection from an ensemble of linear L1-regularized SVMs. It is sufficiently accurate to detect errors in phenotype assignments of microbial genomes. However, some microbial species remained misclassified in our analysis, which indicates that further distinctive genes and pathways for plant biomass degradation are currently poorly represented in the data and could therefore not be identified.

To identify a lignocellulose degrader from the currently available data, the presence of a few domains, many of which are already known, is sufficient. The identification of several protein domains which have so far not been associated with microbial plant biomass degradation in the Pfam-based SVM analyses as being relevant may warrant further scrutiny. A difficulty in our study was to generate a sufficiently large and correctly annotated dataset to reach reliable conclusions. This means that the results could probably be further improved in the future, as more sequences and information on plant biomass degraders become available. The method will probably also be suitable for identifying relevant gene and protein families of other phenotypes.

The prediction and subsequent validation of three Bacteroidales genomes to represent cellulose-degrading species demonstrates the value of our technique for the identification of plant biomass degraders from draft genomes from complex microbial communities, where there is an increasing production of genome assemblages for uncultured microbes. These to our knowledge represent the first cellulolytic Bacteroidetes-affiliated lineages described from herbivore gut environments. This finding has the potential to influence future cellulolytic activity investigations within rumen microbiomes, which has for the greater part been attributed to the metabolic capabilities of species affiliated to the bacterial phyla Firmicutes and Fibrobacteres.

Methods

Annotation

We annotated all protein coding sequences of microbial genomes and metagenomes with Pfam protein domains (Pfam-A 26.0) and Carbohydrate-Active Enzymes (CAZymes) [28,38]. The CAZy database contains information on families of structurally related catalytic modules and carbohydrate binding modules (CBMs) or (functional) domains of enzymes that degrade, modify or create glycosidic bonds. HMMs for the Pfam domains were downloaded from the Pfam database. Microbial and metagenomic protein sequences were retrieved from IMG 3.4 and IMG/M 3.3 [39,40]. HMMER 3 [41] with gathering thresholds was used to annotate the samples with Pfam domains. Each Pfam family has a manually defined gathering threshold for the bit score that was set in such a way that there were no false-positives detected. For annotation of protein sequences with CAZy families, the available annotations from the database were used. For annotations not available in the database, HMMs for the CAZy families were downloaded from dbCAN (http://csbl.bmb.uga.edu/dbcan) [42]. To be considered a valid annotation, matches to Pfam and dbCAN protein domain HMMs in the protein sequences were required to be supported by an e-value of at least 1e-02 and a bit score of at least 25. Additionally, we excluded matches to dbCAN HMMs with an alignment longer than 100 bp that did not exceed an e-value of 1e-04. Multiple matches of one and the same protein sequence against a single Pfam or dbCAN HMM exceeding the thresholds were counted as one annotation.

Phenotype annotation of lignocellulose-degrading and non-degrading microbes

We defined genomes and metagenomes as originating from either lignocellulose-degrading or non-lignocellulose-degrading microbial species based on information provided by IMG/M and in the literature. For every microbial genome and metagenome, we downloaded the genome publication and further available articles (Additional file 1: Table S1). We did not consider genomes for which no publications were available. For cellulose-degrading species annotated in IMG, we verified these assignments based on these publications. We used text search to

identify the keywords “cellulose”, “cellulase”, “carbon source”, “plant cell wall” or “polysaccharide” in the publications for non-cellulose-degrading species. We subsequently read all articles that contained these keywords in detail to classify the respective organism as either cellulose-degrading or non-degrading. Genomes that could not be unambiguously classified in this manner were excluded from our study.

Classification with an ensemble of support vector machine classifiers

The SVM is a supervised learning method that can be used for data classification [20,21]. Here, we use an L1-regularized L2-loss SVM, which solves the following optimization problem for a set of instance-label pairs $(\vec{x_i}, y_i)$, $\vec{x_i} \in R^n$, $y_i \in \{-1, +1\}$, $i = 1, \ldots, l$:

$$\min_{\vec{w}} || \vec{w} ||_1 + C \sum_{i=1}^{l} \left(\max(0, 1 - y_i \vec{w}^T \vec{x_i}) \right)^2, \quad (1)$$

where $C \geq 0$ is a penalty parameter. This choice of the classifier and regularization term results in sparse models, where non-zero components of the weight vector $\vec{w}$ are important for discrimination between the classes [43]. SVM classification was performed using the LIBLINEAR package [44]. The components of $\vec{x_i}$ are either binary valued and represent the presence or absence of protein domains, or continuous-valued and represent the frequency of a particular protein domain or gene family relative to the total number of annotations. All features were normalized by dividing by the sum of all vector entries and subsequently scaled, such that the value of each feature was within the range [0,1]. The label +1 was assigned to genomes and metagenomes of plant biomass-degrading microorganisms, the label -1 to all sequences from non-degrading ones. Classification of the draft genomes assembled from the fiber-adherent microbial community from cow rumen was performed with a voting committee of multiple models with different settings for the penalty parameter C that performed comparably well. A majority vote of the 5 most accurate models was used here obtained in a single cross-validation run with different settings of the penalty parameter C.

Performance evaluation

The assignment accuracy of a classifier was determined with a standard nested cross-validation (nCV) setup [45]. In nCV, an outer cross-validation loop is organized according to the leave-one-out principle: In each step, one data point is left out. In an inner loop, the optimal parameters for the model (here, the penalty parameter C) are sought, in a second cross-validation experiment with the remaining data points. For determination of the best setting for the penalty parameter C, values for $C = 10^x$, x = -3.0, -2.5, -2.25, . . ., 0 were tried. Values of the parameter C larger than 1 were not tested extensively, as we found that they resulted in models with similar accuracies. This is in agreement with the Liblinear tutorial in the appendix of [44] which states that once the parameter C exceeds a certain value, the obtained models have a similar accuracy. The SVM with the penalty parameter setting yielding the best assignment accuracy was used to predict the class membership of the left out data point. The class membership predictions for all data points were used to determine the assignment accuracy of the classifier, based on their agreement with the correct assignments. For this purpose, the result of each leave-one-out experiment was classified as either a true positive (TP - correctly predicted lignocellulose degraders), true negative (TN - correctly predicted non-degraders), false positive (FP - non-degraders predicted to be degraders) or a false negative assignment (FN - degraders predicted to be non-degraders). The recall of the positive class and the true negative rate of the classifier were calculated according to the following equations:

$$\text{Re}call = \frac{TP}{TP + FN} \quad (2)$$

$$True\ negative\ rate = \frac{TN}{TN + FP} \quad (3)$$

The average of the recall and the true negative rate, the macro-accuracy, was used as the assignment accuracy to assess the overall performance:

$$MACC = \frac{\text{Re}call + True\ negative\ rate}{2} \quad (4)$$

Subsequently, we identified the settings for the penalty parameter C with the best macro-accuracy by leave-one-out cross-validation. The parameter settings resulting in the most accurate models were used to each train a separate model on the entire data set. Prediction of the five best models were combined to form a voting committee and used for the classification of novel sequence samples such as the partial genome reconstructions from the cow rumen metagenome of switch-grass adherent microbes (see Additional file 2: Table S2 for an evaluation and meta-parameter settings of these ensembles of classifiers).

Feature selection

An SVM model can be represented by a sparse weight vector $\vec{w}$. The positive and negative components of $\vec{w}$, the ‘feature weights’, specify the relative importance of the protein domains or CAZy families for discrimination between plant biomass-degrading and non-plant

biomass-degrading microorganisms. To determine the most distinctive features for the positive class (that is, the lignocellulose degraders), we selected all features that received a positive weight in weight vectors of the majority of the five most accurate models. This ensemble of models was also used for classification of the cow rumen draft genomes of uncultured microbes (see Classification with a SVM).

Additional files

Additional file 1: Table S1. Isolate strains and metagenome samples used in this study. The signs "+" and "-" indicate availability of CAZy or Pfam annotation data. The symbol * marks strains for which we provide another reference than the genome publication characterizing the metabolic capacities of the respective strain.

Additional file 2: Table S2: Evaluation and meta-parameter settings of the ensembles of classifiers. The ensembles were used for feature selection and phenotype classification of the (draft) genomes and metagenomes. The macro-accuracy for each model for a discrete set of values for the parameter *C* was calculated in cross-validation experiments. The five best models were selected based on macro-accuracy. The mean of the exponentially transformed parameter *C* and the mean macro-accuracy for these five models are shown for all trained classifiers. For details on the different ensemble classifiers, see the Results section in the manuscript.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AW, YT, PBP and ACM designed the study, interpreted the results and wrote the manuscript. AW and YT conducted the experiments under the supervision of ACM. SGAK and AW computed the CAZy family and protein domain annotations. All authors read and approved the final manuscript.

Acknowledgements

YT, AW and ACM were supported by the Max Planck society and Heinrich Heine University Düsseldorf. PBP gratefully acknowledges support from the Research Council of Norway and the Bilateralt Forskningssamarbeid - Prosjektetablering (BILAT) program. The authors are grateful to Angela Rennwanz who helped downloading the articles for the microbial genomes used in our analysis.

Author details

[1]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, Saarbrücken 66123, Germany. [2]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Post Office Box 5003, Ås 1432, Norway. [3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany.

Received: 16 August 2012 Accepted: 12 February 2013
Published: 15 February 2013

References

1. Rubin EM: **Genomics of cellulosic biofuels.** *Nature* 2008, **454:**841–845.
2. Kaylen M, Van Dyne DL, Choi YS, Blasé M: **Economic feasibility of producing ethanol from lignocellulosic feedstocks.** *Biores Technol* 2000, **72:**19–32.
3. Lee J: **Biological conversion of lignocellulosic biomass to ethanol.** *J Biotechnol* 1997, **56:**1–24.
4. Wheals AE, Basso LC, Alves DMG, Amorim HV: **Fuel ethanol after 25 years.** *TIBTECH* 1999, **17:**482–487.
5. Mitchell WJ: **Physiology of carbohydrate to solvent conversion by clostridia.** *Adv Microb Physiol* 1998, **39:**31–130.
6. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD: **Biomass recalcitrance: engineering plants and enzymes for biofuels production.** *Science* 2007, **315:**804–807.
7. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P, Han CS, Lucas S, Misra M, Myers GL, *et al*: **Genome sequence of the cellulolytic gliding bacterium cytophaga hutchinsonii.** *Appl Environ Microbiol* 2007, **73:**3536–3546.
8. Brumm P, Mead D, Boyum J, Drinkwater C, Gowda K, Stevenson D, Weimer P: **Functional annotation of fibrobacter succinogenes S85 carbohydrate active enzymes.** *Appl Biochem Biotechnol* 2010, doi:10.1007/s12010-010-9070-5.
9. Morrison M, Pope PB, Denman SE, McSweeney CS: **Plant biomass degradation by gut microbiomes: more of the same or something new?** *Curr Opin Biotech* 2009, **20:**358–363.
10. Brumm P, Hermanson S, Hochstein B, Boyum J, Hermersmann N, Gowda K, Mead D: **Mining *Dictyoglomus turgidum* for enzymatically active carbohydrases.** *Appl Biochem Biotechnol* 2010, doi:10.1007/s12010-010-9029-6.
11. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng J-F, Hugenholtz P, McSweeney CS, Morrison M: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different to other herbivores.** *Proc Natl Acad Sci USA* 2010, **107:**14793–14798.
12. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, *et al*: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560–565.
13. Brulc JM, Antonopoulos DA, Berg Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, *et al*: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.** *Proc Natl Acad Sci USA* 1948, **2009:**106.
14. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T, *et al*: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen.** *Science* 2011, **331:**463–467.
15. Pope PB, Mackenzie AK, Gregor I, Smith W, Sundset MA, McHardy AC, Morrison M, Eijsink VGH: **Metagenomics of the svalbard reindeer rumen microbiome reveals abundance of polysaccharide utilization loci.** *PLoS One* 2012, doi:10.1371/journal.pone.0038571.
16. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2012, **40:**D13–D25.
17. Beerenwinkel N, Dumer M, Oette M, Korn K, Hoffmann D, Kaiser R, Lengauer T, Selbig J, Walter H: **Geno2Pheno: estimating phenotypic drug resistance from HIV-1 genotypes.** *Nucleic Acids Res* 2003, **31:**3850–3855.
18. Yosef N, Gramm J, Wang Q-F, Noble WS, Karp RM, Sharan R: **Prediction of phenotype information from genotype data.** *Commun Inf Syst* 2010, **10:**99–114.
19. Someya S, Kakuta M, Morita M, Sumikoshi K, Cao W, Ge Z, Hirose O, Nakamura S, Terada T, Shimizu K: **Prediction of carbohydrate-binding proteins from sequences using support vector machines.** *Adv Bioinformatics* 2010, doi:10.1155/2010/289301.
20. Cortes C, Vapnik V: **Support-vector networks.** *Mach Learn* 1995, **20:**273–297.
21. Boser B, Guyon I, Vapnik V: **A training algorithm for optimal margin classifiers.** In *Fifth Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Pittsburgh: ACM; 1992:144–152.
22. Chertkov O, Sikorski J, Nolan M, Lapidus A, Lucas S, Del Rio TG, Tice H, Cheng J-F, Goodwin L, Pitluck S, *et al*: **Complete genome sequence of Thermomonospora curvata type strain (B9).** *Stand Genomic Sci* 2011, **4:**13–22.
23. Anderson I, Abt B, Lykidis A, Klenk HP, Kyrpides N, Ivanova N: **Genomics of aerobic cellulose utilization systems in actinobacteria.** *PLoS One* 2012, **7:**e39331.
24. Aspeborg H, Coutinho PM, Wang Y, Brumer H 3rd, Henrissat B: **Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5).** *BMC Evol Biol* 2012, **12:**186.
25. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Carbohydrate-binding modules: fine-tuning polysaccharide recognition.** *Biochem J* 2004, **15:**769–781.
26. Suen G, Weimer PJ, Stevenson DM, Aylward FO, Boyum J, Deneke J, Drinkwater C, Ivanova NN, Mikhailova N, Chertkov O, *et al*: **The complete genome sequence of fibrobacter succinogenes S85 reveals a cellulolytic and metabolic specialist.** *PLoS One* 2011, **6:**e18814.

27. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: **SMART: a web-based tool for the study of genetically mobile domains.** *Nucleic Acids Res* 2000, **28:**231–234.
28. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40:**D290–D301.
29. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Res* 2001, **29:**41–43.
30. Wilson DB: **Three microbial strategies for plant cell wall degradation.** *Ann N Y Acad Sci* 2008, **1125:**289–297.
31. Olson DG, Tripathi SA, Giannone RJ, Lo J, Caiazza NC, Hogsett DA, Hettich RL, Guss AM, Dubrovsky G, Lynd LR: **Deletion of the Cel48S cellulase from Clostridium thermocellum.** *Proc Natl Acad Sci USA* 2010, doi:10.1073/pnas.1003584107.
32. DeBoy RT, Mongodin EF, Fouts DE, Tailford LE, Khouri H, Emerson JB, Mohamoud Y, Watkins K, Henrissat B, Gilbert HJ, Nelson KE: **Insights into plant cell wall degradation from the genome sequence of the soil bacterium Cellvibrio japonicus.** *J Bacteriol* 2008, **190:**5455–5463.
33. Taylor LE, Henrissat B, Coutinho PM, Ekborg NA, Hutcheson SW, Weiner RM: **Complete cellulase system in the marine bacterium Saccharophagus degradans strain 2-40 T.** *J Bacteriol* 2006, **188:**3849–3861.
34. Hervé C, Rogowski A, Blake AW, Marcus SE, Gilbert HJ, Knox JP: **Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects.** *Proc Natl Acad Sci USA* 2010, **107:**15293–15298.
35. Duan CJ, Feng JX: **Mining metagenomes for novel cellulase genes.** *Biotechnol Lett* 2010, **32:**1765–1775.
36. Wilson DB: **Evidence for a novel mechanism of microbial cellulose degradation.** *Cellulose* 2009, **16:**723–727.
37. Lynd LR, Weimer PJ, van Zyl WH, Pretorius IS: **Microbial cellulose utilization: fundamentals and biotechnology.** *Microbiol Mol Biol Rev* 2002, **66:**506–577.
38. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The carbohydrate-active EnZymes database (CAZy): an expert resource for glycogenomics.** *Nucleic Acids Res* 2009, **37:**D233–D238.
39. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, *et al*: **IMG/M: the integrated metagenome data management and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D123–D129.
40. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, *et al*: **IMG: the integrated microbial genomes database and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D115–D122.
41. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39:**W29–W37.
42. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a web resource for automated carbohydrate-active enzyme annotation.** *Nucleic Acids Res* 2012, doi:10.1093/nar/gks479.
43. Yaun G-X, Chang K-W, Hsieh C-J, Lin C-J: **A comparison of optimization methods for large-scale L1-regularized linear classification.** *J Mach Learn Res* 2010, **11:**3183–3234.
44. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ: **LIBLINEAR: a library for large linear classification.** *J Mach Learn Res* 2008, **9:**1871–1874.
45. Ruschhaupt M, Huber W, Poustka A, Mansmann U: **A compendium to ensure computational reproducibility in high-dimensional classification tasks.** *Stat Appl Genet Mol Biol* 2004, **3:**Article 37.

doi:10.1186/1754-6834-6-24
Cite this article as: Weimann *et al.*: De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes. *Biotechnology for Biofuels* 2013 **6**:24.

Supplementary Table S1. Isolate strains and metagenome samples used in this study. The signs "+" and "-" indicate availability of CAZy or Pfam annotation data. The symbol * marks strains for which we provide another reference than the genome publication characterizing the metabolic capacities of the respective strain.

				Pfam	CAZy			Reference
				Binary/ weighted	a	b	c	
Cellulose-degrading organisms	*Meta-genomes*	1	*Macropus eugenii* gut microbiome (tammar wallaby)	+	–	+	+	Pope et al 2010) [1]
		2	Cow rumen microbiome	–	–	–	+	Brulc et al 2009 [2]
		3	Termite gut microbiome	+	–	–	+	Warnecke et al 2007 [3]
	Genomes	1	*Acidothermus cellulolyticus* 11B	+	+			Barabote et al 2009 [4]
		2	*Anaerocellum thermophilum* Z-1320, DSM 6725	+	+			Kataeva et al 2009 [5]
		3	*Bryantella formatexigens* I-52, DSM 14469	+	–			Wolin et al 2003 [6]
		4	*Caldicellulosiruptor saccharolyticus* DSM 8903	+	+			Rainey et al 1994* [7]
		5	*Cellulomonas flavigena* 134, DSM 20109	+	+			Abt et al 2010 [8]
		6	*Cellvibrio japonicus Ueda* 107	+	+			DeBoy et al 2008 [9]
		7	*Clostridium cellulolyticum* H10	+	+			Petitdemange et al 1984* [10]
		8	*Clostridium phytofermentans* ISDg	+	+			Warnick et al 2002* [11]
		9	*Clostridium thermocellum* ATCC 27405	+	+			Feinberg et al 2011 [12]
		10	*Cytophaga hutchinsonii* ATCC 33406	+	+			Xie et al 2007 [13]
		11	*Dictyoglomus turgidum* DSM 6724	+	+			Brumm et al 2011 [14]
		12	*Fibrobacter succinogenes succinogenes* S85	+	+			Bae et al 1993* [15]
		13	*Postia placenta* Mad-698-R	+	–			Martinez et al 2009 [16]
		14	*Ruminococcus flavefaciens* FD-1	+	–			Berg Miller et al 2009 [17]
		15	*Saccharophagus degradans* 2-40	+	+			Fraiberg et al 2010 [18]

		16	*Teredinibacter turnerae* T7901	+	+	Yang et al 2009 [19]
		17	*Thermobifida fusca* YX	+	+	Lykidis et al 2007 [20]
		18	*Thermomonospora curvata* DSM 43183	+	+	Chertkov et al 2011 [21]
		19	*Xylanimonas cellulosilytica* XIL07, DSM 15894	+	+	Foster et al 2010 [22]
Non-cellulose degrading organisms	*Genomes*	1	*Acetobacter pasteurianus* IFO 3283-01	+	+	Azuma et al 2009 [23]
		2	*Acidimicrobium ferrooxidans* DSM 10331	+	+	Clum et al 2009 [24]
		3	*Acidithiobacillus ferrooxidans* ATCC 23270	+	+	Valdés et al 2008 [25]
		4	*Actinosynnema mirum* DSM 43827	+	+	Land et al 2009 [26]
		5	*Agrobacterium tumefaciens* C58 (Cereon)	+	+	Wood et al 2001 [27]
		6	*Alcanivorax borkumensis* SK2	+	+	Schneiker et al 2006 [28]
		7	*Alkalilimnicola ehrlichei* MLHE-1	+	+	Hoeft et al 2007* [29]*
		8	*Alkaliphilus metalliredigens* QYMF	+	+	Fu et al 2009)* [30]
		9	*Archaeoglobus fulgidus* DSM 4304	+	-	Klenk et al 1997 [31]
		10	*Arthrobacter aurescens* TC1	+	+	Mongodin et al 2006 [32]
		11	*Azoarcus* sp. BH72	+	+	Krause et al 2006 [33]
		12	*Azorhizobium caulinodans* ORS 571	+	+	Liu et al 2011 [34]
		13	*Azotobacter vinelandii* DJ, ATCC BAA-1303	+	+	Setubal et al 2009 [35]
		14	*Beijerinckia indica indica* ATCC 9039	+	+	Tamas et al 2010 [36]
		15	*Candidatus amoebophilus asiaticus* 5a2	+	+	Schmitz-Esser et al 2010 [37]
		16	*Chloroflexus aurantiacus* J-10-fl	+	+	Tang et al 2011 [38]
		17	*Chromobacterium violaceum* ATCC 12472	+	+	Brazilian National Genome Project Consortium 2003 [39]
		18	*Comamonas testosteroni* KF-1	+	-	Ma et al 2009 [40]
		19	*Cupriavidus taiwanensis*	+	-	Amadou et al 2008 [41]
		20	*Cyanothece* sp. ATCC 51142	+	+	Welsh et al 2008 [42]
		21	*Dehalococcoides ethenogenes* 195	+	+	Seshadri et al 2005 [43]
		22	*Desulfatibacillum alkenivorans* AK-01	+	+	Callaghan et al 2012 [44]
		23	*Desulfitobacterium hafniense* DCB-2	+	+	Shinoda et al 2006) [45]

		24	*Desulfohalobium retbaense* DSM 5692	+	+	Spring et al 2010 [46]
		25	*Desulfomicrobium baculatum* DSM 4028	+	+	Copeland et al 2009 [47]
		26	*Desulfotalea psychrophila* LSv54	+	+	Rabus et al 2004 [48]
		27	*Desulfotomaculum reducens* MI-1	+	+	Junier et al 2010 [49]
		28	*Diaphorobacter* sp. TPSYc	+	+	Byrne-Bailey et al 2010 [50]
		29	*Frankia alni* ACN14a	+	+	Normand et al 2007 [51]
		30	*Geobacter bemidjiensis* Bem	+	+	Aklujkar et al 2010 [52]
		31	*Hyperthermus butylicus* DSM 5456	+	-	Brügger et al 2007 [53]
		32	*Klebsiella pneumoniae* 342	+	+	Yi et al 2010 [54]
		33	*Lactobacillus salivarius salivarius* UCC118	+	+	Jimenez et al 2010 [55]
		34	*Magnetococcus* sp. MC-1	+	+	Schübbe et al 2009 [56]
		35	*Marinobacter aquaeolei* VT8	+	+	Singer et al 2011* [57]
		36	*Mesorhizobium loti* MAFF303099	+	+	Kaneko et al 2000 [58]
		37	*Metallosphaera sedula* DSM 5348	+	-	Auernik et al 2008 [59]
		38	*Methanobrevibacter smithii* ATCC 35061	+	-	Hansen et al 2011 [60]
		39	*Methanocaldococcus fervens* AG86	+	-	Galperin and Cochrane 2009 [61]
		40	*Methanococcoides burtonii* DSM 6242	+	-	Saunders et al 2003 [62]
		41	*Methanocorpusculum labreanum* Z	+	-	Anderson et al 2009 [63]
		42	*Methanoculleus marisnigri* JR1	+	-	Anderson et al 2009 [63]
		43	*Methanopyrus kandleri* AV19	+	-	Slesarev et al 2002 [64]
		44	*Methanosarcina acetivorans* C2A	+	-	Galagan et al 2002 [65]
		45	*Methanosphaera stadtmanae* DSM 3091	+	-	Fricke et al 2006 [66]
		46	*Methylibium petroleiphilum* PM1	+	+	Kane et al 2007 [67]
		47	*Methylocella silvestris* BL2	+	+	Chen et al 2010 [68]
		48	*Nautilia profundicola* Am-H	+	+	Campbell et al 2009 [69]
		49	*Nitrobacter hamburgensis* X14	+	+	Starkenburg et al 2008 [70]
		50	*Nitrosococcus oceani* ATCC 19707	+	+	Klotz et al 2006 [71]
		51	*Nitrosomonas europaea* ATCC 19718	+	+	Chain et al 2003 [72]
		52	*Nitrosopumilus maritimus* SCM1	+	-	Walker et al 2010 [73]

		53	*Nitrosospira multiformis* ATCC 25196	+	+	Norton et al 2008 [74]
		54	*Nostoc punctiforme* PCC 73102	+	+	Meeks et al 2001 [75]
		55	*Paracoccus denitrificans* PD1222	+	+	Siddavattam et al 2011 [76]
		56	*Parvibaculum lavamentivorans* DS-1	+	+	Schleheck et al 2007 [77]
		57	*Pelotomaculum thermopropionicum* SI	+	+	Kosaka et al 2008 [78]
		58	*Persephonella marina* EX-H1	+	+	Reysenbach et al 2009 [79]
		59	*Polaromonas naphthalenivorans* CJ2	+	+	Yagi et al 2009 [80]
		60	*Pseudomonas mendocina* ymp	+	+	Guo et al 2011 [81]
		61	*Pyrobaculum aerophilum* IM2	+	-	Fitz-Gibbon et al 2002 [82]
		62	*Pyrococcus abyssi* GE5	+	-	(Cohen et al 2003 [83]
		63	*Rhizobium etli* CFN 42	+	+	Fauvart et al 2011 [84]
		64	*Rhodobacter sphaeroides* KD131	+	+	Porter et al 2011 [85]
		65	*Rhodococcus* sp. RHA1	+	+	Takeda et al 2010 [86]
		66	*Rhodoferax ferrireducens* T118	+	+	Risso et al 2009 [87]
		67	*Rhodospirillum rubrum* ATCC 11170	+	+	Munk et al 2011 [88]
		68	*Sinorhizobium medicae* WSM419	+	+	Reeve et al 2010 [89]
		69	*Slackia heliotrinireducens* DSM 20476	+	+	Pukall et al 2009 [90]
		70	*Streptococcus thermophilus* LMD-9	+	+	Sun et al 2011 [91]
		71	*Sulfolobus acidocaldarius* DSM 639	+	-	Chen et al 2005 [92]
		72	*Sulfurospirillum deleyianum* DSM 6946	+	+	Sikorski et al 2010 [93]
		73	*Synechococcus elongatus* PCC 7942	+	+	Holtman et al 2005 [94]
		74	*Synechococcus* sp. CC9605	+	+	Jenkins et al 2006)* [95]
		75	*Syntrophomonas wolfei wolfei* Goettingen	+	+	Sieber et al 2010 [96]
		76	*Syntrophus aciditrophicus* SB	+	+	McInerney et al 2007 [97]
		77	*Thermotoga lettingae* TMO	+	+	Zhaxybayeva et al 2009 [98]
		78	*Thioalkalivibrio* sp. HL-EbGR7	+	+	Muyzer et al 2011 [99]
		79	*Thiobacillus denitrificans* ATCC 25259	+	+	Beller et al 2006 [100]
		80	*Thiomicrospira crunogena* XCL-2	+	+	Scott et al 2006 [101]
		81	*Thiomicrospira denitrificans* ATCC 33889	+	+	Sievert et al 2008 [102]
		82	*Zymomonas mobilis mobilis* ZM4	+	-	Pappas et al 2011 [103]

References

1. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng JF, Hugenholtz P, McSweeney CS, Morrison M: **Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107:**14793-14798.
2. Brulc JM, Antonopoulos DA, Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, et al: **Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**1948-1953.
3. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, et al: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450:**560-565.
4. Barabote RD, Xie G, Leu DH, Normand P, Necsulea A, Adney WS, Xu XC, Lapidus A, Daubin V, Me C, et al: **Complete genome of the cellulolytic thermophile Acidothermus cellulolyticus 11B provides insights into its ecophysiological and evolutionary adaptations.** *Genome Research* 2009**:**1033-1043.
5. Kataeva Ia, Yang S-J, Dam P, Poole FL, Yin Y, Zhou F, Chou W-c, Xu Y, Goodwin L, Sims DR, et al: **Genome sequence of the anaerobic, thermophilic, and cellulolytic bacterium "Anaerocellum thermophilum" DSM 6725.** *Journal of bacteriology* 2009, **191:**3760-3761.
6. Wolin MJ, Miller TL, Collins MD, Lawson PA: **Formate-dependent growth and homoacetogenic fermentation by a bacterium from human feces: description of Bryantella formatexigens gen. nov., sp. nov.** *Applied and environmental microbiology* 2003, **69:**6321-6326.
7. Rainey FA, Donnison AM, Janssen PH, Saul D, Rodrigo A, Bergquist PL, Daniel RM, Stackebrandt E, Morgan HW: **Description of Caldicellulosiruptor saccharolyticus gen. nov., sp. nov: an obligately anaerobic, extremely thermophilic, cellulolytic bacterium.** *FEMS microbiology letters* 1994, **120:**263-266.
8. Abt B, Foster B, Lapidus A, Clum A, Sun H, Pukall Ru, Lucas S, Glavina Del Rio T, Nolan M, Tice H, et al: **Complete genome sequence of Cellulomonas flavigena**

type strain (134). *Standards in genomic sciences* 2010, **3:**15-25.
9. DeBoy RT, Mongodin EF, Fouts DE, Tailford LE, Khouri H, Emerson JB, Mohamoud Y, Watkins K, Henrissat B, Gilbert HJ, Nelson KE: **Insights into plant cell wall degradation from the genome sequence of the soil bacterium Cellvibrio japonicus.** *Journal of bacteriology* 2008, **190:**5455-5463.
10. Petitdemange E, Biologique LDC, Ce V-l-n, Bacte C: **Clostridium cellulolyticum sp. nov. , a Cellulolytic, Mesophilic Species from Decayed Grass.** *International Journal* 1984**:**155-159.
11. Warnick TA, Methe BA, Leschine SB: **Clostridium phytofermentans sp. nov., a cellulolytic mesophile from forest soil.** *International journal of systematic and evolutionary microbiology* 2002, **52:**1155-1160.
12. Feinberg L, Foden J, Barrett T, Davenport KW, Bruce D, Detter C, Tapia R, Han C, Lapidus A, Lucas S, et al: **Complete genome sequence of the cellulolytic thermophile Clostridium thermocellum DSM1313.** *Journal of bacteriology* 2011, **193:**2906-2907.
13. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P, Han CS, Lucas S, Misra M, Myers GL, et al: **Genome sequence of the cellulolytic gliding bacterium Cytophaga hutchinsonii.** *Applied and environmental microbiology* 2007, **73:**3536-3546.
14. Brumm P, Hermanson S, Hochstein B, Boyum J, Hermersmann N, Gowda K, Mead D: **Mining Dictyoglomus turgidum for enzymatically active carbohydrases.** *Applied biochemistry and biotechnology* 2011, **163:**205-214.
15. Bae HD, McAllister Ta, Yanke J, Cheng KJ, Muir aD: **Effects of Condensed Tannins on Endoglucanase Activity and Filter Paper Digestion by Fibrobacter succinogenes S85.** *Applied and environmental microbiology* 1993, **59:**2132-2138.
16. Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P, et al: **Genome, transcriptome, and secretome analysis of wood decay fungus Postia placenta supports unique mechanisms of lignocellulose conversion.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**1954-1959.
17. Berg Miller ME, Antonopoulos Da, Rincon MT, Band M, Bari A, Akraiko T, Hernandez A, Thimmapuram J, Henrissat B, Coutinho PM, et al: **Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of Ruminococcus flavefaciens FD-1.** *PloS one* 2009, **4:**e6650.
18. Fraiberg M, Borovok I, Weiner RM, Lamed R: **Discovery and characterization of cadherin domains in Saccharophagus degradans 2-40.** *Journal of bacteriology* 2010, **192:**1066-1074.
19. Yang JC, Madupu R, Durkin aS, Ekborg Na, Pedamallu CS, Hostetler JB, Radune D, Toms BS, Henrissat B, Coutinho PM, et al: **The complete genome of Teredinibacter turnerae T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms).** *PloS one* 2009, **4:**e6085.
20. Lykidis A, Mavromatis K, Ivanova N, Anderson I, Land M, DiBartolo G, Martinez M, Lapidus A, Lucas S, Copeland A, et al: **Genome sequence and analysis of the soil cellulolytic actinomycete Thermobifida fusca YX.** *Journal of*

bacteriology 2007, **189:**2477-2486.
21. Chertkov O, Sikorski J, Nolan M, Lapidus A, Lucas S, Del Rio TG, Tice H, Cheng J-F, Goodwin L, Pitluck S, et al: **Complete genome sequence of Thermomonospora curvata type strain (B9).** *Standards in genomic sciences* 2011, **4:**13-22.
22. Foster B, Pukall Ru, Abt B, Nolan M, Glavina Del Rio T, Chen F, Lucas S, Tice H, Pitluck S, Cheng J-F, et al: **Complete genome sequence of Xylanimonas cellulosilytica type strain (XIL07).** *Standards in genomic sciences* 2010, **2:**1-8.
23. Azuma Y, Hosoyama A, Matsutani M, Furuya N, Horikawa H, Harada T, Hirakawa H, Kuhara S, Matsushita K, Fujita N, Shirai M: **Whole-genome analyses reveal genetic instability of Acetobacter pasteurianus.** *Nucleic Acids Res* 2009, **37:**5768-5783.
24. Clum A, Nolan M, Lang E, Glavina Del Rio T, Tice H, Copeland A, Cheng J-F, Lucas S, Chen F, Bruce D, et al: **Complete genome sequence of Acidimicrobium ferrooxidans type strain (ICP).** *Standards in genomic sciences* 2009, **1:**38-45.
25. Valdés J, Pedroso I, Quatrini R, Dodson RJ, Tettelin H, Blake R, Eisen Ja, Holmes DS: **Acidithiobacillus ferrooxidans metabolism: from genome sequence to industrial applications.** *BMC genomics* 2008, **9:**597.
26. Land M, Lapidus A, Mayilraj S, Chen F, Copeland A, Del Rio TG, Nolan M, Lucas S, Tice H, Cheng J-F, et al: **Complete genome sequence of Actinosynnema mirum type strain (101).** *Standards in genomic sciences* 2009, **1:**46-53.
27. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF, et al: **The genome of the natural genetic engineer Agrobacterium tumefaciens C58.** *Science (New York, NY)* 2001, **294:**2317-2323.
28. Schneiker S, Martins dos Santos ViaP, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, et al: **Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis.** *Nature biotechnology* 2006, **24:**997-1004.
29. Hoeft SE, Blum JS, Stolz JF, Tabita FR, Witte B, King GM, Santini JM, Oremland RS: **Alkalilimnicola ehrlichii sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor.** *International journal of systematic and evolutionary microbiology* 2007, **57:**504-512.
30. Fu H-l, Meng Y, Ordóñez E, Villadangos AF, Bhattacharjee H, Gil JA, Mateos LM, Rosen BP: **Properties of arsenite efflux permeases (Acr3) from Alkaliphilus metalliredigens and Corynebacterium glutamicum.** *The Journal of biological chemistry* 2009, **284:**19887-19895.
31. Klenk HP, Clayton Ra, Tomb JF, White O, Nelson KE, Ketchum Ka, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, et al: **The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus.** *Nature* 1997, **390:**364-370.
32. Mongodin EF, Shapir N, Daugherty SC, DeBoy RT, Emerson JB, Shvartzbeyn A, Radune D, Vamathevan J, Riggs F, Grinberg V, et al: **Secrets of soil survival**

revealed by the genome sequence of Arthrobacter aurescens TC1. *PLoS genetics* 2006, **2:**e214.
33. Krause A, Ramakumar A, Bartels D, Battistoni F, Bekel T, Boch J, B\"ohm M, Friedrich F, Hurek T, Krause L, et al: **Complete genome of the mutualistic, N2-fixing grass endophyte Azoarcus sp. strain BH72.** *Nature biotechnology* 2006, **24:**1385-1391.
34. Liu C-T, Lee K-B, Wang Y-S, Peng M-H, Lee K-T, Suzuki S, Suzuki T, Oyaizu H: **Involvement of the azorhizobial chromosome partition gene (parA) in the onset of bacteroid differentiation during Sesbania rostrata stem nodule development.** *Applied and environmental microbiology* 2011, **77:**4371-4382.
35. Setubal JC, dos Santos P, Goldman BS, Ertesvag H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D, Cromes L, et al: **Genome sequence of Azotobacter vinelandii, an obligate aerobe specialized to support diverse anaerobic metabolic processes.** *Journal of bacteriology* 2009, **191:**4534-4545.
36. Tamas I, Dedysh SN, Liesack W, Stott MB, Alam M, Murrell JC, Dunfield PF: **Complete genome sequence of Beijerinckia indica subsp. indica.** *Journal of bacteriology* 2010, **192:**4532-4533.
37. Schmitz-Esser S, Tischler P, Arnold R, Montanaro J, Wagner M, Rattei T, Horn M: **The genome of the amoeba symbiont "Candidatus Amoebophilus asiaticus" reveals common mechanisms for host cell interaction among amoeba-associated bacteria.** *Journal of bacteriology* 2010, **192:**1045-1057.
38. Tang K-h, Barry K, Chertkov O, Dalin E, Han CS, Hauser LJ, Honchak BM, Karbach LE, Land ML, Lapidus A, et al: **Complete genome sequence of the filamentous anoxygenic phototrophic bacterium Chloroflexus aurantiacus.** *BMC genomics* 2011, **12:**334.
39. Brazilian National Genome Project Consortium: **The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100:**11660-11665.
40. Ma Y-F, Zhang Y, Zhang J-y, Chen D-w, Zhu Y, Zheng H, Wang S-y, Jiang C-y, Zhao G-p, Liu S-j: **The complete genome of Comamonas testosteroni reveals its genetic adaptations to changing environments.** *Applied and environmental microbiology* 2009, **75:**6812-6819.
41. Amadou C, Mangenot S, Glew M, Bontemps C, Capela D, Dossat C, Marchetti M, Servin B, Saad M, Schenowitz C, et al: **Genome sequence of the -rhizobium Cupriavidus taiwanensis and comparative genomics of rhizobia.** *Genome Research* 2008**:**1472-1483.
42. Welsh Ea, Liberton M, Stöckel J, Loh T, Elvitigala T, Wang C, Wollam A, Fulton RS, Clifton SW, Jacobs JM, et al: **The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**15094-15099.
43. Seshadri R, Adrian L, Fouts DE, Eisen Ja, Phillippy AM, Methe Ba, Ward NL, Nelson WC, Deboy RT, Khouri HM, et al: **Genome sequence of the PCE-dechlorinating bacterium Dehalococcoides ethenogenes.** *Science (New York, NY)* 2005, **307:**105-108.

44. Callaghan AV, Morris BEL, Pereira IAC, McInerney MJ, Austin RN, Groves JT, Kukor JJ, Suflita JM, Young LY, Zylstra GJ, Wawrik B: **The genome sequence of Desulfatibacillum alkenivorans AK-01: a blueprint for anaerobic alkane oxidation.** *Environmental Microbiology* 2012, **14:**101-113.
45. Shinoda Y, Ikenaga Y, Abe M, Naito K, Inatomi K, Furukawa K, Inui M, Yukawa H: **Complete Genome Sequence of the Dehalorespiring Bacterium Desulfitobacterium hafniense Y51 and Comparison with Dehalococcoides ethenogenes 195.** *Journal of bacteriology* 2006, **188:**2262-2274.
46. Spring S, Nolan M, Lapidus A, Glavina Del Rio T, Copeland A, Tice H, Cheng J-F, Lucas S, Land M, Chen F, et al: **Complete genome sequence of Desulfohalobium retbaense type strain (HR(100)).** *Standards in genomic sciences* 2010, **2:**38-48.
47. Copeland A, Spring S, Göker M, Schneider S, Lapidus A, Del Rio TG, Tice H, Cheng J-F, Chen F, Nolan M, et al: **Complete genome sequence of Desulfomicrobium baculatum type strain (X).** *Standards in genomic sciences* 2009, **1:**29-37.
48. Rabus R, Ruepp A, Frickey T, Rattei T, Fartmann B, Stark M, Bauer M, Zibat A, Lombardot T, Becker I, et al: **The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments.** *Environ Microbiol* 2004, **6:**887-902.
49. Junier P, Junier T, Podell S, Sims DR, Detter JC, Lykidis A, Han CS, Wigginton NS, Gaasterland T, Bernier-Latmani R: **The genome of the Gram-positive metal- and sulfate-reducing bacterium Desulfotomaculum reducens strain MI-1.** *Environmental Microbiology* 2010, **12:**2738-2754.
50. Byrne-Bailey KG, Weber Ka, Chair AH, Bose S, Knox T, Spanbauer TL, Chertkov O, Coates JD: **Completed genome sequence of the anaerobic iron-oxidizing bacterium Acidovorax ebreus strain TPSY.** *Journal of bacteriology* 2010, **192:**1475-1476.
51. Normand P, Lapierre P, Tisa LS, Gogarten JP, Alloisio N, Bagnarol E, Bassi CA, Berry AM, Bickhart DM, Choisne N, et al: **Genome characteristics of facultatively symbiotic Frankia sp. strains reflect host range and host plant biogeography.** *Genome Research* 2007**:**7-15.
52. Aklujkar M, Young ND, Holmes D, Chavan M, Risso C, Kiss HE, Han CS, Land ML, Lovley DR: **The genome of Geobacter bemidjiensis, exemplar for the subsurface clade of Geobacter species that predominate in Fe(III)-reducing subsurface environments.** *BMC genomics* 2010, **11:**490.
53. Brügger K, Chen L, Stark M, Zibat A, Redder P, Ruepp A, Awayez M, She Q, Garrett Ra, Klenk H-P: **The genome of Hyperthermus butylicus: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C.** *Archaea (Vancouver, BC)* 2007, **2:**127-135.
54. Yi H, Xi Y, Liu J, Wang J, Wu J, Xu T, Chen W, Chen B, Lin M, Wang H, et al: **Sequence analysis of pKF3-70 in Klebsiella pneumoniae: probable origin from R100-like plasmid of Escherichia coli.** *PloS one* 2010, **5:**e8601.
55. Jimenez E, Langa S, Martin V, Arroyo R, Martin R, Fernandez L, Rodriguez JM: **Complete genome sequence of Lactobacillus fermentum CECT 5716, a probiotic strain isolated from human milk.** *Journal of bacteriology* 2010,

192:4800.
56. Schübbe S, Williams TJ, Xie G, Kiss HE, Brettin TS, Martinez D, Ross Ca, Sch\"uler D, Cox BL, Nealson KH, Bazylinski Da: **Complete genome sequence of the chemolithoautotrophic marine magnetotactic coccus strain MC-1.** *Applied and environmental microbiology* 2009, **75:**4835-4852.
57. Singer E, Webb EA, Nelson WC, Heidelberg JF, Ivanova N, Pati A, Edwards KJ: **Genomic potential of Marinobacter aquaeolei, a biogeochemical "opportunitroph".** *Applied and environmental microbiology* 2011, **77:**2763-2771.
58. Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe a, Idesawa K, Ishikawa a, Kawashima K, et al: **Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti (supplement).** *DNA research : an international journal for rapid publication of reports on genes and genomes* 2000, **7:**381-406.
59. Auernik KS, Maezato Y, Blum PH, Kelly RM: **The genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon Metallosphaera sedula provides insights into bioleaching-associated metabolism.** *Applied and environmental microbiology* 2008, **74:**682-692.
60. Hansen EE, Lozupone Ca, Rey FE, Wu M, Guruge JL, Narra A, Goodfellow J, Zaneveld JR, McDonald DT, Goodrich Ja, et al: **Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins.** *Proceedings of the National Academy of Sciences of the United States of America* 2011, **108 Suppl** 4599-4606.
61. Galperin MY, Cochrane GR: **Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.** *Nucleic acids research* 2009, **37:**D1-4.
62. Saunders NFW, Thomas T, Curmi PMG, Mattick JS, Kuczek E, Slade R, Davis J, Franzmann PD, Boone D, Rusterholtz K, et al: **Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii.** *Genome Research* 2003, **13:**1580-1588.
63. Anderson I, Ulrich LE, Lupa B, Susanti D, Porat I, Hooper SD, Lykidis A, Sieprawska-Lupa M, Dharmarajan L, Goltsman E, et al: **Genomic characterization of methanomicrobiales reveals three classes of methanogens.** *PloS one* 2009, **4:**e5797.
64. Slesarev AI, Mezhevaya KV, Makarova KS, Polushin NN, Shcherbinina OV, Shakhova VV, Belova GI, Aravind L, Natale Da, Rogozin IB, et al: **The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**4644-4649.
65. Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, et al: **The genome of M. acetivorans reveals extensive metabolic and physiological diversity.** *Genome Research* 2002, **12:**532-542.
66. Fricke WF, Seedorf H, Henne A, Kruer M, Liesegang H, Hedderich R, Gottschalk G, Thauer RK: **The genome sequence of Methanosphaera stadtmanae reveals**

why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. *Journal of bacteriology* 2006, **188:**642-658.
67. Kane SR, Chakicherla AY, Chain PSG, Schmidt R, Shin MW, Legler TC, Scow KM, Larimer FW, Lucas SM, Richardson PM, Hristova KR: **Whole-genome analysis of the methyl tert-butyl ether-degrading beta-proteobacterium Methylibium petroleiphilum PM1.** *Journal of bacteriology* 2007, **189:**1931-1945.
68. Chen Y, Crombie A, Rahman MT, Dedysh SN, Liesack W, Stott MB, Alam M, Theisen AR, Murrell JC, Dunfield PF: **Complete genome sequence of the aerobic facultative methanotroph Methylocella silvestris BL2.** *Journal of bacteriology* 2010, **192:**3840-3841.
69. Campbell BJ, Smith JL, Hanson TE, Klotz MG, Stein LY, Lee CK, Wu D, Robinson JM, Khouri HM, Eisen JA, Cary SC: **Adaptations to submarine hydrothermal environments exemplified by the genome of Nautilia profundicola.** *PLoS genetics* 2009, **5:**e1000362.
70. Starkenburg SR, Larimer FW, Stein LY, Klotz MG, Chain PSG, Sayavedra-Soto La, Poret-Peterson AT, Gentry ME, Arp DJ, Ward B, Bottomley PJ: **Complete genome sequence of Nitrobacter hamburgensis X14 and comparative genomic analysis of species within the genus Nitrobacter.** *Applied and environmental microbiology* 2008, **74:**2852-2863.
71. Klotz MG, Arp DJ, Chain PS, El-Sheikh AF, Hauser LJ, Hommes NG, Larimer FW, Malfatti SA, Norton JM, Poret-Peterson AT, et al: **Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium Nitrosococcus oceani ATCC 19707.** *Applied and environmental microbiology* 2006, **72:**6299-6315.
72. Chain P, Lamerdin J, Larimer F, Regala W, Lao V, Land M, Hauser L, Hooper A, Klotz M, Norton J, et al: **Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph Nitrosomonas europaea.** *Journal of bacteriology* 2003, **185:**2759-2773.
73. Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, Brochier-Armanet C, Chain PSG, Chan PP, Gollabgir a, et al: **Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea.** *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107:**8818-8823.
74. Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PSG, Hauser LJ, Land ML, Larimer FW, Shin MW, Starkenburg SR: **Complete genome sequence of Nitrosospira multiformis, an ammonia-oxidizing bacterium from the soil environment.** *Applied and environmental microbiology* 2008, **74:**3559-3572.
75. Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P, Atlas R: **An overview of the genome of Nostoc punctiforme, a multicellular, symbiotic cyanobacterium.** *Photosynthesis Research* 2001**:**85-106.
76. Siddavattam D, Karegoudar TB, Mudde SK, Kumar N, Baddam R, Avasthi TS, Ahmed N: **Genome of a novel isolate of Paracoccus denitrificans capable of degrading N,N-dimethylformamide.** *Journal of bacteriology* 2011, **193:**5598-5599.
77. Schleheck D, Knepper TP, Eichhorn P, Cook AM: **Parvibaculum lavamentivorans**

DS-1T degrades centrally substituted congeners of commercial linear alkylbenzenesulfonate to sulfophenyl carboxylates and sulfophenyl dicarboxylates. *Applied and environmental microbiology* 2007, **73:**4725-4732.
78. Kosaka T, Kato S, Shimoyama T, Ishii S, Abe T, Watanabe K: **The genome of Pelotomaculum thermopropionicum reveals niche-associated evolution in anaerobic microbiota.** *Genome Res* 2008, **18:**442-448.
79. Reysenbach A-L, Hamamura N, Podar M, Griffiths E, Ferreira S, Hochstein R, Heidelberg J, Johnson J, Mead D, Pohorille a, et al: **Complete and draft genome sequences of six members of the Aquificales.** *Journal of bacteriology* 2009, **191:**1992-1993.
80. Yagi JM, Sims D, Brettin T, Bruce D, Madsen EL: **The genome of Polaromonas naphthalenivorans strain CJ2, isolated from coal tar-contaminated sediment, reveals physiological and metabolic versatility and evolution through extensive horizontal gene transfer.** *Environmental Microbiology* 2009, **11:**2253-2270.
81. Guo W, Wang Y, Song C, Yang C, Li Q, Li B, Su W, Sun X, Song D, Yang X, Wang S: **Complete genome of Pseudomonas mendocina NK-01, which synthesizes medium-chain-length polyhydroxyalkanoates and alginate oligosaccharides.** *Journal of bacteriology* 2011, **193:**3413-3414.
82. Fitz-Gibbon ST, Ladner H, Kim U-J, Stetter KO, Simon MI, Miller JH: **Genome sequence of the hyperthermophilic crenarchaeon Pyrobaculum aerophilum.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99:**984-989.
83. Cohen GN, Barbe Ve, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Qu\'erellou Je, Ripp R, et al: **An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi.** *Molecular microbiology* 2003, **47:**1495-1512.
84. Fauvart M, Sánchez-Rodríguez A, Beullens S, Marchal K, Michiels J: **Genome Sequence of Rhizobium etli CNPAF512, a Nitrogen-Fixing Symbiont Isolated from Bean Root Nodules in Brazil.** *Journal of bacteriology* 2011, **193:**3158-3159.
85. Porter SL, Wilkinson Da, Byles ED, Wadhams GH, Taylor S, Saunders NJ, Armitage JP: **Genome sequence of Rhodobacter sphaeroides Strain WS8N.** *Journal of bacteriology* 2011, **193:**4027-4028.
86. Takeda H, Shimodaira J, Yukawa K, Hara N, Kasai D, Miyauchi K, Masai E, Fukuda M: **Dual two-component regulatory systems are involved in aromatic compound degradation in a polychlorinated-biphenyl degrader, Rhodococcus jostii RHA1.** *Journal of bacteriology* 2010, **192:**4741-4751.
87. Risso C, Sun J, Zhuang K, Mahadevan R, DeBoy R, Ismail W, Shrivastava S, Huot H, Kothari S, Daugherty S, et al: **Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe(III)-reducer Rhodoferax ferrireducens.** *BMC genomics* 2009, **10:**447.
88. Munk aC, Copeland A, Lucas S, Lapidus A, Del Rio TG, Barry K, Detter JC, Hammon N, Israni S, Pitluck S, et al: **Complete genome sequence of Rhodospirillum rubrum type strain (S1).** *Standards in genomic sciences* 2011, **4:**293-302.

89. Reeve W, Chain P, Ardley J, Nandesena K, Tiwari R, Malfatti S, Kiss H, Lapidus A, Co- A, Nolan M, et al: **Complete genome sequence of the Medicago microsymbiont Ensifer (Sinorhizobium) medicae strain WSM419.** *Standards in genomic sciences* 2010**:**77-86.
90. Pukall R, Lapidus A, Nolan M, Copeland A, Glavina Del Rio T, Lucas S, Chen F, Tice H, Cheng J-F, Chertkov O, et al: **Complete genome sequence of Slackia heliotrinireducens type strain (RHS 1).** *Standards in genomic sciences* 2009, **1:**234-241.
91. Sun Z, Chen X, Wang J, Zhao W, Shao Y, Wu L, Zhou Z, Sun T, Wang L, Meng H, et al: **Complete genome sequence of Streptococcus thermophilus strain ND03.** *Journal of bacteriology* 2011, **193:**793-794.
92. Chen L, Brugger K, Skovgaard M, Redder P, She Q, Torarinsson E, Greve B, Awayez M, Zibat A, Klenk HP, Garrett RA: **The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota.** *Journal of bacteriology* 2005, **187:**4992-4999.
93. Sikorski J, Lapidus A, Copeland A, Glavina T, Rio D, Nolan M, Lucas S, Chen F, Tice H, Cheng J-f, et al: **Complete genome sequence of Sulfurospirillum deleyianum type strain (5175T).** *Standards in genomic sciences* 2010**:**149-157.
94. Holtman CK, Chen Y, Sandoval P, Gonzales A, Nalty MS, Thomas TL, Youderian P, Golden SS: **High-Throughput Functional Analysis of the Synechococcus elongatus PCC 7942 Genome.** *DNA research* 2005, **12:**103-115.
95. Jenkins BD, Zehr JP, Gibson A, Campbell L: **Cyanobacterial assimilatory nitrate reductase gene diversity in coastal and oligotrophic marine environments.** *Environmental Microbiology* 2006, **8:**2083-2095.
96. Sieber JR, Sims DR, Han C, Kim E, Lykidis A, Lapidus AL, McDonnald E, Rohlin L, Culley DE, Gunsalus R, McInerney MJ: **The genome of Syntrophomonas wolfei: new insights into syntrophic metabolism and biohydrogen production.** *Environmental Microbiology* 2010, **12:**2289-2301.
97. McInerney MJ, Rohlin L, Mouttaki H, Kim U, Krupp RS, Rios-Hernandez L, Sieber J, Struchtemeyer CG, Bhattacharyya A, Campbell JW, Gunsalus RP: **The genome of Syntrophus aciditrophicus: life at the thermodynamic limit of microbial growth.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**7600-7605.
98. Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbo CL, Doolittle WF, Gogarten JP, Noll KM: **On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales.** *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106:**5865-5870.
99. Muyzer G, Sorokin DY, Mavromatis K, Lapidus A, Clum A, Ivanova N, Pati A, D'Haeseleer P, Woyke T, Kyrpides NC: **Complete genome sequence of "Thioalkalivibrio sulfidophilus" HL-EbGr7.** *Standards in genomic sciences* 2011, **4:**23-35.
100. Beller HR, Chain PSG, Letain TE, Chakicherla A, Larimer FW, Richardson PM, Coleman MA, Wood AP, Kelly DP: **The Genome Sequence of the Obligately Chemolithoautotrophic, Facultatively Anaerobic Bacterium Thiobacillus denitrificans.** *Journal of bacteriology* 2006, **188:**1473-1488.

101. Scott KM, Sievert SM, Abril FN, Ball La, Barrett CJ, Blake Ra, Boller AJ, Chain PSG, Clark Ja, Davis CR, et al: **The genome of deep-sea vent chemolithoautotroph Thiomicrospira crunogena XCL-2.** *PLoS biology* 2006, **4:**e383.
102. Sievert SM, Scott KM, Klotz MG, Chain PSG, Hauser LJ, Hemp J, H\"ugler M, Land M, Lapidus A, Larimer FW, et al: **Genome of the epsilonproteobacterial chemolithoautotroph Sulfurimonas denitrificans.** *Applied and environmental microbiology* 2008, **74:**1145-1156.
103. Pappas KM, Kouvelis VN, Saunders E, Brettin TS, Bruce D, Detter C, Balakireva M, Han CS, Savvakis G, Kyrpides NC, Typas Ma: **Genome sequence of the ethanol-producing Zymomonas mobilis subsp. mobilis lectotype strain ATCC 10988.** *Journal of bacteriology* 2011, **193:**5051-5052.

Supplementary Table S2. Evaluation and meta-parameter settings of the ensembles of classifiers used for feature selection and phenotype classification of (draft) genomes.

The macro-accuracy for each model for a discrete set of values for the parameter *C* was calculated in cross-validation experiments. The five best models were selected based on macro-accuracy. The mean exponentially transformed parameter *C* and the mean macro-accuracy for these five models are shown for all trained classifiers. For details on the different ensemble classifiers, see the result section in the manuscript.

	Mean parameter *C*	Mean macro-Accuracy
$eSVM_{bPFAM}$	$10^{-1.7}$	0.93
$eSVM_{fPFAM}$	$10^{-1.5}$	0.87
$eSVM_{CAZY\_A}$	$10^{-1.0}$	0.95
$eSVM_{CAZY\_B}$	$10^{-1.9}$	0.95
$eSVM_{CAZY\_C}$	$10^{-1.9}$	0.94
$eSVM_{CAZY\_a}$	$10^{-1.1}$	0.93
$eSVM_{CAZY\_b}$	$10^{-1.6}$	0.94
$eSVM_{CAZY\_c}$	$10^{-1.8}$	0.92

From genomes to phenotypes: Traitar, the microbial trait analyzer

Aaron Weimann[1,2,3], Kyra Mooren[1,3], Jeremy Frank[4], Phillip B. Pope[4], Andreas Bremges[1,2] and Alice C. McHardy[1,2,3,†]

[1]Computational Biology of Infection Research,
Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany

[2]German Center for Infection Research (DZIF),
partner site Hannover-Braunschweig, 38124 Braunschweig, Germany

[3]Department for Algorithmic Bioinformatics,
Heinrich Heine University, 40225 Düsseldorf, Germany

[4]Department of Chemistry, Biotechnology and Food Science,
Norwegian University of Life Sciences, Ås, 1432 Norway

[†]Correspondence: alice.mchardy@helmholtz-hzi.de

Abstract

The number of sequenced genomes is growing exponentially, profoundly shifting the bottleneck from data generation to genome interpretation. Traits are often used to characterize and distinguish bacteria, and are likely a driving factor in microbial community composition, yet little is known about the traits of most microbes. We describe Traitar, the microbial trait analyzer, which is a fully automated software package for deriving phenotypes from the genome sequence. Traitar provides phenotype classifiers to predict 67 traits related to the use of various substrates as carbon and energy sources, oxygen requirement, morphology, antibiotic susceptibility, proteolysis and enzymatic activities. Furthermore, it suggests protein families associated with the presence of particular phenotypes. Our method uses L1-regularized L2-loss support vector machines for phenotype assignments based on phyletic patterns of protein families and their evolutionary histories across a diverse set of microbial species. We demonstrate reliable phenotype assignment for Traitar to bacterial genomes from 572 species of 8 phyla, also based on incomplete single-cell genomes and simulated draft genomes. We also showcase its application in metagenomics by verifying and complementing a manual metabolic reconstruction of two novel Clostridiales species based on draft genomes recovered from commercial biogas reactors. Traitar is available at https://github.com/hzi-bifo/traitar.

Introduction

Microbes are often characterized and distinguished by their traits, for instance, in *Bergey's Manual of Systematic Bacteriology* (Goodfellow et al., 2012). A trait or phenotype can vary in complexity; for example, it can refer to the degradation of a specific substrate or the activity of an enzyme inferred in a lab assay, the respiratory mode of an organism, the reaction to Gram staining or antibiotic resistances. Traits are also likely driving factor for microbial community composition (Martiny et al., 2015). Microbial community members with varying metabolic capabilities can aid in waste water treatment, bioremediation of soils and promotion of plant growth (Bai et al., 2015; Narihiro and Sekiguchi, 2007; Olapade and Ronk, 2015); in the cow rumen microbiota, bacterial cellulose degraders influence the ability to process plant biomass material (Hess et al., 2011). In the Tammar wallaby foregut microbiome, the dominant bacterial species is implicated in the lower methane emissions produced by wallaby compared to ruminants (Pope et al., 2011).

In addition to the exponential growth of available sequenced microbial genome isolates, metagenome and single cell genome sequencing further contributes to the increasing number of available genomes. For the recovery of genomes from metagenomes (GFMs), computational methods based on e.g. differential read coverage and *k*-mer usage were developed (Alneberg et al., 2014; Cleary et al., 2015; Gregor et al., 2016; Imelfort et al., 2014; Kang et al., 2015; Nielsen et al., 2014), which allow to recover genomes without the need to obtain microbial isolates in pure cultures (Brown et al., 2015; Hess et al., 2011). In addition, single-cell genomics provides another culture-independent analysis technique and also allows, although often fragmented, genome recovery for less abundant taxa in microbial communities (Lasken and McLean, 2014; Rinke et al., 2013). Together, these

developments profoundly shift the analytical bottleneck from data generation to interpretation.

The genotype–phenotype relationships for some microbial traits have been well studied. For instance, bacterial motility is attributed to the proteins of the flagellar apparatus (Macnab, 2003). We have recently shown that delineating such relationships from microbial genomes and accompanying phenotype information with statistical learning methods enables the accurate prediction of the plant biomass degradation phenotype and the *de novo* discovery of both known and novel protein families that are relevant for the realization of the plant biomass degradation phenotype (Konietzny et al., 2014; Weimann et al., 2013). However, a fully automated software framework for prediction of a broad range of traits from only the genome sequence is currently missing. Additionally, horizontal gene transfer, a common phenomenon across bacterial genomes, has not been utilized to improve trait prediction so far. Traits with their causative genes may be transferred from one bacterium to the other (Ochman et al., 2000; Pal et al., 2005) (e.g. for antibiotic resistances (Martinez, 2008)) and the vertically transferred part of a bacterial genome might be unrelated to the traits under investigation (Barker and Pagel, 2005; Harvey and Pagel, 1991; Martiny et al., 2015).

Here we present Traitar, the microbial trait analyzer: an easy-to-use, fully automated software framework for the accurate prediction of currently 67 phenotypes directly from the genome sequence (Figure 1).

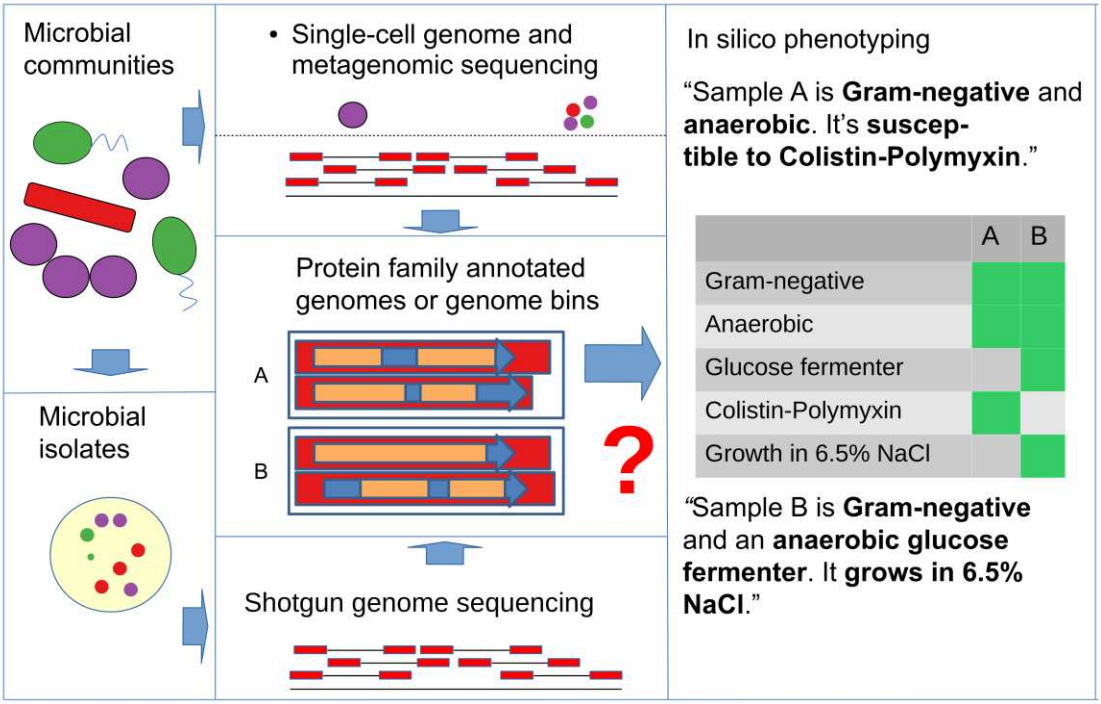


Figure 1: Traitar can be used to phenotype microbial community members based on genomes recovered from single-cell sequencing or (metagenomic) environmental shotgun sequencing data or of microbial isolates. Traitar provides classification models based on protein family annotation for a wide variety of different phenotypes related to the use of various substrates as source of carbon and energy for growth, oxygen requirement, morphology, antibiotic susceptibility and enzymatic activity.

We used phenotype data from the microbiology section of the Global Infectious Disease and Epidemiology Network (GIDEON) – a resource dedicated to the diagnosis, treatment and teaching of infectious diseases and microbiology (Berger, 2005) – for training phenotype classification models on the protein family annotation of a large number of sequenced genomes of microbial isolates (predominantly bacterial pathogens). We investigated the effect of incorporating ancestral protein family gain and losses into the model inference on classification performance, to allow consideration of horizontal gene transfer events in inference of phenotype-related protein families and phenotype classification. We rigorously tested the performance of our software in cross-validation experiments, on further test data sets and for different taxonomic ranks. To test Traitar's applicability beyond the bacteria

represented in GIDEON, we subsequently applied it to several hundred bacteria described in Bergey's systematic bacteriology (Goodfellow et al., 2012). We used Traitar to phenotype bacterial single amplified genomes (SAGs) and simulated incomplete genomes to investigate its potential for phenotyping microbial samples with incomplete genome sequences. We characterized two novel Clostridiales species of a biogas reactor community with Traitar, based on their genomes recovered with metagenomics. This verified and complemented a manual metabolic reconstruction. As Traitar furthermore suggests protein families associated with the presence of a particular phenotype, we discuss the protein families Traitar identified for several phenotypes, namely for 'Motility', 'Nitrate to nitrite' conversion and 'L-arabinose' fermentation.

Traitar is implemented in Python 2.7. It is freely available under the open-source GPL 3.0 license at https://github.com/hzi-bifo/traitar and as a Docker container at https://hub.docker.com/r/aweimann/traitar. A Traitar web service can be accessed at https://research.bifo.helmholtz-hzi.de/traitar.

Results

The Traitar software

We begin with a description of the Traitar software and phenotype classifiers. Traitar predicts the presence or absence of a phenotype, i.e. assigns a phenotype label, for 67 microbial traits to every input sequence sample (Table 1, Supplementary Table 1). For each of these traits, Traitar furthermore suggests candidate protein families associated with its realization, which can be subject of experimental follow-up studies.

For phenotype prediction, Traitar uses one of two different classification models. We trained the first classifier – the phypat classifier – on the protein and phenotype presence & absence labels from 234 bacterial species (Methods – Phenotype models). The second classifier – the phypat+PGL classifier – was trained using the same data and additionally information on evolutionary protein family and phenotype gains and losses. The latter were determined using maximum likelihood inference of their ancestral character states on the species phylogeny (Methods – Ancestral protein family and phenotype gains and losses).

The input to Traitar is either a nucleotide sequence FASTA file for every sample, which is run through gene prediction software, or a protein sequence FASTA file. Traitar then annotates the proteins with protein families. Subsequently, it predicts the presence or absence of each of the 67 traits for every input sequence. Note that Traitar doesn't require a phylogenetic tree for the input samples.

Finally, it associates the predicted phenotypes with the protein families that contributed to these predictions (Figure 2). A parallel execution of Traitar is supported by GNU parallel (Tange, 2011). The Traitar annotation procedure and the training of the phenotype models are described in more detail below (Methods – Traitar software).

Table 1: The 67 traits available in Traitar for phenotyping. We grouped each of these phenotypes into a microbiological or biochemical category.

Phenotype[(a)]	Category[(b)]
Alkaline phosphatase	
Beta hemolysis	
Coagulase production	
Lipase	Enzyme
Nitrate to nitrite	
Nitrite to gas	
Pyrrolidonyl-beta-naphthylamide	

Bile-susceptible	Growth
Colistin-Polymyxin susceptible	
DNase	
Growth at 42°C	
Growth in 6.5% NaCl	
Growth in KCN	
Growth on MacConkey agar	
Growth on ordinary blood agar	
Mucate utilization	
Arginine dihydrolase	Growth: Amino Acid
Indole	
Lysine decarboxylase	
Ornithine decarboxylase	
Acetate utilization	Growth:Carboxylic Acid
Citrate	
Malonate	
Tartrate utilization	
Gas from glucose	Growth:Glucose
Glucose fermenter	
Glucose oxidizer	
Methyl red	
Voges Proskauer	
Cellobiose	Growth:Sugar
D-Mannitol	
D-Mannose	
D-Sorbitol	
D-Xylose	
Esculin hydrolysis	
Glycerol	
Lactose	
L-Arabinose	
L-Rhamnose	
Maltose	
Melibiose	
myo-Inositol	
ONPG (beta galactosidase)$_{(d)}$	
Raffinose	
Salicin	
Starch hydrolysis	
Sucrose	
Trehalose	
Urea hydrolysis	
Bacillus or coccobacillus	Morphology
Coccus	
Coccus - clusters or groups predominate	
Coccus - pairs or chains predominate	

Gram negative	
Gram positive	
Motile	
Spore formation	
Yellow pigment	
Aerobe	Oxygen
Anaerobe	
Capnophilic	
Facultative	
Catalase	Oxygen:Enzyme
Oxidase	
Hydrogen sulfide	Product
Casein hydrolysis	Proteolysis
Gelatin hydrolysis	

(a) GIDEON phenotypes with at least 10 presence and 10 absence labels

(b) Phenotypes assigned to microbiological / biochemical categories

(c) ONPG: o-Nitrophenyl-β-D-galatopyranosid

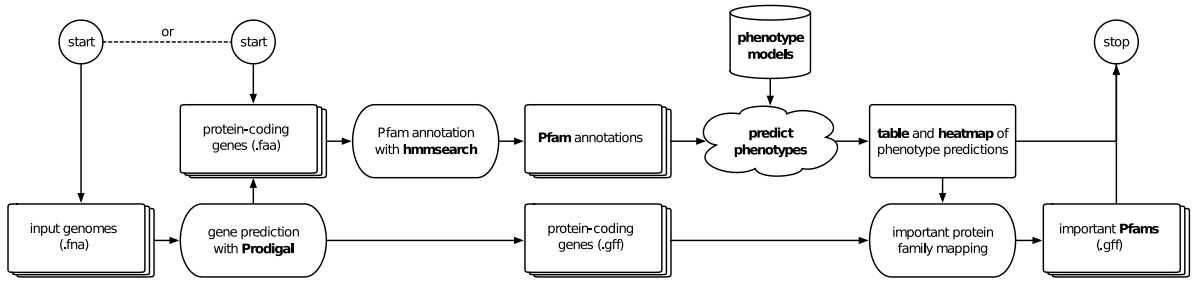


Figure 2: Work flow of Traitar. Input to the software can be genome sequene samples in nucleotide or amino acid FASTA format. Traitar predicts phenotypes based on pre-computed classification models and provides graphical and tabular output. In the case of nucleotide input, the protein families that are important for the phenotype predictions will be further mapped to the predicted protein-coding genes.

Evaluation

We evaluated the two Traitar classifiers using ten-fold nested cross-validation on 234 bacterial species found in GIDEON (GIDEON I). The determined macro-accuracy (the accuracy balanced over all phenotypes) for the 67 GIDEON phenotypes was 82.6% for the phypat classifier and 85.5% for the phypat+PGL classifier; the accuracy (fraction of correct assignments averaged over all tested samples) for

phypat was 88.1%, in comparison to 89.8% for phypat+PGL (Methods – Evaluation metrics; Table 2). Notably, Traitar classified 53 phenotypes with more than 80% macro-accuracy and 26 phenotypes with at least 90% macro-accuracy with one of the two classifiers (Figure 3, Supplementary Table 2). Phenotypes that could be predicted with very high confidence included the outcome of a 'Methyl red' test, 'Spore formation', oxygen requirement (i.e. 'Anaerobe' and 'Aerobe'), 'Growth on MacConkey agar' or 'Catalase'. Some phenotypes proved to be difficult to predict (60-70% macro-accuracy), which included 'DNAse', 'myo-Inositol' or 'Yellow pigment' and 'Tartrate utilization', regardless of which classifier was used. This might be caused by the relatively small number (<20) of positive (phenotype present) examples that were available.

Table 2: We evaluated the Traitar phypat and phypat+PGL phenotype classifiers and a consensus vote of both classifiers for 234 bacteria described in the Global Infectious Disease and Epidemiology Online Network (GIDEON) in a 10-fold nested cross-validation using different evaluation measures (Methods – Evaluation). Subsequently, we tested another 42 bacteria from GIDEON and 296 bacteria described in Bergey's manual of systematic bacteriology for an independent performance assessment of the two classifiers.

Data set (# bacteria)	Classifier	Macro-accuracy	Accuracy	Recall Phenotype+	Recall Phenotype-
GIDEON I (234)	phypat	82.6	88.1	86.1	91.4
	phypat+PGL	**85.5**	**89.8**	**87.8**	90.9
	consensus	83.0	88.8	82.2	**95.4**
GIDEON II (42)	Phypat	85.3	87.5	84.9	90.2
	phypat+PGL	**86.7**	**87.9**	**86.3**	89.7
	consensus	85.7	87.2	80.8	**93.7**
Bergey's (296)	phypat	NA[1]	**72.9**	**74.6**	71.2
	phypat+PGL	NA[1]	72.4	74	70.8
	consensus	NA[1]	**72.9**	66.6	**79.2**

1 We only report the overall accuracy, as insufficient phenotype labels (less than 5 with a negative and positive label, respectively) were available for several phenotypes, to enable a comparable macro-accuracy calculation to the other data sets (Supplementary Table 1).

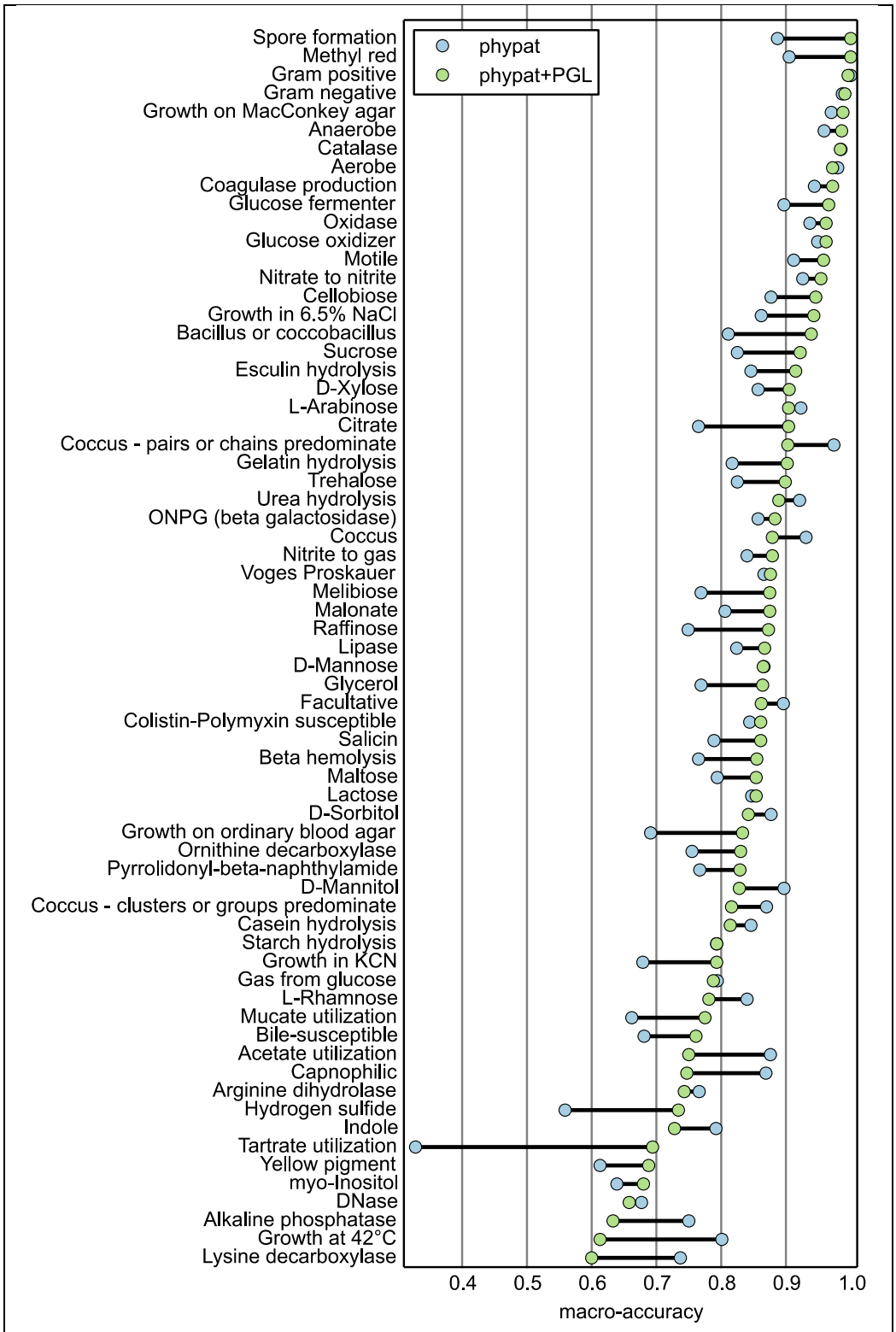


Figure 3: Macro-accuracy for each phenotype for the Traitar phypat and phypat+PGL phenotype classifiers determined in nested cross-validation on 234 bacterial species described in the Global Infectious Disease and Epidemiology Online Network (Methods – Evaluation metrics, Table 1, Supplementary Table 1).

For an independent assessment of Traitar's classification performance we next tested Traitar on 42 bacterial species that had phenotype information available in GIDEON (GIDEON II), but were not used for learning the phenotype models (The Traitar software – Annotation). For calculation of the macro-accuracy, we considered only phenotypes represented by at least five phenotype-positive and five phenotype-negative bacteria. On these data, Traitar predicted the phenotypes with a macro-accuracy of 85.3% with the phypat classifier and 86.7% with the phypat+PGL classifier, and accuracies of 87.5% and 87.9%, respectively (Table 2). To investigate the performance of Traitar for bacterial genomes from a different data source, we next determined from two volumes of Bergey's Manual of Systematic Bacteriology, namly 'The Proteobacteria' and 'The Firmicutes', the phenotypes of further sequenced bacteria that were not in our GIDEON I and II data sets (Supplementary Table 1, 4). In total, we thus identified phenotypes for another 296 sequenced bacterial species (The Traitar software – Annotation). Also for these bacteria, Traitar performed well but was less reliable than before, with accuracies for the phypat classifier of 72.9% and 72.1% for the phypat+PGL classifier (Table 2). This is likely due to the taxonomic differences of bacteria listed in GIDEON and Bergey's and also because most of the bacteria in Bergey's have only draft genomes available for phenotyping.

When combining the predictions of the phypat and phypat+PGL classifiers into a consensus vote, Traitar assigns phenotypes more reliably, while predicting less phenotype labels compared to the individuals classifiers (Table 2). Depending on the use case, Traitar can be used with performance characterized by different trade-offs between the recall of the phenotype-positive and the phenotype-negative classes.

Performance per taxon at different ranks of the taxonomy

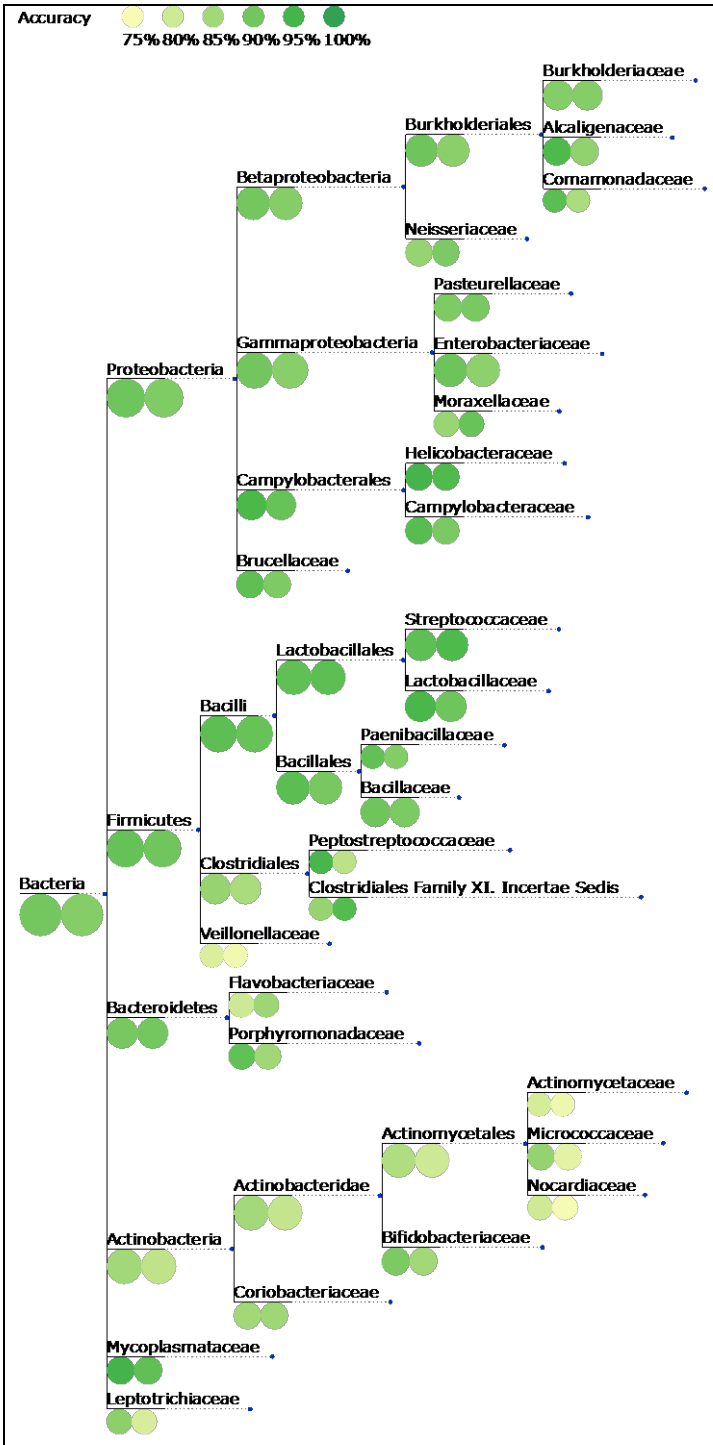


Figure 4: Classification accuracy for each taxon at different ranks of the NCBI taxonomy. For better visualization of names for the internal nodes, the taxon names are displayed on branches leading to the respective taxon node in the tree. The nested cross-validation accuracy obtained with Traitar for 234 bacterial species described in the Global Infectious Disease and Epidemiology Online Network was projected onto the NCBI taxonomy down to the family level. Colored circles at the tree nodes depict the performance of the phypat+PGL classifier (left-hand circles) and the phypat classifier (right-hand circles). The size of the circles reflects the number of species per taxon.

We investigated the performance of Traitar across the part of the bacterial tree of life represented in our data set. For this purpose, we evaluated the nested cross-validation performance of the phypat and phypat+PGL classifiers at different ranks of the NCBI taxonomy. For a given GIDEON taxon, we pooled all bacterial species that are descendants of this taxon. Figure 4 shows the accuracy estimates projected on the NCBI taxonomy from the domain level down to individual families. Notably, the accuracy of the phypat+PGL (phypat) classifier for the phyla covered by at least five bacterial species showed low variance and was high across all phyla, ranging from 84% (81%) for Actinobacteria over 90% (89%) for Bacteroidetes, 89% (90%) for Proteobacteria, 91% (90%) for Firmicutes to 91% (86%) for Tenericutes.

Phenotyping incomplete genomes

GFMs or SAGs are often incomplete and thus we analyzed the effect of missing genome assembly parts onto the performance of Traitar. Rinke *et al.* used a single-cell sequencing approach to analyze poorly characterized parts of the bacterial and archaeal tree of life, the so-called 'microbial dark matter' (Rinke et al., 2013). They pooled 20 SAGs from the candidate phylum Cloacimonetes, formerly known as WWE1, to generate joint – more complete – genome assemblies that had at least a genome-wide average nucleotide identity of 97% and belonged to a single 16S-based operational taxonomic unit, namely *Cloacamonas acidaminovorans (Pelletier et al., 2008)*.

According to our predictions based on the joint assembly of the single-cell genomes, *C. acidaminovorans* is Gram-negative and is adapted to an anaerobic lifestyle, which agrees with the description by Rinke *et al.* (Figure 5). Traitar further predicted 'Arginine dihydrolase' activity, which is in line with the characterization of the species as an amino acid degrader (Rinke et al., 2013). Remarkably, the prediction of a bacil-

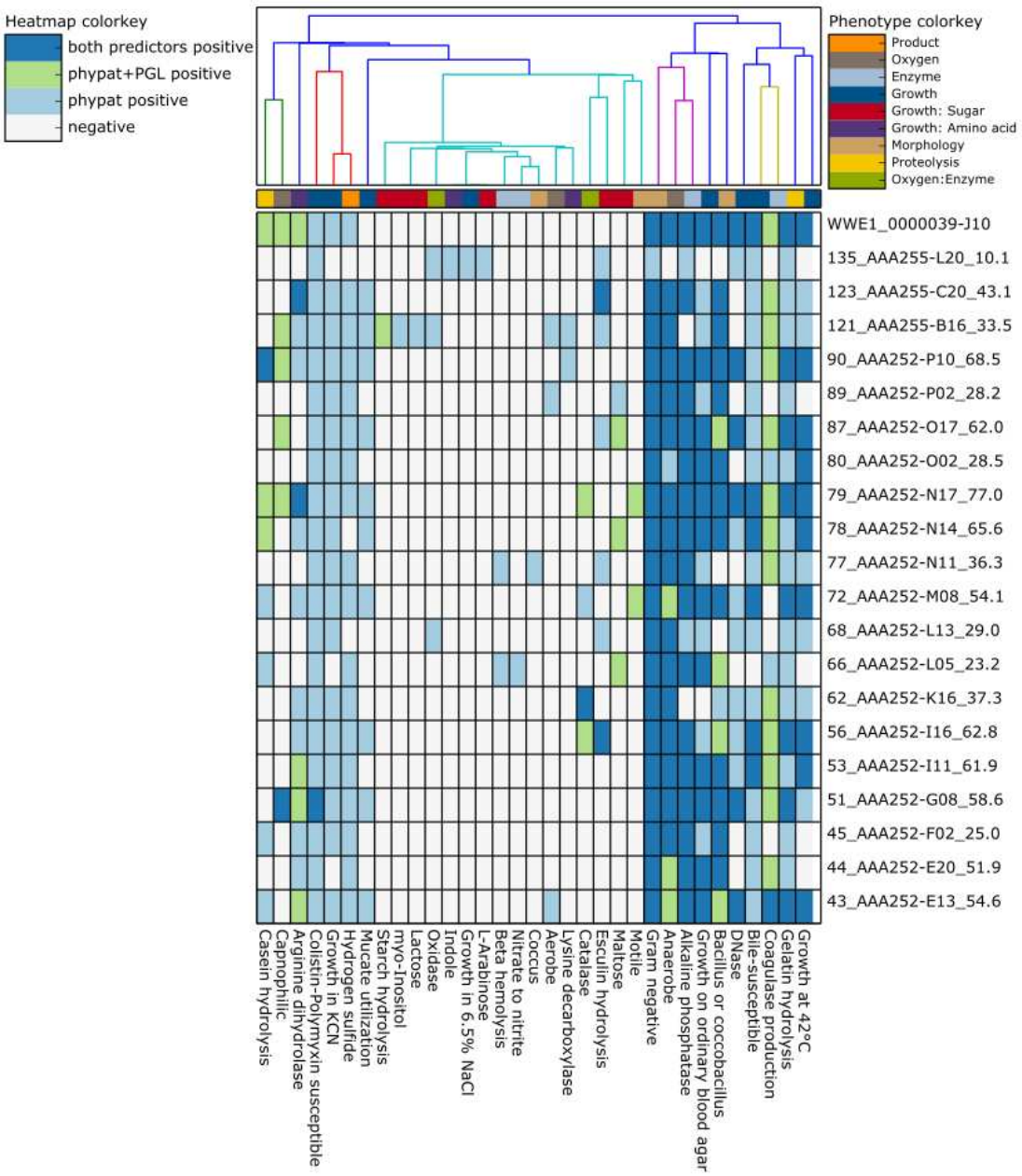


Figure 5: Single-cell phenotyping with Traitar. We used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with Traitar. Shown is a heatmap of assembly samples vs. phenotypes, which is the standard visualization for phenotype predictions in Traitar. The origin of the phenotype's prediction (Traitar phypat and/or Traitar phypat+PGL classifier) determines the color of the heatmap entries. The sample labels have their genome completeness estimates as suffixes. The colors of the dendrogram indicate similar phenotype distributions across samples, as determined by a hierarchical clustering with SciPy[1].

[1] http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

lus or coco-bacillus shape agrees with the results of Limam *et al.* (Limam et al., 2014), who used a WWE1-specific probe and characterized the samples with fluorescence *in situ* hybridization. They furthermore report that members of the Cloacimonetes candidate phylum are implicated in anaerobic digestion of cellulose, primarily in early hydrolysis, which is in line with the very limited carbohydrate degradation spectrum found by Traitar.

Subsequently, we compared the predicted phenotypes for the SAGs to the predictions for the joint assembly. The phypat classifier recalled more of the phenotype predictions of the joint assembly based on the SAGs than the phypat+PGL classifier. However, the phypat+PGL classifier made fewer false positive predictions (Figure 6 a).

In the next experiment, we inferred phenotypes based on simulated GFMs, by subsampling from the coding sequences of each of the 42 bacterial genomes (GIDEON II). Starting with the complete set of coding sequences we randomly deleted genes from the genomes. For the obtained draft genomes with different degrees of completeness, we re-ran the Traitar classification and computed the accuracy measures, as before. We observed that the average fraction of phenotypes identified (macro-recall for the positive class) of the phypat+PGL classifier dropped more quickly with more missing coding sequences than that of the phypat classifier (Figure 6 b). However, at the same time, the recall of the negative class of the phypat+PGL classifier improved with a decreasing number of coding sequences, meaning that fewer but more reliable predictions were made.

Overall, the tradeoffs in the recall of the phenotype-positive and the phenotype-negative classes of the two classifiers resulted in a similar overall macro-accuracy across the range of tested genome completeness.

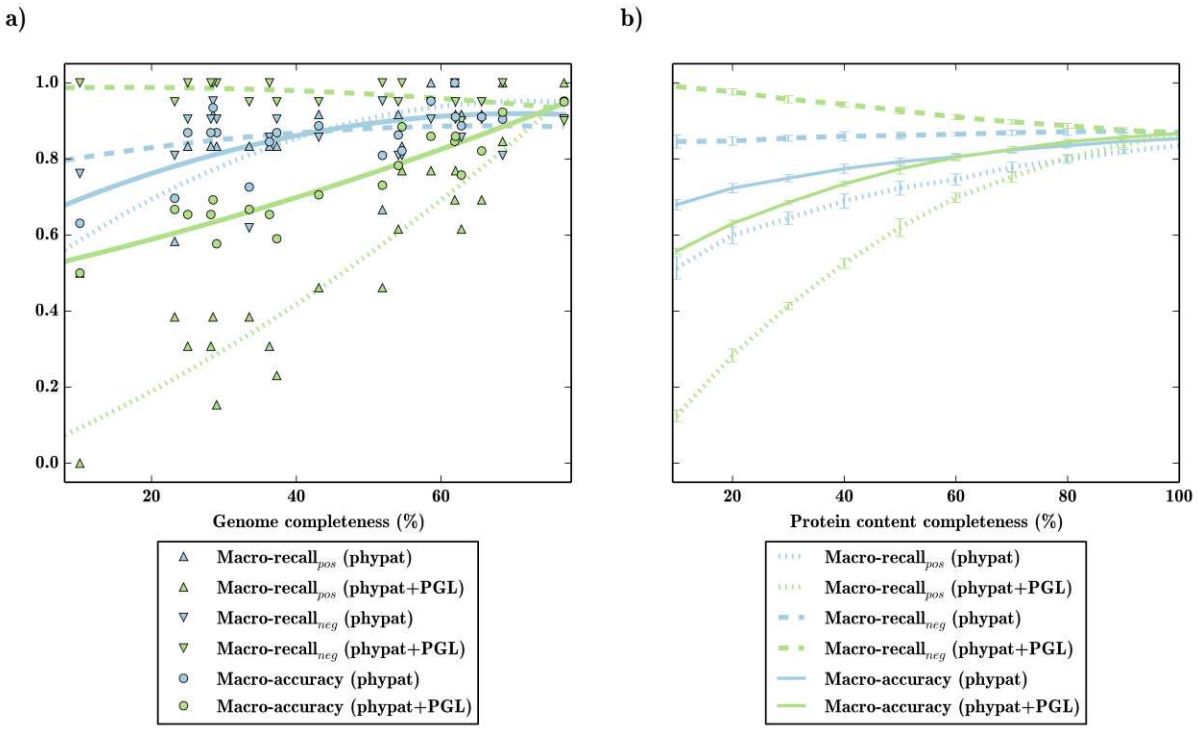


Figure 6: Phenotyping simulated draft genomes and single cell genomes. In (a) we used 20 genome assemblies with varying degrees of completeness from single cells of the Cloacimonetes candidate phylum and a joint assembly for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. Shown is the performance of the phenotype prediction vs. the genome completeness of the single cells with respect to the joint assembly. In (b) we simulated draft genomes based on an independent test set of 42 microbial (pan)genomes. The coding sequences of these genomes were down-sampled (10 replications per sampling point) and the resulting simulated draft genomes were used for phenotyping with the Traitar phypat and the Traitar phypat+PGL classifiers. We plotted various performance estimates (mean center values and and s.d. error bars shown) against the protein content completeness.

Thus, depending on the intended usage, a particular classifier can be chosen: we expect that the reliable predictions inferred with the phypat+PGL classifier and the more abundant, but less reliable predictions made with the phypat classifier will complement one another in different use cases for partial genomes recovered from metagenomic data.

By analyzing the protein families with assigned weights and the bias terms of the two classifiers, we found the phypat+PGL classifier to base its predictions primarily on the presence of protein families that were typical for the phenotypes. In contrast, the

phypat classifier also took typically absent protein families from phenotype-positive genomes into account in its decision. More technically, the positive weights in models of the phypat classifier are balanced out by negative weights, whereas for the phypat+PGL classifier, they are balanced out by the bias term. By down-weighting the bias term for the phypat+PGL classifier by the protein content completeness, we could show that the accuracy of the phypat classifier could be increased over that of the phypat+PGL, regardless of the protein content completeness (data not shown). However, this requires knowledge of the protein content completeness for each genomic sample, which could be indirectly estimated using methods such as checkM (Parks et al., 2015).

Traitar as a resource for gene target discovery

In addition to phenotype assignment, Traitar suggests the protein families relevant for the assignment of a phenotype (Methods – Majority feature selection, Table 3). We exemplarily demonstrate this capability here for three phenotypes that are already well-studied, namely 'Motile', 'Nitrate to nitrite' conversion and 'L-arabinose' metabolism. These phenotypes represent one each from the phenotype categories morphology, enzymatic activity and growth on sugar.

In general, we observed that the protein families important for classification can be seen to be gained and lost jointly with the respective phenotypes within the microbial phylogeny. Among the selected Pfam families that are important for classifying the motility phenotype were proteins of the flagellar apparatus and chemotaxis-related proteins (Table 3). Motility allows bacteria to colonize their preferred environmental niches. Genetically, it is mainly attributed to the flagellum, which is a molecular motor, and is closely related to chemotaxis, a process that lets bacteria sense chemicals in their surroundings. Motility also plays a role in bacterial pathogenicity, as it enables

bacteria to establish and maintain an infection. For example, pathogens can use flagella to adhere to their host and they have been reported to be less virulent if they lack flagella (Josenhans and Suerbaum, 2002). Of 48 flagellar proteins described in (Liu and Ochman, 2007), four proteins (FliS, MotB, FlgD and FliJ) were sufficient for accurate classification of the motility phenotype and were selected by our classifier, as well as FlaE, which was not included in this collection. FliS (PF02561) is a known export chaperone that inhibits early polymerization of the flagellar filament FliC in the cytosol (Lam et al., 2010). MotB (PF13677), part of the membrane proton-channel complex, acts as the stator of the bacterial flagellar motor (Hosking et al., 2006). Traitar also identified further protein families related to chemotaxis, such as CZB (PF13682), a family of chemoreceptor zinc-binding domains found in many bacterial signal transduction proteins involved in chemotaxis and motility (Draper et al., 2011), and the P2 response regulator-binding domain (PF07194). The latter is connected to the chemotaxis kinase CheA and is thought to enhance the phosphorylation signal of the signaling complex (Dutta et al., 1999).

Nitrogen reduction in nitrate to nitrite conversion is an important step of the nitrogen cycle and has a major impact on agriculture and public health. Two types of nitrate reductases are found in bacteria: the membrane-bound Nar and the periplasmic Nap nitrate reductase (Moreno-Vivian et al., 1999), which we found both to be relevant for the classification of the phenotype: we identified all subunits of the Nar complex as being relevant for the ‘Nitrate to nitrite’ conversion phenotype (i.e. the gamma and delta subunit (PF02665, PF02613)), as well as Fer4_11 (PF13247), which is in the iron–sulfur center of the beta subunit of Nar. The delta subunit is involved in the assembly of the Nar complex and is essential for its stability, but probably is not directly part of it (Pantel et al., 1998). Traitar also identified the Molybdopterin oxido-

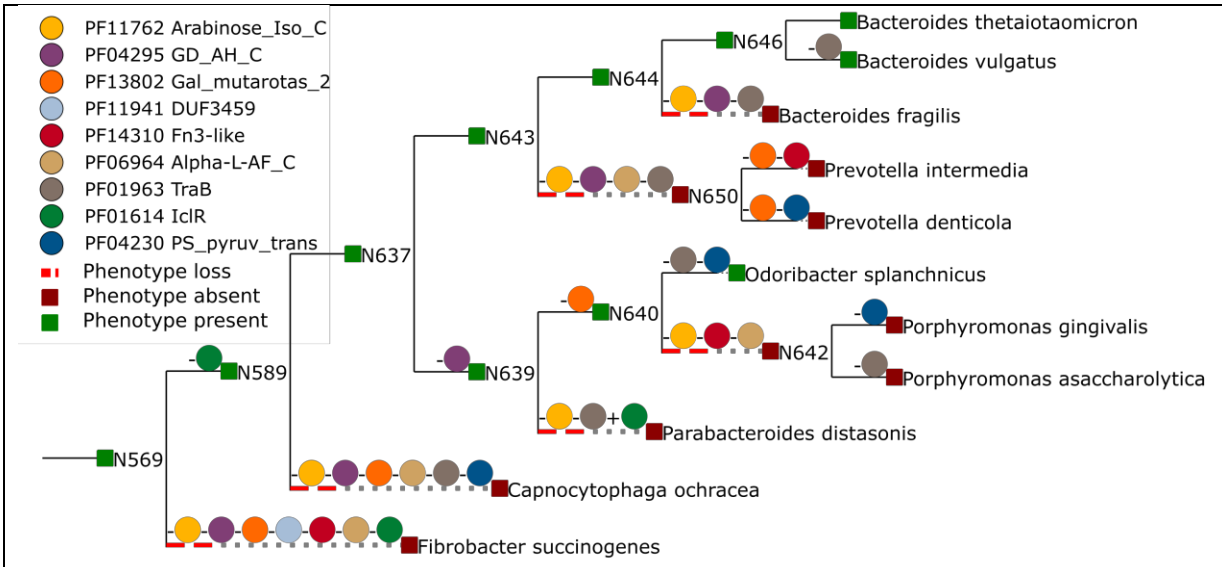


Figure 7: Phenotype gain and loss dynamics match protein family dynamics. We show the phenotype–protein family gain and loss dynamics for families identified as important by Traitar for the L-arabinose phenotype. Signed colored circles along the tree branches depict protein family gains (+) or losses (-). Taxon nodes are colored according to their inferred (ancestral) phenotype state.

reductase Fe4S4 domain (PF04879), which is bound to the alpha subunit of the nitrate reductase complex (Pantel et al., 1998). Traitor furthermore suggested NapB (PF03892) as relevant, which is a subunit of the periplasmic Nap protein and NapD (PF03927), which is an uncharacterized protein implicated in forming Nap (Moreno-Vivian et al., 1999).

Table 3: The most relevant Pfam families for classification of three important phenotypes: 'Nitrate to Nitrite', 'Motility' and 'L-Arabinose'. We ranked the Pfam families with positive weights in the Traitar SVM classifiers by the correlation of the Pfam families with the respective phenotype labels across 234 bacteria described in the Global Infectious Disease and Epidemiology Online Network. Shown are the 10 highest ranking Pfam families along with their descriptions and a description of their phenotype-related function, where we found one.

Accession	Phenotype	Pfam description	Remarks
PF13677	Motile	Membrane MotB of proton-channel complex MotA/MotB	Flagellar protein
PF03963	Motile	Flagellar hook capping protein N-terminal region	Flagellar protein
PF02561	Motile	Flagellar protein FliS	Flagellar protein
PF02050	Motile	Flagellar FliJ protein	Flagellar protein
PF07559	Motile	Flagellar basal body protein FlaE	Flagellar protein
PF13682	Motile	Chemoreceptor zinc-binding domain	Chemotaxis-related
PF03350	Motile	Uncharacterized protein family, UPF0114	

PF05226	Motile	CHASE2 domain	Chemotaxis-related
PF07194	Motile	P2 response regulator binding domain	Chemotaxis-related
PF04982	Motile	HPP family	
PF03927	Nitrate to nitrite	NapD protein	Involved in Nar formation
PF13247	Nitrate to nitrite	4Fe-4S dicluster domain	Iron-sulfur cluster center of the beta subunit of Nar
PF03892	Nitrate to nitrite	Nitrate reductase cytochrome c-type subunit (NapB)	Periplasmic Nap subunit
PF02613	Nitrate to nitrite	Nitrate reductase delta subunit	Nap subunit
PF01127	Nitrate to nitrite	Succinate dehydrogenase/Fumarate reductase transmembrane subunit	
PF01292	Nitrate to nitrite	Prokaryotic cytochrome b561	
PF03459	Nitrate to nitrite	TOBE domain	
PF03824	Nitrate to nitrite	High-affinity nickel transport protein	
PF04879	Nitrate to nitrite	Molybdopterin oxidoreductase Fe4S4 domain	Bound to the alpha subunit of Nar
PF02665	Nitrate to nitrite	Nitrate reductase gamma subunit	Nar subunit
PF11762	L-Arabinose	L-arabinose isomerase C-terminal domain	Catalyzes first reaction in L-arabinose metabolism
PF04295	L-Arabinose	D-galactarate dehydratase / Altronate hydrolase, C terminus	
PF13802	L-Arabinose	Galactose mutarotase-like	
PF11941	L-Arabinose	Domain of unknown function (DUF3459)	
PF14310	L-Arabinose	Fibronectin type III-like domain	
PF06964	L-Arabinose	Alpha-L-arabinofuranosidase C-terminus	Acts on L-arabinose side chains in pectins
PF01963	L-Arabinose	TraB family	
PF01614	L-Arabinose	Bacterial transcriptional regulator	
PF06276	L-Arabinose	Ferric iron reductase FhuF-like transporter	
PF04230	L-Arabinose	Polysaccharide pyruvyl transferase	

L-arabinose is major constituent of plant polysaccharides, which is located, for instance, in pectin side chains and is an important microbial carbon source (Martinez et al., 2008). Traitar identified the L-arabinose isomerase *C*-terminal domain (PF11762), which catalyzes the first step in L-arabinose metabolism – the conversion of L-arabinose into L-ribulose (Sa-Nogueira et al., 1997), as being important for realizing the L-arabinose metabolism. It furthermore suggested the *C*-terminal domain of Alpha-L-arabinofuranosidase (PF06964), which cleaves nonreducing terminal alpha-L-arabinofuranosidic linkages in L-arabinose-containing polysaccharides (Gilead and Shoham, 1995) and is also part of the well-studied L-arabinose operon in *Escherichia coli* (Sa-Nogueira et al., 1997).

Phenotyping biogas reactor population genomes

We used Traitar to phenotype two novel Clostridiales species (unClos_1, unFirm_1) based on their genomic information reconstructed from metagenome samples. These were taken from a commercial biogas reactor operating with municipal waste (Frank et al., 2015). The genomes of unClos_1 and unFirm_1 were estimated to be 91% complete and 60% complete based on contigs ≥5 kb, respectively. Traitar predicted unClos_1 to utilize a broader spectrum of carbohydrates than unFirm_1 (Table 4). We cross-referenced our predictions with a metabolic reconstruction conducted by Frank *et al.* (under review; supplementary material). We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. The manual reconstruction and predictions inferred with Traitar agreed to a great extent (Table 4). Traitar recalled 87.5% (6/7) of the phenotypes inferred via the metabolic reconstruction and also agreed to 81.8% (9/11) on the absent phenotypes. Notable exceptions were that Traitar only found a weak signal for 'D-xylose' utilization. A weak signal means that only a minority of the classifiers in the voting committee assigned these samples to the phenotype-positive class (Methods – Phenotype models). However, the metabolic reconstruction was also inconclusive with respect to xylose fermentation. Furthermore, Traitar only found a weak signal for 'Glucose fermentation' for unFirm_1. Whilst genomic analysis of unFirm_1 revealed the Embden–Meyerhof–Parnas (EMP) pathway, which would suggest glucose fermentation, gene-centric and metaproteomic analysis of this phylotype indicated that the EMP pathway was probably employed in an anabolic direction (gluconeogenesis); therefore unFirm_1 is also unlikely to ferment D-Mannose. This suggests that unFirm_1 is unlikely to ferment sugars and instead metabolizes acetate (also predicted by Traitar, Table 4) via a syntrophic interaction with hydrogen-utilizing methanogens.

Traitar predicted further phenotypes for both species that were not targeted by the manual reconstruction. One of these predictions was an anaerobic lifestyle, which is likely to be accurate, as the genomes were isolated from an anaerobic bioreactor environment. It also predicted them to be Gram-positive, which is probably correct, as the Gram-positive sortase protein family can be found in both genomes.

Table 4 Phenotype predictions for two novel Clostridiales species with genomes reconstructed from a commercial biogas reactor metagenome. Traitar output (yes, no, weak) was cross-referenced with phenotypes manually reconstructed based on Kyoto Encyclopedia of Genes and Genomes orthology annotation (Frank *et al.* submitted; supplementary material), which are primarily the fermentation phenotypes of various sugars. We considered all phenotype predictions that Traitar inferred with either the phypat or the phypat+PGL classifier. A weak prediction means that only a minority of the classifiers in the Traitar voting committee assigned this sample to the phenotype-positive class (Traitar phenotype). Table entries colored in red show a difference between the prediction and the reconstruction, whereas green denotes an overlap; yellow is inconclusive.

	unClos_1	unFirm_1
Glucose	yes	weak
Acetate utilization	no	yes
Mannitol	yes	no
Starch hydrolysis	no	no
Xylose	weak	no
L-Arabinose	yes	no
Capnophilic	yes	no
Sucrose	yes	no
D-Mannose	yes	no
Maltose	yes	no
Arginine dihydrolase	no	yes

This is a Gram-positive biomarker (Paterson and Mitchell, 2004). Furthermore, all Firmicutes known so far are Gram-positive (Goodfellow et al., 2012). Additionally, Traitar assigned ‘Motile’ and ‘Spore formation’ to unFirm_1, based on the presence of several flagellar proteins (e.g. FliM, MotB, FliS and FliJ) and the sporulation proteins CoatF and YunB.

Discussion

We have developed Traitar, a software framework for predicting phenotypes from the protein family profiles of bacterial genomes. Traitar provides a quick and fully automated way of assigning 67 different phenotypes to bacteria based on the protein family content of their genomes.

Microbial trait prediction from phyletic patterns has been proposed in previous studies for a limited number of phenotypes (Feldbauer et al., 2015; Kastenmuller et al., 2009; Konietzny et al., 2014; Lingner et al., 2010; MacDonald and Beiko, 2010; Weimann et al., 2013). To our knowledge, the only currently available software for microbial genotype-phenotype inference is PICA, which is based on learning associations of clusters of orthologous genes (Tatusov et al., 2001) with traits (MacDonald and Beiko, 2010). Recently, PICA was extended by Feldbauer *et al.* for predicting eleven traits overall, optimized for large datasets and tested on incomplete genomes (Feldbauer et al., 2015). Traitar allows prediction of 67 phenotypes, including 60 entirely novel ones. It furthermore includes different prediction modes, one based on phyletic patterns, one additionally including a statistical model of protein family evolution for its predictions. Traitar also suggest associations between phenotypes and protein families. For three traits, we showed that several of these associations are to known key families of establishment of a particular trait, and that furthermore candidate families were suggested, that might serve as targets for experimental studies. Some of the phenotypes annotated in GIDEON are specific for the human habitat (such as ‘coagulase production’ or ‘growth on ordinary blood agar’) and the genetic underpinnings learned by Traitar could be interesting to study for infection disease research.

In cross-validation experiments with phenotype data from the GIDEON database, we showed that the Traitar phypat classifier has high accuracy in phenotyping bacterial samples. Considering ancestral protein family gains and losses in the classification, which is implemented in the Traitar phypat+PGL classifier, improves the accuracy compared to prediction from phyletic patterns only, both for individual phenotypes and overall. Barker *et al.* were first to note the phylogenetic dependence of genomic samples and how this can lead to biased conclusions (Barker and Pagel, 2005). MacDonald *et al.* selected protein families based on correlations with a phenotype and corrected for the taxonomy (MacDonald and Beiko, 2010). Here we accounted for the evolutionary history of the phenotype and the protein families in the classifier training itself to automatically improve phenotype assignment. We additionally demonstrated the reliability of the performance estimates by phenotyping, with a similar accuracy, an independent test dataset with bacteria described in GIDEON, which we did not use in the cross-validation. Traitar also reliably phenotyped a large and heterogenic collection of bacteria that we extracted from Bergey's Manual of Systematic Bacteriology – mostly with only draft genomes available. We didn't observe any bias towards specific taxa in GIDEON, but some of the phenotypes might be realized with different protein families in taxa that are less well represented indicated by the around 15% - 20% less reliable phenotyping results for bacteria described in Bergey's manual of systematic bacteriology. We expect that the accuracy of the phenotype classification models already available in Traitar will further improve the more data will become available and can be incorporated into its training.

We found that Traitar can provide reliable insights into the metabolic capabilities of microbial community members even from partial genomes, which are very common

for genomes recovered from single cells or metagenomes. One obvious limitation being for incomplete genomes, the absence of a phenotype prediction may be due to the absence of the relevant protein families from the input genomes. The analysis of both the SAGs and simulated genomes led us to the same conclusions: the phypat classifier is more suitable for exploratory analysis, as it assigned more phenotypes to incomplete genomes, at the price of more false positive predictions. In contrast, the phypat+PGL classifier assigned fewer phenotypes, but also made fewer false assignments. At the moment, genotype–phenotype inference with Traitar only takes into account the presence and absence of protein families of the bacteria analyzed. This information can be readily computed from the genomic and metagenomic data. Future research could focus also on integration of other 'omics' data to allow even more accurate phenotype assignments. Additionally, expert knowledge of the biochemical pathways that are used in manual metabolic reconstructions, for example, could be integrated as prior knowledge into the model in future studies.

For the phenotyping of novel microbial species, generating a detailed (manual) metabolic reconstruction such as the one by Frank *et al.* (submitted; supplementary material) is time-intensive. Furthermore, such reconstructions are usually focused on specific pathways and are dependent on the research question. This is not an option for studies with 10–50+ genomes, which are becoming more and more common in microbiology (Brown et al., 2015; Hess et al., 2011; Rinke et al., 2013). Traitar thus is likely to be particularly helpful for multi-genome studies. It furthermore may pick up on things outside of the original research focus and could serve as a seed or a first-pass method for a detailed metabolic reconstruction in future studies.

Methods

The Traitar software

In this section we first describe the Traitar annotation procedure. We proceed with the genome and phenotype data used for the training of Traitar phenotype models; afterwards we explain the training and illustrate how we considered ancestral protein family gains and losses in the models. Finally, we specify the requirements for running the Traitar software.

Annotation

In the case of nucleotide DNA sequence input, Traitar uses Prodigal (Hyatt et al., 2010) for gene prediction prior to Pfam family annotation. The amino acid sequences are then annotated in Traitar with protein families (Pfams) from the Pfam database (version 27.0) (Finn et al., 2014) using the hmmsearch command of HMMER 3.0 (Finn et al., 2011).

Each Pfam family has a hand-curated threshold for the bit score, which is set in such a way that no false positive is included (Punta et al., 2012). A fixed threshold of 25 is then applied to the bit score (the log-odds score) and all Pfam domain hits with an E-value above 10^{-2} are discarded. The resulting Pfam family counts (phyletic patterns) are turned into presence or absence values, as we found this representation to yield a favorable classification performance (Weimann et al., 2013).

Genome and phenotype data

We obtained our phenotype data from the GIDEON database (Berger, 2005). In GIDEON a bacterium is labeled either as phenotype-positive, -negative or strain-specific. In the latter case we discarded this phenotype label. The GIDEON traits can be grouped into the categories the use of various substrates as source of carbon and energy for growth, oxygen requirement, morphology, antibiotic susceptibility and

enzymatic activity (Table 1, Supplementary Table 1). We only considered phenotypes that were available in GIDEON for at least 20 bacteria, with a minimum of 10 bacteria annotated as positive (phenotype presence) for a given phenotype and 10 as negative (phenotype absence) to enable a robust and reliable analysis of the respective phenotypes. Furthermore, to be included in the analysis, we required each bacterial sample to have:

a) at least one annotated phenotype,
b) at least one sequenced strain,
c) a representative in the sTOL.

In total, we extracted 234 species-level bacterial samples with 67 phenotypes with sufficient total, positive and negative labels from GIDEON (GIDEON I). GIDEON associates these bacteria with 9305 individual phenotype labels, 2971 being positive and 6334 negative (Supplementary Table 1, 3). GIDEON species that had at least one sequenced strain available but were not part of the sTOL tree were set aside for a later independent assessment of the classification accuracy. In total, this additional dataset comprised further 42 unique species with 58 corresponding sequenced bacterial strains (GIDEON II, Supplementary Table 1, 4). We obtained 1836 additional phenotype labels for these bacteria, consisting of 574 positive and 1262 negative ones. We searched the Firmicutes and Proteobacteria volumes of Bergey's systematic bacteriology specifically for further bacteria not represented so far in the GIDEON data sets (Goodfellow et al., 2012). In total, we obtained phenotype data from Bergey's for 206 Firmicutes and 90 Proteobacteria with a total of 1152 positive labels and 1376 negative labels (Supplementary Table 1, 5). As in GIDEON, in Bergey's the phenotype information is usually given on the species level.

We downloaded the coding sequences of all complete bacterial genomes that were available via the NCBI FTP server under ftp://ftp.ncbi.nlm.nih.gov/genomes/ as of 11 May 2014 and genomes from the PATRIC data base as of September 2015 (Wattam et al., 2014). These were annotated with Traitar. For bacteria with more than one sequenced strain available, we chose the union of the Pfam family annotation of the single genomes to represent the pangenome Pfam family annotation, as in (Liu et al., 2006).

Phenotype models

We represented each phenotype from the set of GIDEON phenotypes across all genomes as a vector ***yp***, and solved a binary classification problem using the matrix of Pfam phyletic patterns *XP* across all genomes as input features and ***yp*** *a*s the binary target variable (Supplementary Figure 1). For classification, we relied on support vector machines (SVMs), which are a well-established machine learning method (Boser et al., 1992). Specifically, we used a linear L1-regularized L2-loss SVM for classification as implemented in the LIBLINEAR library (Fan et al., 2008). For many datasets, linear SVMs achieve comparable accuracy to SVMs with a non-linear kernel but allow faster training. The weight vector of the separating hyperplane provides a direct link to the Pfam families that are relevant for the classification. L1-regularization enables feature selection, which is useful when applied to highly correlated and high-dimensional datasets, as used in this study (Zou and Hastie, 2005). We used the interface to LIBLINEAR implemented in scikit-learn (Pedregosa et al., 2011). For classification of unseen data points – genomes without available phenotype labels supplied by the user – Traitar uses a voting committee of five SVMs with the best single cross-validation accuracy (Methods – Nested cross-validation). Traitar then assigns each unseen data point to the majority class (phenotype presence or absence class) of the voting committee.

Ancestral protein family and phenotype gains and losses

We constructed an extended classification problem by including ancestral protein family gains and losses, as well as the ancestral phenotype gains and losses in our analysis, as implemented in GLOOME (Cohen and Pupko, 2011). Barker *et al.* report that common methods for inferring functional links between genes, that do not take the phylogeny into account, suffer from high rates of false positives (Barker and Pagel, 2005). Here, we jointly derived the classification models from the observable phyletic patterns and phenotype labels, and from phylogenetically unbiased ancestral protein family and phenotype gains and losses, that we inferred via a maximum likelihood approach from the observable phyletic patterns on a phylogenetic tree, showing the relationships among the samples. (Supplementary Figure 1). Ancestral character state evolution in GLOOME is modeled via a continuous-time Markov process with exponential waiting times. The gain and loss rates are sampled from two independent gamma distributions (Cohen and Pupko, 2010).

GLOOME needs a binary phylogenetic tree with branch lengths as input. The taxonomy of the National Center for Biointechnology Information (NCBI) and other taxonomies are not suitable, because they provide no branch length information. We used the sequenced tree of life (sTOL) (Fang et al., 2013), which is bifurcating and was inferred with a maximum likelihood approach based on unbiased sampling of structural protein domains from whole genomes of all sequenced organisms (Gough et al., 2001). We employed GLOOME with standard settings to infer posterior probabilities for the phenotype and Pfam family gains and losses from the Pfam phyletic patterns of all NCBI bacteria represented in the sTOL and the GIDEON phenotypes. Each GIDEON phenotype p is available for a varying number of bacteria. Therefore, for each phenotype, we pruned the sTOL to those bacteria that

were both present in the NCBI database and had a label for the respective phenotype in GIDEON. The posterior probabilities of ancestral Pfam gains and losses were then mapped onto this GIDEON phenotype-specific tree (Gps-sTOL, Supplementary Figure 2).

Let *B* be the set of all branches in the sTOL and *P* be the set of all Pfam families. We then denote the posterior probability g_{ij} of an event a for a Pfam family pf to be a gain event on branch b in the sTOL computed with GLOOME as:

$$g_{ij} = P(a = gain | i = b, j = pf) \; \forall \; i \; \in B, \forall \; j \; \in P,$$

and the posterior probability of a to be a loss event for a Pfam family p on branch b as:

$$l_{ij} = P(a = loss | i = b, j = pf) \; \forall \; i \; \in B, \forall \; j \; \in \; P.$$

We established a mapping $f: B' \rightarrow B$ between the branches of the sTOL B and the set of branches B' of the Gps-sTOL (Supplementary Figure 2). This was achieved by traversing the tree from the leaves to the root.

There are two different scenarios for a branch b' in B' to map to the branches in B:

a) Branch b' in the Gps-sTOL derives from a single branch b in the sTOL: $f(b') = \{b\}$. The posterior probability of a Pfam gain inferred in the Gps-sTOL on branch b' consequently is the same as that on branch b in the sTOL $g_{b'j} = g_{bj} \forall \; j \epsilon P$.

b) Branch b' in the Gps-sTOL derives from *m* branches $b_1, \ldots, b_m$ in the sTOL:

$f(b') = \{b_1, \ldots, b_m\}$ (Supplementary figure 2). In this case, we iteratively calculated the posterior probabilities for at least one Pfam gain g' on branch b' from the posterior probabilities for a gain ${g'}_{b_1 j}$ from the posterior probabilities $g_1, \ldots, g_m$of a gain on branches $b_1, \ldots, b_m$ with the help of h:

$$\begin{aligned} h_1 &= g_{b_1 j} \\ h_{n+1} &= (1 - h_n) \cdot g_{b_{n+1} j} \\ {g'}_{b_1 j} &= h_m \; \forall j \in P. \end{aligned}$$

Inferring the Gps-sTOL Pfam posterior loss probabilities ${l'}_{ij}$ from the sTOL posterior Pfam loss probabilities is analogous to deriving the gain probabilities. The posterior probability for a phenotype p to be gained ${g_{ip}}'$ or lost ${l_{ip}}'$ can be directly defined for the Gps-sTOL in the same way as for the Pfam probabilities.

For classification, we did not distinguish between phenotype or Pfam gains or losses, assuming that the same set of protein families gained with a phenotype will also be lost with the phenotype. This assumption simplified the classification problem. Specifically, we proceeded in the following way:

1. We computed the joint probability x_{ij} of a Pfam family gain or loss on branch b' and the joint probability y_j of a phenotype gain or loss on branch b':

$$\begin{aligned} x_{ij} &= {g'}_{ij} {l'}_{ij} + \left(1 - {g'}_{ij}\right) \cdot {l'}_{ij} + \left(1 - {l'}_{ij}\right) \cdot {g'}_{ij} \; \forall i \in B', \forall j \in P \\ &= {g'}_{ij} + \left(1 - {g'}_{ij}\right) \cdot {l'}_{ij} \end{aligned}$$

$$y_i = g'_{ip} + \left(1 - {g_{ip}}'\right) \cdot \; l'_{ip} \quad \forall i \in B'.$$

2. Let $\boldsymbol{x_i}$ be a vector representing the probabilities x_{ij} for all Pfam families $j \in P$ on

branch b_i. We discarded any samples $(\boldsymbol{x}_i, y_{\mathrm{i}})$ that had a probability for a phenotype gain or loss y_{i} above the reporting threshold of GLOOME but below a threshold t. We set the threshold t to 0.5.

This defines the matrix X and the vector **y** as:

$$(X, \boldsymbol{y}) = \{(\boldsymbol{x}_{\mathbf{i}}, y_{\mathrm{i}}) \mid y_i = 0 \vee y_i \geq t, i \in B'\} ,$$

By this means, we avoided presenting the classifier with samples corresponding to uncertain phenotype gain or loss events and used only confident labels in the subsequent classifier training instead.

3. We inferred discrete phenotype labels $\boldsymbol{y}'$ by applying this threshold t to the joint probability y_i for a phenotype gain or loss to set up a well-defined classification problem with a binary target variable. Whenever the probability for a phenotype to be gained or lost on a specific branch was larger than t, the event was considered to have happened:

$$\boldsymbol{y}' = \begin{cases} 1, \text{if } y_i \geq t \\ 0, \quad \text{otherwise} \end{cases} \forall i \in B'.$$

4. Finally, we formulated a joint binary classification problem for each target phenotype $\boldsymbol{yp}$ and the corresponding gain and loss events $\boldsymbol{y}'$, the phyletic patterns XP, and the Pfam gain and loss events X, which we solved again with a linear L1-regularized L2-loss SVM. We applied this procedure for all GIDEON phenotypes under investigation.

Software Requirements

Traitar can be run on a standard laptop with Linux/Unix. The runtime (wallclock time) for annotating and phenotyping a typical microbial genome with 3 Mbp is 9 minutes (3

min/Mbp) on an Intel(R) Core(TM) i5-2410M dual core processor with 2.30 GHz, requiring only a few megabytes of memory.

Cross-validation

We employed cross-validation to assess the performance of the classifiers individually for each phenotype. For a given phenotype, we divided the bacterial samples that were annotated with that phenotype into ten folds. Each fold was selected once for testing the model, which was trained on the remaining folds. The optimal regularization parameter *C* needed to be determined independently in each step of the cross-validation; therefore, we employed a further inner cross-validation using the following range of values for the parameter *C*: 10^{-3}, $10^{-2} \cdot 0.7$, $10^{-2} \cdot 0.5$, $10^{-2} \cdot 0.2$, $10^{-2} \cdot 0.1$, …, 1. In other words, for each fold kept out for testing in the outer cross-validation, we determined the value of the parameter *C* that gave the best accuracy in an additional tenfold cross-validation on the remaining folds. This value was then used to train the SVM model in the current outer cross-validation step. Whenever we proceeded to a new cross-validation fold, we re-computed the ancestral character state reconstruction of the phenotype with only the training samples included (Ancestral protein family and phenotype gains and losses). This procedure is known as nested cross-validation (Ruschhaupt et al., 2004).

The bacterial samples in the training folds imply a Gps-sTOL in each step of the inner and outer cross-validation without the samples in the test fold. We used the same procedure as before to map the Pfam gains and losses inferred previously on the Gps-sTOL onto the tree defined by the current cross-validation training folds. Importantly, the test error is only estimated on the observed phenotype labels rather than on the inferred phenotype gains and losses.

Evaluation metrics

We used evaluation metrics from multi-label classification theory for performance evaluation (Manning et al., 2008). We determined the performance for the individual phenotype-positive and the phenotype-negative classes based on the confusion matrix of true positive (*TP*), true negative (*TN*), false negative (*FN*) and false positive (*FP*) samples from their binary classification equivalents by averaging over all n phenotypes. We utilized two different accuracy measures for assessing multi-class classification performance (i.e. the accuracy pooled over all classification decisions and the macro-accuracy). Macro-accuracy represents an average over the accuracy of the individual binary classification problems and we computed this from the macro-recall of the phenotype-positive and the phenotype-negative classes as follows:

$$Macro\text{-}recall_{Pos} = \left(\sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i}\right)/n$$

$$Macro\text{-}recall_{Neg} = \left(\sum_{i=1}^{n} \frac{TN_i}{FP_i + TN_i}\right)/n$$

$$Macro\text{-}accuracy = \left(Macro\text{-}recall_{Pos} + Macro\text{-}recall_{Neg}\right)/2.$$

However, if there are only few available labels for some phenotypes, the variance of the macro-accuracy will be high and this measure cannot be reliably computed anymore; it cannot be computed at all if no labels are available. The accuracy only assesses the overall classification performance without consideration of the information about specific phenotypes. Large classes dominate small classes (Manning et al., 2008).

$$Recall_{Pos} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$$

$$Recall_{Neg} = \frac{\sum_{i=1}^{n} TN_i}{\sum_{i=1}^{n} TN_i + \sum_{i=1}^{n} FP_i}$$

$$Accuracy = \left(Recall_{Pos} + Recall_{Neg}\right)/2$$

Majority feature selection

The weights in linear SVMs can directly be linked to features that are relevant for the classification. We identified the most important protein families used as features from the voting committee of SVMs consisting of the five most accurate models, which were also used for classifying new samples. If the majority, which is at least three predictors, included a positive value for a given protein family, we added this feature to the list of important features. We further ranked these protein families features by their correlation with the phenotype using Pearson's correlation coefficient.

Acknowledgements

We thank Andreas Klötgen, David Lähnemann, Susanne Reimering and Alexander Sczyrba for providing helpful comments on the manuscript; Johannes Dröge and Jens Loers for reviewing the Traitar software and Gary Robertson for helping to set up the Traitar web service. JAF and PBP are supported by a grant from the European Research Council (336355-MicroDE).

References

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nat Methods doi:10.1038/nmeth.3103.

Bai, Y., Muller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Munch, P.C., Spaepen, S., Remus-Emsermann, M., *et al*. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. Nature doi:10.1038/nature16192.

Barker, D., and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS computational biology doi:10.1371/journal.pcbi.0010003.

Berger, S.A. (2005). GIDEON: a comprehensive Web-based resource for geographic medicine. International journal of health geographics doi:10.1186/1476-072X-4-10.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. Paper presented at: Proceedings of the fifth annual workshop on computational learning theory (Association for Computing Machinery).

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature doi:10.1038/nature14486.

Cleary, B., Brito, I.L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E.J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol doi:10.1038/nbt.3329.

Cohen, O., and Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. Molecular biology and evolution doi:10.1093/molbev/msp240.

Cohen, O., and Pupko, T. (2011). Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony--a simulation study. Genome biology and evolution doi:10.1093/gbe/evr101.

Draper, J., Karplus, K., and Ottemann, K.M. (2011). Identification of a chemoreceptor zinc-binding domain common to cytoplasmic bacterial chemoreceptors. Journal of bacteriology doi:10.1128/JB.05140-11.

Dutta, R., Qin, L., and Inouye, M. (1999). Histidine kinases: diversity of domain organization. Molecular microbiology *34,* 633-640

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. J Mach Learn Res *9,* 1871-1874

Fang, H., Oates, M.E., Pethica, R.B., Greenwood, J.M., Sardar, A.J., Rackham, O.J., Donoghue, P.C., Stamatakis, A., de Lima Morais, D.A., and Gough, J. (2013). A daily-updated tree of (sequenced) life as a reference for genome research. Scientific reports doi:10.1038/srep02015.

Feldbauer, R., Schulz, F., Horn, M., and Rattei, T. (2015). Prediction of microbial phenotypes based on comparative genomics. BMC bioinformatics doi:10.1186/1471-2105-16-S14-S1.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al*. (2014). Pfam: the protein families database. Nucleic acids research doi:10.1093/nar/gkt1223.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research doi:10.1093/nar/gkr367.

Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G.H., McHardy, A.C., Nederbragt, A.J., and Pope, P.B. (2015). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. bioRxiv doi:10.1101/026922.

Gilead, S., and Shoham, Y. (1995). Purification and characterization of alpha-L-arabinofuranosidase from Bacillus stearothermophilus T-6. Applied and environmental microbiology *61*, 170-174

Goodfellow, M., Kämpfer, P., Busse, H.-J., Trujillo, M.E., Suzuki, K.-i., Ludwig, W., and Whitman, W.B. (2012). Bergey's manual of systematic bacteriology (Springer New York).

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology doi:10.1006/jmbi.2001.5080.

Gregor, I., Droge, J., Schirmer, M., Quince, C., and McHardy, A.C. (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ doi:10.7717/peerj.1603.

Harvey, P.H., and Pagel, M.D. (1991). The comparative method in evolutionary biology, Vol 239 (Oxford University Press Oxford).

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., *et al*. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science doi:10.1126/science.1200387.

Hosking, E.R., Vogt, C., Bakker, E.P., and Manson, M.D. (2006). The Escherichia coli MotAB proton channel unplugged. Journal of molecular biology doi:10.1016/j.jmb.2006.09.035.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics doi:10.1186/1471-2105-11-119.

Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., and Tyson, G.W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ doi:10.7717/peerj.603.

Josenhans, C., and Suerbaum, S. (2002). The role of motility as a virulence factor in bacteria. International journal of medical microbiology : IJMM doi:10.1078/1438-4221-00173.

Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ doi:10.7717/peerj.1165.

Kastenmuller, G., Schenk, M.E., Gasteiger, J., and Mewes, H.W. (2009). Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. Genome biology doi:10.1186/gb-2009-10-3-r28.

Konietzny, S.G., Pope, P.B., Weimann, A., and McHardy, A.C. (2014). Inference of phenotype-defining functional modules of protein families for microbial plant biomass degraders. Biotechnology for biofuels doi:10.1186/s13068-014-0124-8.

Lam, W.W., Woo, E.J., Kotaka, M., Tam, W.K., Leung, Y.C., Ling, T.K., and Au, S.W. (2010). Molecular interaction of flagellar export chaperone FliS and cochaperone HP1076 in Helicobacter pylori. FASEB journal : official publication of the Federation of American Societies for Experimental Biology doi:10.1096/fj.10-155242.

Lasken, R.S., and McLean, J.S. (2014). Recent advances in genomic DNA sequencing of microbial species from single cells. Nat Rev Genet doi:10.1038/nrg3785.

Limam, R.D., Chouari, R., Mazeas, L., Wu, T.D., Li, T., Grossin-Debattista, J., Guerquin-Kern, J.L., Saidi, M., Landoulsi, A., Sghir, A., *et al*. (2014). Members of the uncultured bacterial candidate division WWE1 are implicated in anaerobic digestion of cellulose. MicrobiologyOpen doi:10.1002/mbo3.144.

Lingner, T., Muhlhausen, S., Gabaldon, T., Notredame, C., and Meinicke, P. (2010). Predicting phenotypic traits of prokaryotes from protein domain frequencies. BMC bioinformatics doi:10.1186/1471-2105-11-481.

Liu, R., and Ochman, H. (2007). Stepwise formation of the bacterial flagellar system. Proceedings of the National Academy of Sciences of the United States of America doi:10.1073/pnas.0700266104.

Liu, Y., Li, J., Sam, L., Goh, C.S., Gerstein, M., and Lussier, Y.A. (2006). An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. PLoS computational biology doi:10.1371/journal.pcbi.0020159.

MacDonald, N.J., and Beiko, R.G. (2010). Efficient learning of microbial genotype-phenotype association rules. Bioinformatics doi:10.1093/bioinformatics/btq305.

Macnab, R.M. (2003). How bacteria assemble flagella. Annual review of microbiology doi:10.1146/annurev.micro.57.030502.090832.

Manning, C.D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval, Vol 1 (Cambridge university press Cambridge, UK).

Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D., *et al*. (2008). Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). Nat Biotechnol doi:10.1038/nbt1403.

Martinez, J.L. (2008). Antibiotics and antibiotic resistance genes in natural environments. Science doi:10.1126/science.1159483.

Martiny, J.B., Jones, S.E., Lennon, J.T., and Martiny, A.C. (2015). Microbiomes in light of traits: A phylogenetic perspective. Science doi:10.1126/science.aac9323.

Moreno-Vivian, C., Cabello, P., Martinez-Luque, M., Blasco, R., and Castillo, F. (1999). Prokaryotic nitrate reduction: molecular properties and functional distinction among bacterial nitrate reductases. Journal of bacteriology *181*, 6573-6584

Narihiro, T., and Sekiguchi, Y. (2007). Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update. Current opinion in biotechnology doi:10.1016/j.copbio.2007.04.003.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D.R., Gautier, L., Pedersen, A.G., Le Chatelier, E., *et al*. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol doi:10.1038/nbt.2939.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature doi:10.1038/35012500.

Olapade, O.A., and Ronk, A.J. (2015). Isolation, characterization and community diversity of indigenous putative toluene-degrading bacterial populations with catechol-2,3-dioxygenase genes in contaminated soils. Microbial ecology doi:10.1007/s00248-014-0466-6.

Pal, C., Papp, B., and Lercher, M.J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature genetics doi:10.1038/ng1686.

Pantel, I., Lindgren, P.E., Neubauer, H., and Gotz, F. (1998). Identification and characterization of the Staphylococcus carnosus nitrate reductase operon. Molecular & general genetics : MGG *259*, 105-114

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research doi:10.1101/gr.186072.114.

Paterson, G.K., and Mitchell, T.J. (2004). The biology of Gram-positive sortase enzymes. Trends in microbiology doi:10.1016/j.tim.2003.12.007.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. J Mach Learn Res *12*, 2825-2830

Pelletier, E., Kreimeyer, A., Bocs, S., Rouy, Z., Gyapay, G., Chouari, R., Riviere, D., Ganesan, A., Daegelen, P., Sghir, A., *et al*. (2008). "Candidatus Cloacamonas acidaminovorans": genome sequence reconstruction provides a first glimpse of a new bacterial division. Journal of bacteriology doi:10.1128/JB.01248-07.

Pope, P.B., Smith, W., Denman, S.E., Tringe, S.G., Barry, K., Hugenholtz, P., McSweeney, C.S., McHardy, A.C., and Morrison, M. (2011). Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. Science doi:10.1126/science.1205760.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al*. (2012). The Pfam protein families database. Nucleic acids research doi:10.1093/nar/gkr1065.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., *et al*. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature doi:10.1038/nature12352.

Ruschhaupt, M., Huber, W., Poustka, A., and Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. Statistical applications in genetics and molecular biology doi:10.2202/1544-6115.1078.

Sa-Nogueira, I., Nogueira, T.V., Soares, S., and de Lencastre, H. (1997). The Bacillus subtilis L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression. Microbiology *143 (Pt 3)*, 957-969

Tange, O. (2011). GNU parallel-the command-line power tool. USENIX *36*, 42-47

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic acids research doi:10.1093/nar/29.1.22.

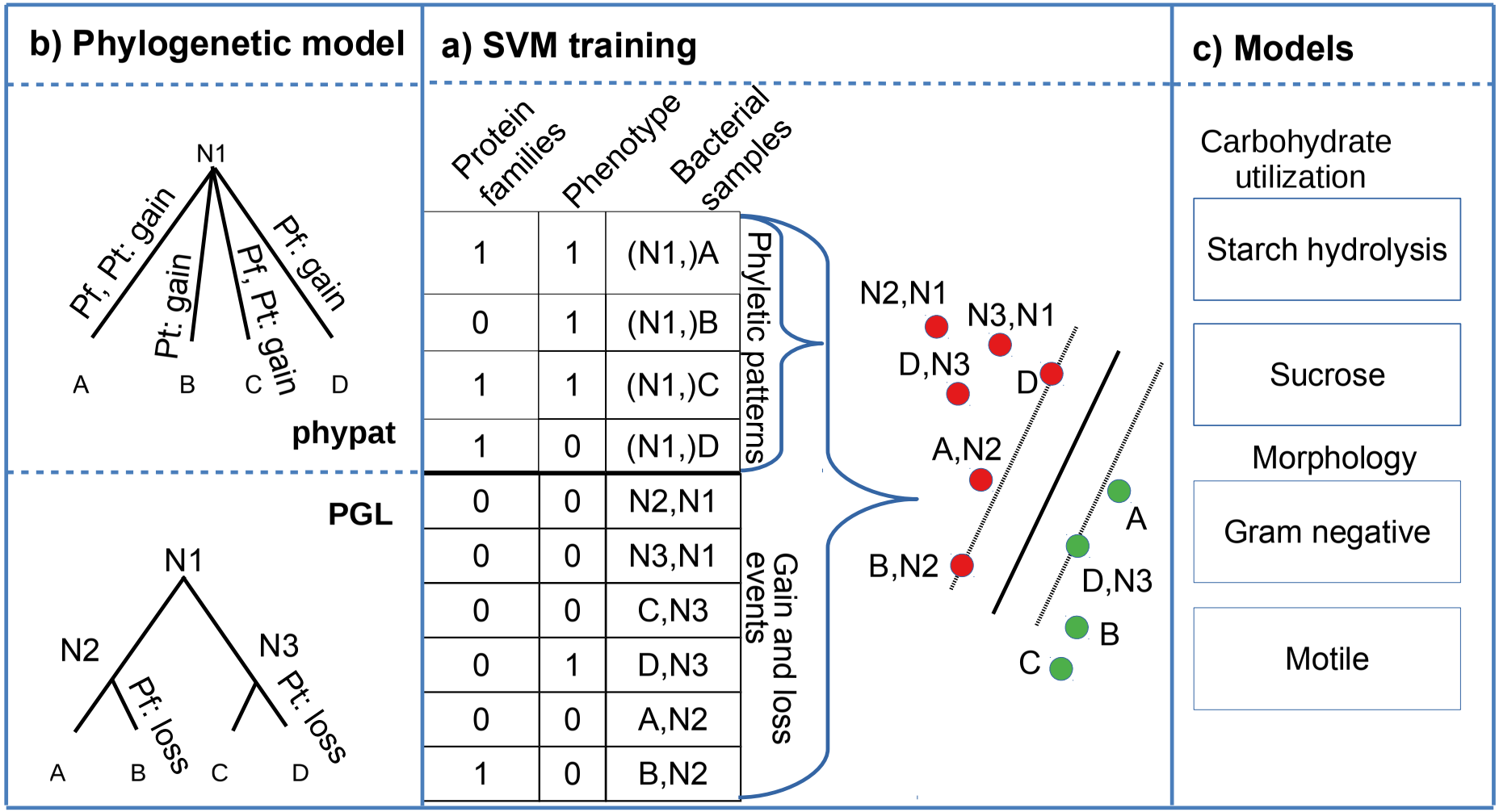
Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., *et al*. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids research doi:10.1093/nar/gkt1099.

Weimann, A., Trukhina, Y., Pope, P.B., Konietzny, S.G., and McHardy, A.C. (2013). De novo prediction of the genomic components and capabilities for microbial plant biomass degradation from (meta-)genomes. Biotechnology for biofuels doi:10.1186/1754-6834-6-24.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J R Stat Soc A *67*, 301-320

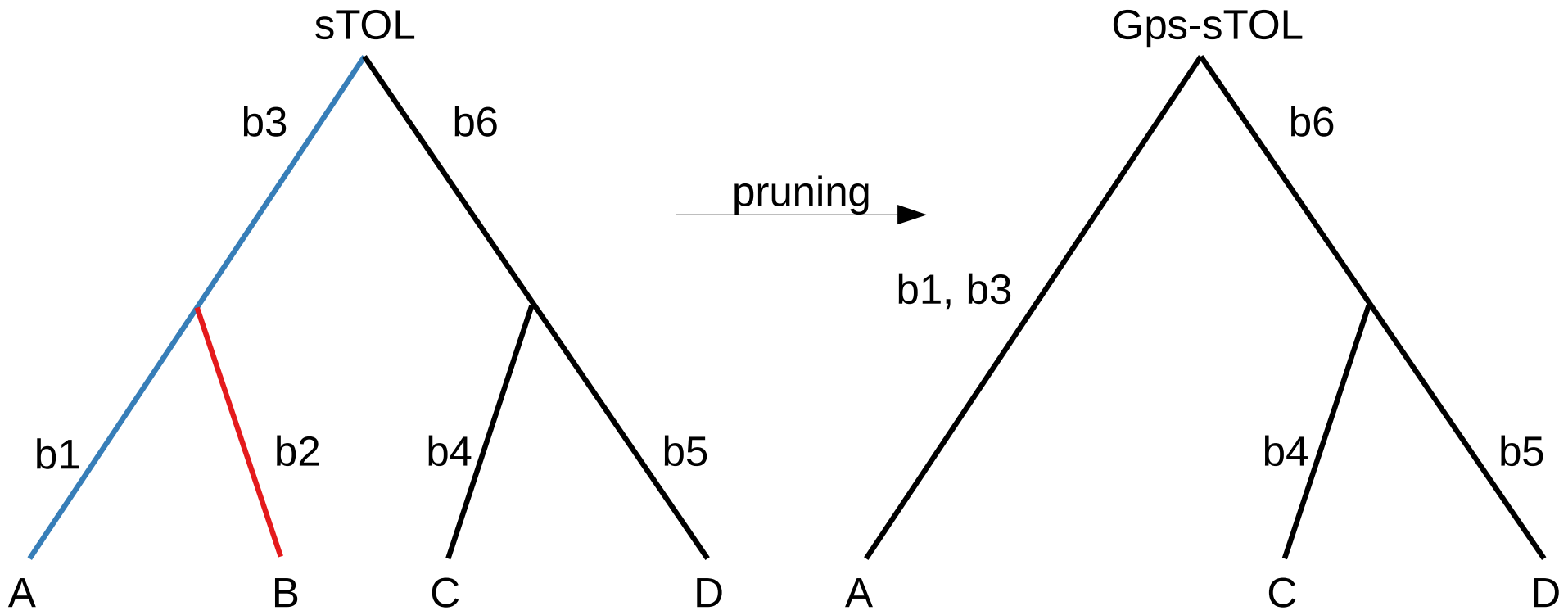
Supplementary Figure 1

Schematic overview of the Traitar phenotype model training. (a) The phenotype and Pfam protein family phyletic patterns correspond to gain events on a star-shaped phylogenetic tree. Alternatively, we reconstructed the ancestral Pfam family and phenotype gain and loss events on the sequenced Tree of Life. (b) We trained a support vector machine classifier either on the phyletic patterns and on the ancestral gain and loss events, or solely on the phyletic patterns. (c) In this way, we inferred classification models for all available phenotypes.



Supplementary Figure 2

Sequenced Tree of Life (sTOL) and GIDEON phenotype-specific Tree of Life (Gps-sTOL) correspondence. The phenotype label for Sample B is not available. Consequently, only branches b_4, b_5 and b_6 are also found in the Gps-sTOL. The posterior probabilities for a Pfam gain or loss are the same for b_4, b_5 and b_6 in both trees. Branches b_1 and b_3 (blue) are collapsed into a single branch. The posterior probability for a gain on branch $b_{1,3}$, $g_{b1,3}$ is computed from the posterior probability for a Pfam gain for b_1 and b_3 as follows: $g_{b1,3} = g_{b1} + (1 - g_{b1}) \cdot g_{b3}$ Branch b_2 (red) in the sTOL does not have an analog in the Gps-sTOL.



Supplementary table S1 Detailed information on the 67 phenotypes used in this study

Phenotype[(a)]	Test type[(b)]	Test description[(c)]	GIDEON I+[(e)]	GIDEON I-[(d)]	GIDEON I total[(f)]	GIDEON II+[(h)]	GIDEON II-[(g)]	GIDEON II total[(i)]	Bergey's+[(k)]	Bergey's-[(l)]	Bergey's total[(m)]
Acetate utilization	General test	A variety of commercial kits are satisfactory. Includes late reactions for gram-positive and non-fermentative gram-negative rods	27	19	46	5	2	7	7	10	17
Aerobe	Basic test	Organisms which grow only in the presence of air.	64	167	231	7	35	42	25	0	25
Alkaline phosphatase	General test	Most kits utilize p-nitrophenyl phosphate as substrate. Insure thorough washing if phosphate buffers are employed	30	15	45	7	3	10	12	21	33
Anaerobe	Basic test	Organisms which do not grow in the presence of oxygen	46	169	215	14	27	41	16	11	27
Arginine dihydrolase	General test	For most organisms, Moeller medium interpreted with control after 18 hours (or longer for gram-positive and non-fermentative gram-negative rods)	46	88	134	8	15	23	35	55	90
Bacillus or coccobacillus	Basic test	Bacilli predominate	160	48	208	28	12	40	0	0	0
Beta hemolysis	General test	Sheep blood	18	134	152	4	17	21	4	5	9

Bile-susceptible	General test	Media and bile concentrations vary according to the species tested; Presumpto Plates useful for anaerobes	16	35	51	2	7	9	0	0	0
Capnophilic	General test	Exogenous carbon dioxide (5 to 7%) must be present for growth; Gas Pak or cylinder gas are preferred	22	175	197	2	33	35	0	0	0
Casein hydrolysis	General test	Standard skim-milk/nutrient agar halo test; Presumpto Plates useful for anaerobes	13	47	60	2	7	9	19	19	38
Catalase	Basic test	Perform on young colonies (up to 24 hours) using 3% hydrogen peroxide (alternative technique for Mycobacteria); Presumpto Plates useful for anaerobes	123	91	214	18	20	38	34	32	66
Cellobiose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	33	63	96	4	7	11	42	50	92

Citrate	General test	Simmons citrate medium using a light inoculum (avoid stabbing the agar). Includes late reactions for gram-positive and non-fermentative gram-negative rods	35	81	116	11	16	27	16	19	35
Coagulase production	General test	Standard or commercial slide plasma tests acceptable; tube tests may help differentiate Staphylococcus aureus from other taxa	30	45	75	3	7	10	2	3	5
Coccus	Basic test	Cocci predominate.	44	170	214	12	28	40	0	0	0
Coccus - clusters or groups predominate	Basic test	The predominant forms are cocci, in clusters or irregular groups.	12	204	216	2	39	41	0	0	0
Coccus - pairs or chains predominate	Basic test	The predominant forms are cocci in chains or pairs.	28	185	213	10	30	40	0	0	0
Colistin-Polymyxin susceptible	General test	Standard disk diffusion technique; recommended media and disk potency may vary for specific taxa	38	55	93	3	11	14	0	0	0

D-Mannitol	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	53	128	181	12	19	31	49	58	107
D-Mannose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	85	60	145	14	12	26	70	20	90
DNase	General test	Standard commercial agar tests; Presumpto Plates useful for anaerobes	15	71	86	2	9	11	2	0	2

D-Sorbitol	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	18	127	145	5	24	29	18	60	78
D-Xylose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	44	118	162	10	22	32	32	77	109
Esculin hydrolysis	General test	For most organisms, Moeller medium interpreted with control after 18 hours (or longer for gram-positive and non-fermentative gram-negative rods)	52	93	145	17	12	29	49	31	80
Facultative	Basic test	Organisms which grow both in the presence and absence of air.	97	130	227	19	23	42	0	0	0

Gas from glucose	General test	Gas produced from D-glucose; Durham tube or gas bubbles noted in commercial kits	22	106	128	3	15	18	4	6	10
Gelatin hydrolysis	General test	Commercial or self prepared (or X-ray film) tests interpreted after 24-48 hours against control at lowered temperature; Presumpto Plates useful for anaerobes	30	111	141	3	19	22	30	31	61
Glucose fermenter	Basic test	Commercial phenol red techniques are generally acceptable; Andrades or more sensitive indicators necessary for organisms producing more subtle pH changes; specific acidification techniques applied for Neisseria; Presumpto Plates useful for anaerobes	131	78	209	26	9	35	0	0	0
Glucose oxidizer	Basic test	Hugh and Leifson method; in most cases 'positive' indicates nonfermentative organism which oxidizes glucose	22	192	214	2	38	40	0	0	0

Glycerol	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	24	86	110	8	14	22	28	31	59
Gram negative	Basic test	Gram-negative forms predominate	117	110	227	17	24	41	0	0	0
Gram positive	Basic test	Gram-positive forms predominate	100	127	227	23	18	41	14	8	22
Growth at 42 degrees C	General test	Media vary according to the species tested	43	27	70	10	2	12	0	4	4
Growth in 6.5% NaCl	General test	Media vary according to the species tested	29	67	96	6	9	15	0	0	0
Growth in KCN	General test	Commercial kits based on 1:13,000 KCN are suggested	10	14	24	2	0	2	18	10	28
Growth on MacConkey agar	Basic test	Visible growth within 48 hours; or within 7 days for gram positive and non-fermentative gram negative rods	55	165	220	7	31	38	0	2	2
Growth on ordinary blood agar	Basic test	Visible growth on sheep blood agar within 48 hours	211	18	229	39	3	42	0	0	0

Hydrogen sulfide	General test	TSI for enterobacteriaceae and most other species; Presumpto Plates useful for anaerobes. Includes late appearance of hydrogen sulfide for gram-positive and non-fermentative gram-negative rods	13	116	129	4	16	20	0	2	2
Indole	General test	A spot test is acceptable for most organisms; overnight testing with a paper strip is helpful in confirming negative reactions; Presumpto Plates useful for anaerobes	20	145	165	2	24	26	18	47	65
Lactose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	61	104	165	11	17	28	48	55	103

L-Arabinose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	44	114	158	8	23	31	24	84	108
Lipase	General test	Standard egg yolk agar test; Presumpto Plates useful for anaerobes	20	76	96	2	13	15	5	9	14
L-Rhamnose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	20	109	129	2	24	26	10	43	53
Lysine decarboxylase	General test	For most organisms, Moeller medium interpreted with control after 18 hours (or longer for gram-positive and non-fermentative gram-negative rods)	11	63	74	4	5	9	6	20	26

Malonate	General test	Standard test based on maintenance of alkaline pH (bromthymol blue) in the presence of glucose and malonate; commercial kits are acceptable	11	29	40	1	4	5	0	0	0
Maltose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	105	66	171	18	9	27	86	32	118
Melibiose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	25	88	113	9	18	27	18	45	63
Methyl red	General test	Commercial or self-prepared media are generally acceptable	23	29	52	2	5	7	5	7	12

Motile	General test	Standard hanging drop on fresh broth isolates for most purposes; perform at 22 to 25 degrees C if Listeria suspected	70	144	214	11	26	37	60	32	92
Mucate utilization	General test	A variety of commercial kits are satisfactory. Includes late reactions for gram-positive and non-fermentative gram-negative rods	11	23	34	2	6	8	3	8	11
myo-Inositol	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	10	107	117	5	16	21	3	29	32
Nitrate to nitrite	General test	Commercial and self-prepared media are acceptable; alternative techniques used for mycobacteria	75	92	167	8	19	27	31	43	74
Nitrite to gas	General test	Standard zinc dust test applied to 'nitrate-negative' organisms	14	76	90	1	11	12	0	0	0

ONPG (beta galactosidase)	General test	Commercial kits are generally satisfactory; suggest a heavy inoculum in buffered medium; yellow pigmented organisms may not be suitable for testing	45	70	115	8	9	17	4	11	15
Ornithine decarboxylase	General test	For most organisms, Moeller medium interpreted with control after 18 hours (or longer for gram-positive and non-fermentative gram-negative rods)	17	67	84	2	5	7	13	23	36
Oxidase	Basic test	Paper strip test from appropriate media	56	130	186	5	24	29	19	19	38
Pyrrolidonyl-beta-naphthylamide	General test	L-pyrrolidonyl-beta-naphthylamide - PYR (or pyrrolidonyl arylamidase - PYRA) - reagents commercially available; read color within 10 seconds (2 minutes for Carr-Scarborough reagent)	23	56	79	5	10	15	0	0	0

Raffinose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	30	107	137	10	17	27	23	66	89
Salicin	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	41	93	134	9	12	21	28	24	52
Spore formation	Basic test	Note that spores may only appear in vitro, and may not be seen in clinical material	18	216	234	2	40	42	3	2	5
Starch hydrolysis	General test	Standard starch hydrolysis or Mueller-Hilton agar tests developed with iodine solutions; Presumpto Plates useful for anaerobes	27	65	92	5	15	20	38	44	82

Sucrose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	84	93	177	20	11	31	85	44	129
Tartrate utilization	General test	A variety of commercial kits are satisfactory. Includes late reactions for gram-positive and non-fermentative gram-negative rods	10	18	28	2	3	5	4	3	7
Trehalose	Fermentation or acidification	Commercial phenol red techniques usually acceptable; Andrades or more sensitive indicators necessary for organisms with more subtle pH changes. Includes late reactions for gram positive & non-fermentative gram negative rods	75	63	138	14	15	29	66	42	108

Urea hydrolysis	General test	Christensen agar testing for most taxa; other techniques for mycobacteria, ureaplasma and other organisms as recommended. Includes late reactions for gram-positive and non-fermentative gram-negative rods	28	127	155	7	22	29	15	56	71
Voges Proskauer	General test	Commercial or self-prepared media are generally acceptable; the test is most reliable when performed on cultures no older than three days	28	68	96	8	8	16	10	19	29
Yellow pigment	General test	Yellow pigment noted on sheep blood or other primary isolation agar. Includes late appearance of pigment for gram-positive and non-fermentative gram-negative rods	13	161	174	1	28	29	1	4	5

(a) GIDEON phenotypes
(b) Type of test required for the phenotype determination in the wet lab according to GIDEON
(c) Remarks on wet lab test for determination of the phenotype according to GIDEON
(d) Number of phenotype-positive bacteria in the GIDEON I dataset
(e) Number of phenotype-negative bacteria in the GIDEON I dataset
(f) Total number of bacteria with phenotype labels in the GIDEON I dataset
(g) Number of phenotype-positive bacteria in the GIDEON II dataset
(h) Number of phenotype-negative bacteria in the GIDEON II dataset
(i) Total number of bacteria with phenotype labels in the GIDEON II dataset

(j) Number of phenotype-positive bacteria in the Bergey dataset
(k) Number of phenotype-negative bacteria in the Bergey dataset
(l) Total number of bacteria with phenotype labels in the Bergey dataset

Supplementary Table S2 Macro-accuracy of the phypat and phypat+PGL classifiers obtained in cross-validation experiments for the 67 GIDEON phenotypes

Phenotype[(a)]	phypat+PGL[(b)]	phypat[(c)]
Spore formation	1	0.887
Methyl red	1	0.905
Gram positive	0.996	1
Gram negative	0.991	0.987
Growth on MacConkey agar	0.988	0.97
Anaerobe	0.986	0.959
Catalase	0.984	0.985
Aerobe	0.972	0.98
Coccus - pairs or chains predominate	0.903	0.974
Coagulase production	0.972	0.944
Glucose fermenter	0.966	0.897
Glucose oxidizer	0.962	0.949
Oxidase	0.962	0.937
Motile	0.958	0.912
Nitrate to nitrite	0.954	0.926
Cellobiose	0.946	0.877
Growth in 6.5% NaCl	0.943	0.862
Bacillus or coccobacillus	0.939	0.811
Coccus	0.879	0.931
L-Arabinose	0.904	0.923
Sucrose	0.922	0.825
Urea hydrolysis	0.889	0.921
Esculin hydrolysis	0.915	0.846
D-Xylose	0.905	0.857
Citrate	0.904	0.765
Gelatin hydrolysis	0.902	0.817
Glycerol	0.864	0.769
Trehalose	0.899	0.825
D-Mannitol	0.828	0.897
Facultative	0.862	0.896

ONPG (beta galactosidase)	0.883	0.857
Nitrite to gas	0.879	0.84
D-Sorbitol	0.842	0.877
Voges Proskauer	0.876	0.866
Acetate utilization	0.75	0.876
Malonate	0.875	0.806
Melibiose	0.875	0.769
Raffinose	0.873	0.749
Coccus - clusters or groups predominate	0.816	0.87
Capnophilic	0.747	0.869
Lipase	0.867	0.824
D-Mannose	0.865	0.866
Salicin	0.861	0.789
Colistin-Polymyxin susceptible	0.861	0.844
Beta hemolysis	0.855	0.765
Lactose	0.854	0.847
Maltose	0.854	0.794
Casein hydrolysis	0.814	0.846
L-Rhamnose	0.781	0.84
Growth on ordinary blood agar	0.833	0.691
Ornithine decarboxylase	0.83	0.755
Pyrrolidonyl-beta-naphthylamide	0.829	0.767
Growth at 42 degrees C	0.613	0.801
Gas from glucose	0.788	0.794
Starch hydrolysis	0.793	0.793
Growth in KCN	0.793	0.679
Indole	0.728	0.792
Lysine decarboxylase	0.6	0.737
Mucate utilization	0.775	0.662
Arginine dihydrolase	0.743	0.766
Bile-susceptible	0.761	0.681
Alkaline phosphatase	0.633	0.75
Hydrogen sulfide	0.734	0.559

Tartrate utilization	0.694	0.328
Yellow pigment	0.688	0.613
myo-Inositol	0.68	0.639
DNase	0.658	0.677

(a) Phenotypes sorted by the maximal macro-accuracy determined from a 10-fold nested cross-validation from 234 bacteria described in the Global Infectious Disease and Epidemiology Online Network

(b) Macro accuracy for the phypat+PGL classifier

(c) Macro accuracy for the phypat classifier

Supplementary Table S3 Mapping of bacterial strains to 234 species described in the Global Infectious Disease and Epidemiology Online Network with links to the National Center for Biotechnology Information (NCBI) databases

Strain[(a)]	Species[(b)]	Bioproject id[(c)]	NCBI taxonomy id[(d)]
Acholeplasma laidlawii PG-8A	Acholeplasma laidlawii	19259	441768
Achromobacter xylosoxidans A8	Achromobacter xylosoxidans	762376	59899
Acidaminococcus fermentans DSM 20731	Acidaminococcus fermentans	591001	43471
Acidaminococcus intestini RyC-MR95	Acidaminococcus intestini	568816	74445
Acidovorax avenae subsp. avenae ATCC 19860	Acidovorax avenae	643561	42497
Acinetobacter baumannii 1656-2	Acinetobacter baumannii	400667	58731
Acinetobacter baumannii AB0057	Acinetobacter baumannii	405416	58765
Acinetobacter baumannii AB307-0294	Acinetobacter baumannii	480119	59083
Acinetobacter baumannii ACICU	Acinetobacter baumannii	497978	158685
Acinetobacter baumannii ATCC 17978	Acinetobacter baumannii	509170	61601
Acinetobacter baumannii AYE	Acinetobacter baumannii	509173	61637
Acinetobacter baumannii MDR-TJ	Acinetobacter baumannii	557600	59271
Acinetobacter baumannii MDR-ZJ06	Acinetobacter baumannii	696749	158677
Acinetobacter baumannii SDF	Acinetobacter baumannii	889738	162739
Acinetobacter baumannii TCDC-AB0715	Acinetobacter baumannii	980514	158679
Acinetobacter calcoaceticus PHEA-2	Acinetobacter calcoaceticus	871585	83123
Aerococcus urinae ACS-120-V-Col10a	Aerococcus urinae	866775	64757
Aeromonas hydrophila subsp. hydrophila ATCC 7966	Aeromonas hydrophila	380703	58617
Aeromonas salmonicida subsp. salmonicida A449	Aeromonas salmonicida	382245	58631
Aggregatibacter actinomycetemcomitans ANH9381	Aggregatibacter actinomycetemcomitans	694569	46989
Aggregatibacter actinomycetemcomitans D7S-1	Aggregatibacter actinomycetemcomitans	754507	80743
Aggregatibacter aphrophilus NJ8700	Aggregatibacter aphrophilus	634176	59407
Anaerococcus prevotii DSM 20548	Anaerococcus prevotii	525919	59219
Arcanobacterium haemolyticum DSM 20595	Arcanobacterium haemolyticum	644284	49489
Arcobacter butzleri ED-1	Arcobacter butzleri	367737	58557
Arcobacter butzleri RM4018	Arcobacter butzleri	944546	158699
Arthrobacter aurescens TC1	Arthrobacter aurescens	290340	58109
Atopobium parvulum DSM 20469	Atopobium parvulum	521095	59195
Bacillus anthracis str. A0248	Bacillus anthracis	198094	57909
Bacillus anthracis str. Ames	Bacillus anthracis	260799	58091
Bacillus anthracis str. 'Ames Ancestor'	Bacillus anthracis	261594	58083
Bacillus anthracis str. CDC 684	Bacillus anthracis	568206	59303

Bacillus anthracis str. H9401	Bacillus anthracis	592021	59385
Bacillus anthracis str. Sterne	Bacillus anthracis	768494	162021
Bacillus cereus 03BB102	Bacillus cereus	222523	57673
Bacillus cereus AH187	Bacillus cereus	226900	57975
Bacillus cereus AH820	Bacillus cereus	288681	58103
Bacillus cereus ATCC 10987	Bacillus cereus	334406	82815
Bacillus cereus ATCC 14579	Bacillus cereus	347495	83611
Bacillus cereus B4264	Bacillus cereus	361100	58529
Bacillus cereus biovar anthracis str. CI	Bacillus cereus	405531	58759
Bacillus cereus E33L	Bacillus cereus	405532	58757
Bacillus cereus F837/76	Bacillus cereus	405534	58753
Bacillus cereus G9842	Bacillus cereus	405535	58751
Bacillus cereus NC7401	Bacillus cereus	572264	59299
Bacillus cereus Q1	Bacillus cereus	637380	50615
Bacillus coagulans 2-6	Bacillus coagulans	345219	54335
Bacillus coagulans 36D1	Bacillus coagulans	941639	68053
Bacillus licheniformis DSM 13 = ATCC 14580	Bacillus licheniformis	279010	58097
Bacillus megaterium DSM 319	Bacillus megaterium	1006007	159841
Bacillus megaterium WSH-002	Bacillus megaterium	592022	48371
Bacillus pumilus SAFR-032	Bacillus pumilus	315750	59017
Bacillus subtilis BSn5	Bacillus subtilis	1052585	73967
Bacillus subtilis subsp. spizizenii str. W23	Bacillus subtilis	1052588	158879
Bacillus subtilis subsp. spizizenii TU-B-10	Bacillus subtilis	224308	57675
Bacillus subtilis subsp. subtilis str. 168	Bacillus subtilis	655816	51879
Bacillus subtilis subsp. subtilis str. RO-NN-1	Bacillus subtilis	936156	62463
Bacillus thuringiensis BMB171	Bacillus thuringiensis	281309	58089
Bacillus thuringiensis serovar chinensis CT-43	Bacillus thuringiensis	412694	58795
Bacillus thuringiensis serovar finitimus YBT-020	Bacillus thuringiensis	541229	158151
Bacillus thuringiensis serovar konkukian str. 97-27	Bacillus thuringiensis	714359	49135
Bacillus thuringiensis str. Al Hakam	Bacillus thuringiensis	930170	158875
Bacteroides fragilis 638R	Bacteroides fragilis	272559	57639
Bacteroides fragilis NCTC 9343	Bacteroides fragilis	295405	58195
Bacteroides fragilis YCH46	Bacteroides fragilis	862962	84217
Bacteroides thetaiotaomicron VPI-5482	Bacteroides thetaiotaomicron	226186	62913
Bacteroides vulgatus ATCC 8482	Bacteroides vulgatus	435590	58253
Bartonella bacilliformis KC583	Bartonella bacilliformis	360095	58533
Bartonella clarridgeiae 73	Bartonella clarridgeiae	696125	62131

Bartonella grahamii as4aup	Bartonella grahamii	634504	59405
Bartonella henselae str. Houston-1	Bartonella henselae	283166	57745
Bartonella quintana str. Toulouse	Bartonella quintana	283165	57635
Bifidobacterium adolescentis ATCC 15703	Bifidobacterium adolescentis	367928	58559
Bifidobacterium bifidum BGN4	Bifidobacterium bifidum	484020	167988
Bifidobacterium bifidum PRL2010	Bifidobacterium bifidum	702459	59883
Bifidobacterium bifidum S17	Bifidobacterium bifidum	883062	59545
Bifidobacterium breve ACS-071-V-Sch8b	Bifidobacterium breve	866777	158863
Bifidobacterium dentium Bd1	Bifidobacterium dentium	401473	43091
Bifidobacterium longum DJO10A	Bifidobacterium longum	1035817	158861
Bifidobacterium longum NCC2705	Bifidobacterium longum	205913	58833
Bifidobacterium longum subsp. infantis 157F	Bifidobacterium longum	206672	57939
Bifidobacterium longum subsp. infantis ATCC 15697 = JCM 1222	Bifidobacterium longum	391904	159865
Bifidobacterium longum subsp. longum BBMN68	Bifidobacterium longum	565040	62693
Bifidobacterium longum subsp. longum JCM 1217	Bifidobacterium longum	565042	62695
Bifidobacterium longum subsp. longum JDM301	Bifidobacterium longum	759350	49131
Bifidobacterium longum subsp. longum KACC 91563	Bifidobacterium longum	890402	60163
Bordetella avium 197N	Bordetella avium	360910	61563
Bordetella bronchiseptica RB50	Bordetella bronchiseptica	257310	57613
Bordetella parapertussis 12822	Bordetella parapertussis	257311	57615
Bordetella pertussis CS	Bordetella pertussis	1017264	158859
Bordetella pertussis Tohama I	Bordetella pertussis	257313	57617
Bordetella petrii DSM 12804	Bordetella petrii	340100	61631
Brachyspira pilosicoli 95/1000	Brachyspira pilosicoli	759914	50609
Brevibacillus brevis NBRC 100599	Brevibacillus brevis	358681	59175
Brucella abortus A13334	Brucella abortus	1104320	83615
Brucella abortus bv. 1 str. 9-941	Brucella abortus	262698	58019
Brucella abortus S19	Brucella abortus	430066	58873
Brucella canis ATCC 23365	Brucella canis	1104321	83613
Brucella canis HSK A52141	Brucella canis	483179	59009
Brucella melitensis ATCC 23457	Brucella melitensis	1029825	158853
Brucella melitensis biovar Abortus 2308	Brucella melitensis	224914	57735
Brucella melitensis bv. 1 str. 16M	Brucella melitensis	359391	62937
Brucella melitensis M28	Brucella melitensis	546272	59241
Brucella melitensis M5-90	Brucella melitensis	703352	158855
Brucella melitensis NI	Brucella melitensis	941967	158857
Brucella suis 1330	Brucella suis	1112912	83617

Brucella suis ATCC 23445	Brucella suis	204722	159871
Brucella suis VBI22	Brucella suis	470137	59015
Burkholderia ambifaria AMMD	Burkholderia ambifaria	339670	58303
Burkholderia ambifaria MC40-6	Burkholderia ambifaria	398577	58701
Burkholderia cenocepacia AU 1054	Burkholderia cenocepacia	216591	57953
Burkholderia cenocepacia HI2424	Burkholderia cenocepacia	331271	58371
Burkholderia cenocepacia J2315	Burkholderia cenocepacia	331272	58369
Burkholderia cenocepacia MC0-3	Burkholderia cenocepacia	406425	58769
Burkholderia gladioli BSR3	Burkholderia gladioli	999541	66301
Burkholderia mallei ATCC 23344	Burkholderia mallei	243160	57725
Burkholderia mallei NCTC 10229	Burkholderia mallei	320388	58387
Burkholderia mallei NCTC 10247	Burkholderia mallei	320389	58385
Burkholderia mallei SAVP1	Burkholderia mallei	412022	58383
Burkholderia multivorans ATCC 17616	Burkholderia multivorans	395019	58697
Burkholderia pseudomallei 1026b	Burkholderia pseudomallei	272560	57733
Burkholderia pseudomallei 1106a	Burkholderia pseudomallei	320372	58391
Burkholderia pseudomallei 1710b	Burkholderia pseudomallei	320373	58389
Burkholderia pseudomallei 668	Burkholderia pseudomallei	357348	58515
Burkholderia pseudomallei K96243	Burkholderia pseudomallei	884204	162511
Burkholderia thailandensis E264	Burkholderia thailandensis	271848	58081
Burkholderia vietnamiensis G4	Burkholderia vietnamiensis	269482	58075
Campylobacter concisus 13826	Campylobacter concisus	360104	58667
Campylobacter curvus 525.92	Campylobacter curvus	360105	58669
Campylobacter hominis ATCC BAA-381	Campylobacter hominis	360107	58981
Campylobacter lari RM2100	Campylobacter lari	306263	58115
Capnocytophaga canimorsus Cc5	Capnocytophaga canimorsus	860228	70727
Capnocytophaga ochracea DSM 7271	Capnocytophaga ochracea	521097	59197
Chromobacterium violaceum ATCC 12472	Chromobacterium violaceum	243365	58001
Citrobacter koseri ATCC BAA-895	Citrobacter koseri	290338	58143
Citrobacter rodentium ICC168	Citrobacter rodentium	637910	43089
Clostridium beijerinckii NCIMB 8052	Clostridium beijerinckii	290402	58137
Clostridium botulinum A2 str. Kyoto	Clostridium botulinum	413999	61579
Clostridium botulinum A3 str. Loch Maree	Clostridium botulinum	441770	58927
Clostridium botulinum A str. ATCC 19397	Clostridium botulinum	441771	58931
Clostridium botulinum A str. ATCC 3502	Clostridium botulinum	441772	58929
Clostridium botulinum A str. Hall	Clostridium botulinum	498213	59147
Clostridium botulinum B1 str. Okra	Clostridium botulinum	498214	59149

Clostridium botulinum Ba4 str. 657	Clostridium botulinum	508765	59159
Clostridium botulinum BKT015925	Clostridium botulinum	508767	59157
Clostridium botulinum B str. Eklund 17B	Clostridium botulinum	515621	59173
Clostridium botulinum E3 str. Alaska E43	Clostridium botulinum	536232	59229
Clostridium botulinum F str. 230613	Clostridium botulinum	758678	159513
Clostridium botulinum F str. Langeland	Clostridium botulinum	929506	66203
Clostridium botulinum H04402 065	Clostridium botulinum	941968	162091
Clostridium difficile 630	Clostridium difficile	272563	57679
Clostridium difficile BI1	Clostridium difficile	645462	41017
Clostridium difficile CD196	Clostridium difficile	645463	40921
Clostridium difficile R20291	Clostridium difficile	699034	158363
Clostridium perfringens ATCC 13124	Clostridium perfringens	195102	57681
Clostridium perfringens SM101	Clostridium perfringens	195103	57901
Clostridium perfringens str. 13	Clostridium perfringens	289380	58117
Clostridium tetani E88	Clostridium tetani	212717	57683
Comamonas testosteroni CNB-2	Comamonas testosteroni	688245	62961
Corynebacterium aurimucosum ATCC 700975	Corynebacterium aurimucosum	548476	59409
Corynebacterium diphtheriae 241	Corynebacterium diphtheriae	257309	57691
Corynebacterium diphtheriae 31A	Corynebacterium diphtheriae	698962	84309
Corynebacterium diphtheriae BH8	Corynebacterium diphtheriae	698963	84313
Corynebacterium diphtheriae C7 (beta)	Corynebacterium diphtheriae	698964	84303
Corynebacterium diphtheriae CDCE 8392	Corynebacterium diphtheriae	698965	84295
Corynebacterium diphtheriae HC01	Corynebacterium diphtheriae	698966	83607
Corynebacterium diphtheriae HC02	Corynebacterium diphtheriae	698967	84297
Corynebacterium diphtheriae HC03	Corynebacterium diphtheriae	698968	84317
Corynebacterium diphtheriae HC04	Corynebacterium diphtheriae	698969	84299
Corynebacterium diphtheriae INCA 402	Corynebacterium diphtheriae	698970	84301
Corynebacterium diphtheriae NCTC 13129	Corynebacterium diphtheriae	698971	84305
Corynebacterium diphtheriae PW8	Corynebacterium diphtheriae	698972	83605
Corynebacterium diphtheriae VA01	Corynebacterium diphtheriae	698973	84311
Corynebacterium jeikeium K411	Corynebacterium jeikeium	306537	58399
Corynebacterium kroppenstedtii DSM 44385	Corynebacterium kroppenstedtii	645127	59411
Corynebacterium pseudotuberculosis 1002	Corynebacterium pseudotuberculosis	1074485	89381
Corynebacterium pseudotuberculosis 1/06-A	Corynebacterium pseudotuberculosis	1087451	162167
Corynebacterium pseudotuberculosis 258	Corynebacterium pseudotuberculosis	1087452	83609
Corynebacterium pseudotuberculosis 267	Corynebacterium pseudotuberculosis	1087453	159669
Corynebacterium pseudotuberculosis 31	Corynebacterium pseudotuberculosis	1087454	159665

Corynebacterium pseudotuberculosis 316	Corynebacterium pseudotuberculosis	1089446	162175
Corynebacterium pseudotuberculosis 3/99-5	Corynebacterium pseudotuberculosis	1117942	157909
Corynebacterium pseudotuberculosis 42/02-A	Corynebacterium pseudotuberculosis	1161911	168258
Corynebacterium pseudotuberculosis C231	Corynebacterium pseudotuberculosis	1168865	167260
Corynebacterium pseudotuberculosis CIP 52.97	Corynebacterium pseudotuberculosis	679896	159677
Corynebacterium pseudotuberculosis Cp162	Corynebacterium pseudotuberculosis	681645	159675
Corynebacterium pseudotuberculosis FRC41	Corynebacterium pseudotuberculosis	765874	50585
Corynebacterium pseudotuberculosis I19	Corynebacterium pseudotuberculosis	889513	159673
Corynebacterium pseudotuberculosis P54B96	Corynebacterium pseudotuberculosis	935298	159671
Corynebacterium pseudotuberculosis PAT10	Corynebacterium pseudotuberculosis	935697	159667
Corynebacterium resistens DSM 45100	Corynebacterium resistens	662755	50555
Corynebacterium ulcerans 0102	Corynebacterium ulcerans	945711	159659
Corynebacterium ulcerans 809	Corynebacterium ulcerans	945712	68291
Corynebacterium ulcerans BR-AD22	Corynebacterium ulcerans	996634	169879
Corynebacterium urealyticum DSM 7109	Corynebacterium urealyticum	504474	61639
Cronobacter sakazakii ATCC BAA-894	Cronobacter sakazakii	1138308	167045
Cronobacter sakazakii ES15	Cronobacter sakazakii	290339	58145
Cronobacter turicensis z3032	Cronobacter turicensis	693216	40821
Cryptobacterium curtum DSM 15641	Cryptobacterium curtum	469378	59041
Cupriavidus metallidurans CH34	Cupriavidus metallidurans	266264	57815
Desulfovibrio desulfuricans ND132	Desulfovibrio desulfuricans	525146	59213
Desulfovibrio desulfuricans subsp. desulfuricans str. ATCC 27774	Desulfovibrio desulfuricans	641491	63159
Desulfovibrio vulgaris DP4	Desulfovibrio vulgaris	391774	58679
Desulfovibrio vulgaris RCH1	Desulfovibrio vulgaris	573059	161961
Desulfovibrio vulgaris str. Hildenborough	Desulfovibrio vulgaris	882	57645
Desulfovibrio vulgaris str. 'Miyazaki F'	Desulfovibrio vulgaris	883	59089
Dichelobacter nodosus VCS1703A	Dichelobacter nodosus	246195	57643
Edwardsiella tarda EIB202	Edwardsiella tarda	498217	41819
Edwardsiella tarda FL6-60	Edwardsiella tarda	718251	159657
Eggerthella lenta DSM 2243	Eggerthella lenta	479437	59079
Enterobacter aerogenes KCTC 2190	Enterobacter aerogenes	1028307	68103
Enterobacter asburiae LF7a	Enterobacter asburiae	640513	72793
Enterococcus faecalis D32	Enterococcus faecalis	1206105	171261
Enterococcus faecalis OG1RF	Enterococcus faecalis	226185	57669
Enterococcus faecalis V583	Enterococcus faecalis	474186	54927
Enterococcus faecium Aus0004	Enterococcus faecium	1155766	87025
Enterococcus faecium DO	Enterococcus faecium	333849	55353

Enterococcus hirae ATCC 9790	Enterococcus hirae	768486	70619
Escherichia coli 042	Escherichia coli	1033813	162139
Escherichia coli 536	Escherichia coli	1048689	162153
Escherichia coli 55989	Escherichia coli	1072459	162115
Escherichia coli ABU 83972	Escherichia coli	155864	57831
Escherichia coli APEC O1	Escherichia coli	199310	57915
Escherichia coli ATCC 8739	Escherichia coli	216592	161985
Escherichia coli BL21(DE3)	Escherichia coli	316385	58979
Escherichia coli 'BL21-Gold(DE3)pLysS AG'	Escherichia coli	316401	161993
Escherichia coli B str. REL606	Escherichia coli	316407	161931
Escherichia coli BW2952	Escherichia coli	331111	58395
Escherichia coli CFT073	Escherichia coli	331112	58393
Escherichia coli DH1	Escherichia coli	362663	58531
Escherichia coli E24377A	Escherichia coli	364106	58541
Escherichia coli ED1a	Escherichia coli	386585	57781
Escherichia coli ETEC H10407	Escherichia coli	405955	58623
Escherichia coli HS	Escherichia coli	409438	59425
Escherichia coli IAI1	Escherichia coli	413997	58803
Escherichia coli IAI39	Escherichia coli	431946	161939
Escherichia coli IHE3034	Escherichia coli	439855	58919
Escherichia coli KO11FL	Escherichia coli	444450	59091
Escherichia coli LF82	Escherichia coli	469008	161947
Escherichia coli NA114	Escherichia coli	481805	58783
Escherichia coli O103:H2 str. 12009	Escherichia coli	511145	57779
Escherichia coli O111:H- str. 11128	Escherichia coli	536056	161951
Escherichia coli O127:H6 str. E2348/69	Escherichia coli	544404	59235
Escherichia coli O157:H7 str. EC4115	Escherichia coli	566546	162011
Escherichia coli O157:H7 str. EDL933	Escherichia coli	573235	41021
Escherichia coli O157:H7 str. Sakai	Escherichia coli	574521	59343
Escherichia coli O157:H7 str. TW14359	Escherichia coli	585034	59377
Escherichia coli O26:H11 str. 11368	Escherichia coli	585035	62979
Escherichia coli O55:H7 str. CB9615	Escherichia coli	585055	59383
Escherichia coli O55:H7 str. RM12579	Escherichia coli	585056	62981
Escherichia coli O7:K1 str. CE10	Escherichia coli	585057	59381
Escherichia coli O83:H1 str. NRG 857C	Escherichia coli	585395	41013
Escherichia coli P12b	Escherichia coli	585396	41023
Escherichia coli S88	Escherichia coli	585397	59379

Escherichia coli SE11	Escherichia coli	591946	161965
Escherichia coli SE15	Escherichia coli	595495	162099
Escherichia coli SMS-3-5	Escherichia coli	595496	59391
Escherichia coli str. 'clone D i14'	Escherichia coli	655817	161975
Escherichia coli str. 'clone D i2'	Escherichia coli	685038	161987
Escherichia coli str. K-12 substr. DH10B	Escherichia coli	696406	161991
Escherichia coli str. K-12 substr. MG1655	Escherichia coli	701177	46655
Escherichia coli str. K-12 substr. W3110	Escherichia coli	714962	162007
Escherichia coli UM146	Escherichia coli	741093	163995
Escherichia coli UMN026	Escherichia coli	866768	59245
Escherichia coli UMNK88	Escherichia coli	869729	162043
Escherichia coli UTI89	Escherichia coli	885275	162049
Escherichia coli W	Escherichia coli	885276	162047
Escherichia coli Xuzhou21	Escherichia coli	910348	162061
Escherichia fergusonii ATCC 35469	Escherichia fergusonii	585054	59375
Eubacterium eligens ATCC 27750	Eubacterium eligens	515620	59171
Eubacterium limosum KIST612	Eubacterium limosum	903814	59777
Eubacterium rectale ATCC 33656	Eubacterium rectale	515619	59169
Fibrobacter succinogenes subsp. succinogenes S85	Fibrobacter succinogenes	59374	161919
Filifactor alocis ATCC 35896	Filifactor alocis	546269	46625
Finegoldia magna ATCC 29328	Finegoldia magna	334413	58867
Francisella philomiragia subsp. philomiragia ATCC 25017	Francisella philomiragia	484022	59105
Francisella tularensis subsp. holarctica FTNF002-00	Francisella tularensis	1001534	89373
Francisella tularensis subsp. holarctica LVS	Francisella tularensis	1001542	89379
Francisella tularensis subsp. holarctica OSU18	Francisella tularensis	177416	57589
Francisella tularensis subsp. mediasiatica FSC147	Francisella tularensis	376619	58595
Francisella tularensis subsp. tularensis FSC198	Francisella tularensis	393011	58687
Francisella tularensis subsp. tularensis NE061598	Francisella tularensis	393115	58693
Francisella tularensis subsp. tularensis SCHU S4	Francisella tularensis	418136	58811
Francisella tularensis subsp. tularensis TI0902	Francisella tularensis	441952	58939
Francisella tularensis subsp. tularensis TIGB03	Francisella tularensis	458234	58999
Francisella tularensis subsp. tularensis WY96-3418	Francisella tularensis	510831	161973
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	Fusobacterium nucleatum	190304	57885
Gardnerella vaginalis 409-05	Gardnerella vaginalis	1009464	162045
Gardnerella vaginalis ATCC 14019	Gardnerella vaginalis	525284	55487
Gardnerella vaginalis HMP9231	Gardnerella vaginalis	553190	43211
Gordonia bronchialis DSM 43247	Gordonia bronchialis	526226	41403

Gordonia polyisoprenivorans VH2	Gordonia polyisoprenivorans	1112204	86651
Haemophilus ducreyi 35000HP	Haemophilus ducreyi	233412	57625
Haemophilus influenzae 10810	Haemophilus influenzae	262727	161921
Haemophilus influenzae 86-028NP	Haemophilus influenzae	262728	161923
Haemophilus influenzae F3031	Haemophilus influenzae	281310	58093
Haemophilus influenzae F3047	Haemophilus influenzae	374930	58591
Haemophilus influenzae PittEE	Haemophilus influenzae	374931	58593
Haemophilus influenzae PittGG	Haemophilus influenzae	71421	57771
Haemophilus influenzae R2846	Haemophilus influenzae	862964	86647
Haemophilus influenzae R2866	Haemophilus influenzae	866630	62123
Haemophilus influenzae Rd KW20	Haemophilus influenzae	935897	62097
Haemophilus parainfluenzae T3T1	Haemophilus parainfluenzae	862965	72801
Helicobacter bizzozeronii CIII-1	Helicobacter bizzozeronii	1002804	68141
Helicobacter cinaedi PAGU611	Helicobacter cinaedi	1172562	162219
Helicobacter felis ATCC 49179	Helicobacter felis	936155	61409
Helicobacter pylori 2017	Helicobacter pylori	102608	161139
Helicobacter pylori 2018	Helicobacter pylori	1055527	158157
Helicobacter pylori 26695	Helicobacter pylori	1055528	159611
Helicobacter pylori 35A	Helicobacter pylori	1055529	161157
Helicobacter pylori 51	Helicobacter pylori	1055530	159615
Helicobacter pylori 83	Helicobacter pylori	1127122	165869
Helicobacter pylori 908	Helicobacter pylori	1163739	162205
Helicobacter pylori B38	Helicobacter pylori	1163740	162207
Helicobacter pylori B8	Helicobacter pylori	1163741	162209
Helicobacter pylori Cuz20	Helicobacter pylori	1163742	162211
Helicobacter pylori ELS37	Helicobacter pylori	1163743	162213
Helicobacter pylori F16	Helicobacter pylori	290847	161925
Helicobacter pylori F30	Helicobacter pylori	357544	58517
Helicobacter pylori F32	Helicobacter pylori	512562	59165
Helicobacter pylori F57	Helicobacter pylori	563041	59305
Helicobacter pylori G27	Helicobacter pylori	570508	59327
Helicobacter pylori Gambia94/24	Helicobacter pylori	585535	49903
Helicobacter pylori HPAG1	Helicobacter pylori	585538	161153
Helicobacter pylori HUP-B14	Helicobacter pylori	592205	59415
Helicobacter pylori India7	Helicobacter pylori	637913	159639
Helicobacter pylori J99	Helicobacter pylori	693745	49873
Helicobacter pylori Lithuania75	Helicobacter pylori	765962	53541

Helicobacter pylori P12	Helicobacter pylori	765963	53539
Helicobacter pylori PeCan18	Helicobacter pylori	765964	159987
Helicobacter pylori PeCan4	Helicobacter pylori	794851	159467
Helicobacter pylori Puno120	Helicobacter pylori	85962	178201
Helicobacter pylori Puno135	Helicobacter pylori	85963	57789
Helicobacter pylori Sat464	Helicobacter pylori	866344	161145
Helicobacter pylori Shi112	Helicobacter pylori	866345	159991
Helicobacter pylori Shi169	Helicobacter pylori	866346	161143
Helicobacter pylori Shi417	Helicobacter pylori	869727	159985
Helicobacter pylori Shi470	Helicobacter pylori	907237	159491
Helicobacter pylori SJM180	Helicobacter pylori	907238	161149
Helicobacter pylori SNT49	Helicobacter pylori	907239	159989
Helicobacter pylori SouthAfrica7	Helicobacter pylori	907240	159493
Helicobacter pylori v225d	Helicobacter pylori	985080	161159
Helicobacter pylori XZ274	Helicobacter pylori	985081	161151
Klebsiella oxytoca E718	Klebsiella oxytoca	1006551	83159
Klebsiella oxytoca KCTC 1686	Klebsiella oxytoca	1191061	170256
Klebsiella variicola At-22	Klebsiella variicola	640131	42113
Kocuria rhizophila DC2201	Kocuria rhizophila	378753	59099
Kytococcus sedentarius DSM 20547	Kytococcus sedentarius	478801	59071
Lactobacillus acidophilus 30SC	Lactobacillus acidophilus	272621	57685
Lactobacillus acidophilus NCFM	Lactobacillus acidophilus	891391	63605
Lactobacillus brevis ATCC 367	Lactobacillus brevis	387344	57989
Lactobacillus buchneri NRRL B-30929	Lactobacillus buchneri	511437	66205
Lactobacillus casei ATCC 334	Lactobacillus casei	321967	57985
Lactobacillus casei BD-II	Lactobacillus casei	498216	50673
Lactobacillus casei BL23	Lactobacillus casei	543734	59237
Lactobacillus casei LC2W	Lactobacillus casei	998820	162119
Lactobacillus casei str. Zhang	Lactobacillus casei	999378	162121
Lactobacillus crispatus ST1	Lactobacillus crispatus	748671	48359
Lactobacillus fermentum CECT 5716	Lactobacillus fermentum	334390	58865
Lactobacillus fermentum IFO 3956	Lactobacillus fermentum	712938	162003
Lactobacillus gasseri ATCC 33323	Lactobacillus gasseri	324831	57687
Lactobacillus johnsonii DPC 6026	Lactobacillus johnsonii	257314	58029
Lactobacillus johnsonii FI9785	Lactobacillus johnsonii	633699	41735
Lactobacillus johnsonii NCC 533	Lactobacillus johnsonii	909954	162057
Lactobacillus plantarum JDM1	Lactobacillus plantarum	220668	62911

Lactobacillus plantarum subsp. plantarum ST-III	Lactobacillus plantarum	644042	59361
Lactobacillus plantarum WCFS1	Lactobacillus plantarum	889932	53537
Lactobacillus reuteri DSM 20016	Lactobacillus reuteri	491077	55357
Lactobacillus reuteri JCM 1112	Lactobacillus reuteri	557433	58875
Lactobacillus reuteri SD2112	Lactobacillus reuteri	557436	58471
Lactobacillus rhamnosus ATCC 8530	Lactobacillus rhamnosus	1088720	162169
Lactobacillus rhamnosus GG	Lactobacillus rhamnosus	568703	161983
Lactobacillus rhamnosus Lc 705	Lactobacillus rhamnosus	568704	59315
Lactobacillus salivarius CECT 5713	Lactobacillus salivarius	362948	58233
Lactobacillus salivarius UCC118	Lactobacillus salivarius	712961	162005
Lactococcus garvieae ATCC 49156	Lactococcus garvieae	420889	73413
Lactococcus garvieae Lg2	Lactococcus garvieae	420890	161935
Laribacter hongkongensis HLHK9	Laribacter hongkongensis	557598	59265
Legionella longbeachae NSW150	Legionella longbeachae	661367	46099
Legionella pneumophila 2300/99 Alcoy	Legionella pneumophila	272624	57609
Legionella pneumophila str. Corby	Legionella pneumophila	297245	58209
Legionella pneumophila str. Lens	Legionella pneumophila	297246	58211
Legionella pneumophila str. Paris	Legionella pneumophila	400673	58733
Legionella pneumophila subsp. pneumophila ATCC 43290	Legionella pneumophila	423212	48801
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	Legionella pneumophila	933093	86885
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130	Leptospira interrogans	189518	57881
Leptospira interrogans serovar Lai str. 56601	Leptospira interrogans	267671	58065
Leptospira interrogans serovar Lai str. IPAV	Leptospira interrogans	573825	161957
Leptotrichia buccalis C-1013-b	Leptotrichia buccalis	523794	59211
Leuconostoc citreum KM20	Leuconostoc citreum	349519	58481
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	Leuconostoc mesenteroides	1107880	84337
Leuconostoc mesenteroides subsp. mesenteroides J18	Leuconostoc mesenteroides	203120	57919
Listeria innocua Clip11262	Listeria innocua	272626	61567
Listeria monocytogenes 07PF0776	Listeria monocytogenes	1030009	162131
Listeria monocytogenes 08-5578	Listeria monocytogenes	1126011	162185
Listeria monocytogenes 08-5923	Listeria monocytogenes	169963	61583
Listeria monocytogenes 10403S	Listeria monocytogenes	265669	57689
Listeria monocytogenes EGD-e	Listeria monocytogenes	393126	54441
Listeria monocytogenes Finland 1998	Listeria monocytogenes	393127	54443
Listeria monocytogenes FSL R2-561	Listeria monocytogenes	393130	54459
Listeria monocytogenes HCC23	Listeria monocytogenes	393133	54461
Listeria monocytogenes J0161	Listeria monocytogenes	552536	59203

Listeria monocytogenes M7	Listeria monocytogenes	568819	59317
Listeria monocytogenes serotype 4b str. CLIP 80459	Listeria monocytogenes	637381	43727
Listeria monocytogenes serotype 4b str. F2365	Listeria monocytogenes	653938	43671
Listeria seeligeri serovar 1/2b str. SLCC3954	Listeria seeligeri	683837	46215
Listeria welshimeri serovar 6b str. SLCC5334	Listeria welshimeri	386043	61605
Lysinibacillus sphaericus C3-41	Lysinibacillus sphaericus	444177	58945
Micrococcus luteus NCTC 2665	Micrococcus luteus	465515	59033
Mobiluncus curtisii ATCC 43063	Mobiluncus curtisii	548479	49695
Moraxella catarrhalis RH4	Moraxella catarrhalis	749219	48809
Mycobacterium leprae Br4923	Mycobacterium leprae	272631	57697
Mycobacterium leprae TN	Mycobacterium leprae	561304	59293
Mycoplasma fermentans JER	Mycoplasma fermentans	637387	53543
Mycoplasma fermentans M64	Mycoplasma fermentans	943945	62099
Mycoplasma genitalium G37	Mycoplasma genitalium	243273	57707
Mycoplasma penetrans HF-2	Mycoplasma penetrans	272633	57729
Mycoplasma pneumoniae 309	Mycoplasma pneumoniae	1112856	85495
Mycoplasma pneumoniae FH	Mycoplasma pneumoniae	272634	57709
Mycoplasma pneumoniae M129	Mycoplasma pneumoniae	722438	162027
Neisseria gonorrhoeae FA 1090	Neisseria gonorrhoeae	242231	57611
Neisseria gonorrhoeae NCCP11945	Neisseria gonorrhoeae	521006	59191
Neisseria gonorrhoeae TCDC-NG08107	Neisseria gonorrhoeae	940296	161097
Neisseria lactamica 020-06	Neisseria lactamica	489653	60851
Neisseria meningitidis 053442	Neisseria meningitidis	122586	57817
Neisseria meningitidis 8013	Neisseria meningitidis	122587	57819
Neisseria meningitidis alpha14	Neisseria meningitidis	272831	57825
Neisseria meningitidis alpha710	Neisseria meningitidis	374833	58587
Neisseria meningitidis FAM18	Neisseria meningitidis	604162	161967
Neisseria meningitidis G2136	Neisseria meningitidis	630588	161971
Neisseria meningitidis H44/76	Neisseria meningitidis	662598	61649
Neisseria meningitidis M01-240149	Neisseria meningitidis	909420	162083
Neisseria meningitidis M01-240355	Neisseria meningitidis	935588	162075
Neisseria meningitidis M04-240196	Neisseria meningitidis	935589	162077
Neisseria meningitidis MC58	Neisseria meningitidis	935591	162079
Neisseria meningitidis NZ-05/33	Neisseria meningitidis	935593	162081
Neisseria meningitidis WUE 2594	Neisseria meningitidis	935599	162085
Neisseria meningitidis Z2491	Neisseria meningitidis	942513	162093
Nocardia farcinica IFM 10152	Nocardia farcinica	247156	58203

Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111	Nocardiopsis dassonvillei	446468	49483
Ochrobactrum anthropi ATCC 49188	Ochrobactrum anthropi	439375	58921
Odoribacter splanchnicus DSM 220712	Odoribacter splanchnicus	709991	63397
Olsenella uli DSM 7084	Olsenella uli	633147	51367
Paenibacillus polymyxa E681	Paenibacillus polymyxa	1052684	162159
Paenibacillus polymyxa M1	Paenibacillus polymyxa	349520	53477
Paenibacillus polymyxa SC2	Paenibacillus polymyxa	886882	59583
Parabacteroides distasonis ATCC 8503	Parabacteroides distasonis	435591	58301
Pasteurella multocida 36950	Pasteurella multocida	1075089	86887
Pasteurella multocida subsp. multocida str. 3480	Pasteurella multocida	1132496	156881
Pasteurella multocida subsp. multocida str. HN06	Pasteurella multocida	272843	57627
Pasteurella multocida subsp. multocida str. Pm70	Pasteurella multocida	584721	161955
Pediococcus pentosaceus ATCC 25745	Pediococcus pentosaceus	278197	57981
Porphyromonas asaccharolytica DSM 20707	Porphyromonas asaccharolytica	879243	66603
Porphyromonas gingivalis ATCC 33277	Porphyromonas gingivalis	1030843	67407
Porphyromonas gingivalis TDC60	Porphyromonas gingivalis	242619	57641
Porphyromonas gingivalis W83	Porphyromonas gingivalis	431947	58879
Prevotella denticola F0289	Prevotella denticola	767031	65091
Prevotella intermedia 17	Prevotella intermedia	246198	163151
Prevotella melaninogenica ATCC 25845	Prevotella melaninogenica	553174	51377
Propionibacterium acnes 266	Propionibacterium acnes	1031709	162137
Propionibacterium acnes 6609	Propionibacterium acnes	1091045	162177
Propionibacterium acnes ATCC 11828	Propionibacterium acnes	1114966	80745
Propionibacterium acnes KPA171202	Propionibacterium acnes	1114967	80735
Propionibacterium acnes SK137	Propionibacterium acnes	1114969	80733
Propionibacterium acnes TypeIA2 P.acn17	Propionibacterium acnes	267747	58101
Propionibacterium acnes TypeIA2 P.acn31	Propionibacterium acnes	553199	48071
Propionibacterium acnes TypeIA2 P.acn33	Propionibacterium acnes	909952	162059
Proteus mirabilis HI4320	Proteus mirabilis	529507	61599
Providencia stuartii MRSN 2154	Providencia stuartii	1157951	162193
Pseudomonas aeruginosa DK2	Pseudomonas aeruginosa	1089456	162173
Pseudomonas aeruginosa LESB58	Pseudomonas aeruginosa	1093787	168996
Pseudomonas aeruginosa M18	Pseudomonas aeruginosa	208963	57977
Pseudomonas aeruginosa NCGM2.S1	Pseudomonas aeruginosa	208964	57945
Pseudomonas aeruginosa PA7	Pseudomonas aeruginosa	381754	58627
Pseudomonas aeruginosa PAO1	Pseudomonas aeruginosa	557722	59275
Pseudomonas aeruginosa UCBPP-PA14	Pseudomonas aeruginosa	941193	162089

Pseudomonas fluorescens A506	Pseudomonas fluorescens	1037911	165185
Pseudomonas fluorescens F113	Pseudomonas fluorescens	1114970	87037
Pseudomonas fluorescens Pf0-1	Pseudomonas fluorescens	205922	57591
Pseudomonas fluorescens SBW25	Pseudomonas fluorescens	216595	158693
Pseudomonas mendocina NK-01	Pseudomonas mendocina	1001585	66299
Pseudomonas mendocina ymp	Pseudomonas mendocina	399739	58723
Pseudomonas putida BIRD-1	Pseudomonas putida	1042876	68747
Pseudomonas putida DOT-T1E	Pseudomonas putida	1196325	171260
Pseudomonas putida F1	Pseudomonas putida	160488	57843
Pseudomonas putida GB-1	Pseudomonas putida	231023	167583
Pseudomonas putida KT2440	Pseudomonas putida	351746	58355
Pseudomonas putida ND6	Pseudomonas putida	390235	58651
Pseudomonas putida S16	Pseudomonas putida	76869	58735
Pseudomonas putida W619	Pseudomonas putida	931281	162055
Pseudomonas stutzeri A1501	Pseudomonas stutzeri	1123519	170940
Pseudomonas stutzeri ATCC 17588 = LMG 11199	Pseudomonas stutzeri	1196835	168379
Pseudomonas stutzeri CCUG 29243	Pseudomonas stutzeri	379731	58641
Pseudomonas stutzeri DSM 10701	Pseudomonas stutzeri	96563	68749
Pseudomonas stutzeri DSM 4166	Pseudomonas stutzeri	996285	162113
Rahnella aquatilis CIP 78.65 = ATCC 33071	Rahnella aquatilis	1151116	158049
Rahnella aquatilis HX2	Rahnella aquatilis	745277	86855
Ralstonia pickettii 12D	Ralstonia pickettii	402626	58737
Ralstonia pickettii 12J	Ralstonia pickettii	428406	58859
Rhodococcus equi 103S	Rhodococcus equi	685727	60171
Rhodococcus erythropolis PR4	Rhodococcus erythropolis	234621	59019
Roseburia hominis A2-183	Roseburia hominis	585394	73419
Rothia dentocariosa ATCC 17931	Rothia dentocariosa	762948	49331
Rothia mucilaginosa DY-18	Rothia mucilaginosa	680646	43093
Salmonella bongori NCTC 12419	Salmonella bongori	218493	70155
Sebaldella termitidis ATCC 33386	Sebaldella termitidis	526218	41865
Selenomonas sputigena ATCC 35185	Selenomonas sputigena	546271	55329
Serratia plymuthica AS9	Serratia plymuthica	768492	67313
Shigella boydii CDC 3083-94	Shigella boydii	300268	58215
Shigella boydii Sb227	Shigella boydii	344609	58415
Shigella dysenteriae Sd197	Shigella dysenteriae	300267	58213
Shigella flexneri 2002017	Shigella flexneri	198214	62907
Shigella flexneri 2a str. 2457T	Shigella flexneri	198215	57991

Shigella flexneri 2a str. 301	Shigella flexneri	373384	58583
Shigella flexneri 5 str. 8401	Shigella flexneri	591020	159233
Shigella sonnei Ss046	Shigella sonnei	300269	58217
Slackia heliotrinireducens DSM 20476	Slackia heliotrinireducens	471855	59051
Staphylococcus aureus 04-02981	Staphylococcus aureus	1006543	159861
Staphylococcus aureus RF122	Staphylococcus aureus	1028799	88071
Staphylococcus aureus subsp. aureus 11819-97	Staphylococcus aureus	1074252	162163
Staphylococcus aureus subsp. aureus 71193	Staphylococcus aureus	1118959	88065
Staphylococcus aureus subsp. aureus COL	Staphylococcus aureus	1123523	159981
Staphylococcus aureus subsp. aureus ECT-R 2	Staphylococcus aureus	1155084	162141
Staphylococcus aureus subsp. aureus ED133	Staphylococcus aureus	158878	57835
Staphylococcus aureus subsp. aureus ED98	Staphylococcus aureus	158879	57837
Staphylococcus aureus subsp. aureus HO 5096 0412	Staphylococcus aureus	196620	57903
Staphylococcus aureus subsp. aureus JH1	Staphylococcus aureus	273036	57661
Staphylococcus aureus subsp. aureus JH9	Staphylococcus aureus	282458	57839
Staphylococcus aureus subsp. aureus JKD6159	Staphylococcus aureus	282459	57841
Staphylococcus aureus subsp. aureus LGA251	Staphylococcus aureus	359786	58455
Staphylococcus aureus subsp. aureus M013	Staphylococcus aureus	359787	58457
Staphylococcus aureus subsp. aureus MRSA252	Staphylococcus aureus	418127	58817
Staphylococcus aureus subsp. aureus MSHR1132	Staphylococcus aureus	426430	58839
Staphylococcus aureus subsp. aureus MSSA476	Staphylococcus aureus	451515	58555
Staphylococcus aureus subsp. aureus Mu3	Staphylococcus aureus	451516	58925
Staphylococcus aureus subsp. aureus Mu50	Staphylococcus aureus	546342	159855
Staphylococcus aureus subsp. aureus MW2	Staphylococcus aureus	548473	159859
Staphylococcus aureus subsp. aureus N315	Staphylococcus aureus	663951	159241
Staphylococcus aureus subsp. aureus NCTC 8325	Staphylococcus aureus	681288	41455
Staphylococcus aureus subsp. aureus str. JKD6008	Staphylococcus aureus	685039	159689
Staphylococcus aureus subsp. aureus str. Newman	Staphylococcus aureus	703339	161969
Staphylococcus aureus subsp. aureus T0131	Staphylococcus aureus	869816	159691
Staphylococcus aureus subsp. aureus TCH60	Staphylococcus aureus	889933	159389
Staphylococcus aureus subsp. aureus TW20	Staphylococcus aureus	93061	57795
Staphylococcus aureus subsp. aureus USA300_FPR3757	Staphylococcus aureus	93062	57797
Staphylococcus aureus subsp. aureus USA300_TCH1516	Staphylococcus aureus	985002	89393
Staphylococcus aureus subsp. aureus VC40	Staphylococcus aureus	985006	159391
Staphylococcus carnosus subsp. carnosus TM300	Staphylococcus carnosus	396513	59401
Staphylococcus epidermidis ATCC 12228	Staphylococcus epidermidis	176279	57663
Staphylococcus epidermidis RP62A	Staphylococcus epidermidis	176280	57861

Staphylococcus haemolyticus JCSC1435	279808	62919
Staphylococcus lugdunensis HKU09-01	1034809	162143
Staphylococcus lugdunensis N920143	698737	46233
Staphylococcus pseudintermedius ED99	937773	62125
Staphylococcus pseudintermedius HKU10-03	984892	162109
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	342451	58411
Stenotrophomonas maltophilia D457	1163399	162199
Stenotrophomonas maltophilia JV3	391008	58657
Stenotrophomonas maltophilia K279a	522373	61647
Stenotrophomonas maltophilia R551-3	868597	72473
Streptobacillus moniliformis DSM 12112	519441	41863
Streptococcus agalactiae 2603V/R	205921	57935
Streptococcus agalactiae A909	208435	57943
Streptococcus agalactiae NEM316	211110	61585
Streptococcus dysgalactiae subsp. equisimilis ATCC 12394	486410	59103
Streptococcus dysgalactiae subsp. equisimilis GGS_124	663954	161979
Streptococcus equi subsp. equi 4047	40041	59261
Streptococcus equi subsp. zooepidemicus	552526	59263
Streptococcus equi subsp. zooepidemicus MGCS10565	553482	59259
Streptococcus gallolyticus subsp. gallolyticus ATCC 43143	637909	46061
Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069	981539	162103
Streptococcus gallolyticus UCN34	990317	63617
Streptococcus gordonii str. Challis substr. CH1	467705	57667
Streptococcus infantarius subsp. infantarius CJ18	1069533	87033
Streptococcus intermedius JTH08	591365	168614
Streptococcus mitis B6	365659	46097
Streptococcus oralis Uo5	927666	65449
Streptococcus pasteurianus ATCC 43144	981540	68019
Streptococcus pneumoniae 670-6B	1130804	162191
Streptococcus pneumoniae 70585	170187	57857
Streptococcus pneumoniae AP200	171101	57859
Streptococcus pneumoniae ATCC 700669	189423	52533
Streptococcus pneumoniae CGSP14	373153	58581
Streptococcus pneumoniae D39	487213	59119
Streptococcus pneumoniae G54	487214	59117
Streptococcus pneumoniae Hungary19A-6	488221	59125
Streptococcus pneumoniae INV104	488222	59121

Streptococcus pneumoniae INV200	Streptococcus pneumoniae	488223	59123
Streptococcus pneumoniae JJA	Streptococcus pneumoniae	512566	59167
Streptococcus pneumoniae OXC141	Streptococcus pneumoniae	516950	59181
Streptococcus pneumoniae P1031	Streptococcus pneumoniae	525381	49735
Streptococcus pneumoniae R6	Streptococcus pneumoniae	561276	59287
Streptococcus pneumoniae ST556	Streptococcus pneumoniae	574093	52453
Streptococcus pneumoniae Taiwan19F-14	Streptococcus pneumoniae	869215	162037
Streptococcus pneumoniae TCH8431/19A	Streptococcus pneumoniae	869216	162035
Streptococcus pneumoniae TIGR4	Streptococcus pneumoniae	869269	162039
Streptococcus pseudopneumoniae IS7493	Streptococcus pseudopneumoniae	1054460	71153
Streptococcus pyogenes Alab49	Streptococcus pyogenes	1010840	158061
Streptococcus pyogenes M1 GAS	Streptococcus pyogenes	160490	57845
Streptococcus pyogenes MGAS10270	Streptococcus pyogenes	160491	57847
Streptococcus pyogenes MGAS10394	Streptococcus pyogenes	186103	57871
Streptococcus pyogenes MGAS10750	Streptococcus pyogenes	193567	57895
Streptococcus pyogenes MGAS15252	Streptococcus pyogenes	198466	57911
Streptococcus pyogenes MGAS1882	Streptococcus pyogenes	286636	58105
Streptococcus pyogenes MGAS2096	Streptococcus pyogenes	293653	58337
Streptococcus pyogenes MGAS315	Streptococcus pyogenes	319701	58335
Streptococcus pyogenes MGAS5005	Streptococcus pyogenes	370551	58569
Streptococcus pyogenes MGAS6180	Streptococcus pyogenes	370552	58571
Streptococcus pyogenes MGAS8232	Streptococcus pyogenes	370553	58573
Streptococcus pyogenes MGAS9429	Streptococcus pyogenes	370554	58575
Streptococcus pyogenes NZ131	Streptococcus pyogenes	471876	59035
Streptococcus pyogenes SSI-1	Streptococcus pyogenes	487215	162171
Streptococcus pyogenes str. Manfredo	Streptococcus pyogenes	798300	158037
Streptococcus salivarius 57.I	Streptococcus salivarius	1046629	162151
Streptococcus salivarius JIM8777	Streptococcus salivarius	347253	162145
Streptococcus uberis 0140J	Streptococcus uberis	218495	57959
Streptomyces griseus subsp. griseus NBRC 13350	Streptomyces griseus	455632	58983
Treponema denticola ATCC 35405	Treponema denticola	243275	57583
Treponema pallidum subsp. pallidum DAL-1	Treponema pallidum	243276	208669
Treponema pallidum subsp. pallidum SS14	Treponema pallidum	455434	58977
Treponema pallidum subsp. pallidum str. Chicago	Treponema pallidum	491078	87069
Treponema pallidum subsp. pallidum str. Nichols	Treponema pallidum	491079	87051
Treponema pallidum subsp. pertenue str. CDC2	Treponema pallidum	491080	87067
Treponema pallidum subsp. pertenue str. Gauthier	Treponema pallidum	491081	87065

Treponema pallidum subsp. pertenue str. SamoaD	Treponema pallidum	666714	159543
Tropheryma whipplei str. Twist	Tropheryma whipplei	203267	57705
Tropheryma whipplei TW08/27	Tropheryma whipplei	218496	57961
Tsukamurella paurometabola DSM 20162	Tsukamurella paurometabola	521096	48829
Ureaplasma parvum serovar 3 str. ATCC 27815	Ureaplasma parvum	273119	57711
Ureaplasma parvum serovar 3 str. ATCC 700970	Ureaplasma parvum	505682	58887
Ureaplasma urealyticum serovar 10 str. ATCC 33699	Ureaplasma urealyticum	565575	59011
Veillonella parvula DSM 2008	Veillonella parvula	479436	41927
Vibrio cholerae IEC224	Vibrio cholerae	1134456	89389
Vibrio cholerae LMA3984-4	Vibrio cholerae	243277	57623
Vibrio cholerae M66-2	Vibrio cholerae	345073	159869
Vibrio cholerae MJ-1236	Vibrio cholerae	579112	59355
Vibrio cholerae O1 biovar El Tor str. N16961	Vibrio cholerae	593588	59387
Vibrio cholerae O1 str. 2010EL-1786	Vibrio cholerae	914149	78933
Vibrio cholerae O395	Vibrio cholerae	935297	159541
Vibrio furnissii NCTC 11218	Vibrio furnissii	903510	82347
Vibrio parahaemolyticus RIMD 2210633	Vibrio parahaemolyticus	223926	57969
Vibrio vulnificus CMCP6	Vibrio vulnificus	196600	58007
Vibrio vulnificus MO6-24/O	Vibrio vulnificus	216895	62909
Vibrio vulnificus YJ016	Vibrio vulnificus	914127	62243
Weeksella virosa DSM 16922	Weeksella virosa	865938	63627
Wolinella succinogenes DSM 1740	Wolinella succinogenes	273121	61591
Yersinia enterocolitica subsp. enterocolitica 8081	Yersinia enterocolitica	393305	57741
Yersinia enterocolitica subsp. palearctica 105.5R(r)	Yersinia enterocolitica	930944	162069
Yersinia enterocolitica subsp. palearctica Y11	Yersinia enterocolitica	994476	63663
Yersinia pestis A1122	Yersinia pestis	1035377	158119
Yersinia pestis Angola	Yersinia pestis	187410	57875
Yersinia pestis Antiqua	Yersinia pestis	214092	57621
Yersinia pestis biovar Medievalis str. Harbin 35	Yersinia pestis	229193	58037
Yersinia pestis biovar Microtus str. 91001	Yersinia pestis	349746	58485
Yersinia pestis CO92	Yersinia pestis	360102	58607
Yersinia pestis D106004	Yersinia pestis	377628	58609
Yersinia pestis D182038	Yersinia pestis	386656	58619
Yersinia pestis KIM10+	Yersinia pestis	547048	158537
Yersinia pestis Nepal516	Yersinia pestis	637382	158071
Yersinia pestis Pestoides F	Yersinia pestis	637385	158073
Yersinia pestis Z176003	Yersinia pestis	637386	47317

Yersinia pseudotuberculosis IP 31758	Yersinia pseudotuberculosis	273123	58157
Yersinia pseudotuberculosis IP 32953	Yersinia pseudotuberculosis	349747	58487
Yersinia pseudotuberculosis PB1/+	Yersinia pseudotuberculosis	502800	59151
Yersinia pseudotuberculosis YPIII	Yersinia pseudotuberculosis	502801	59153

(a) Strain designation
(b) Species name
(c) NCBI bioproject id of the sequencing project
(d) NCBI taxonomy strain id

Supplementary Table S4 Mapping of bacterial strains to 42 species in the Global Infectious Disease and Epidemiology Online Network with links to the National Center for Biotechnology Information (NCBI) databases

Strain[(a)]	Species[(b)]	NCBI Bioproject id[(c)]	NCBI taxonomy id[(d)]
Actinobacillus suis H91-0380	Actinobacillus suis	176363	696748
Adlercreutzia equolifaciens DSM 19450	Adlercreutzia equolifaciens	223286	1384484
Alistipes shahii WAL 8301	Alistipes shahii	197175	717959
Bacillus infantis NRRL B-14911	Bacillus infantis	222804	1367477
Bacteroides xylanisolvens XB1A	Bacteroides xylanisolvens	197168	657309
Burkholderia cepacia GG4	Burkholderia cepacia	173858	1009846
Butyrivibrio fibrisolvens	Butyrivibrio fibrisolvens	197155	831
Campylobacter coli 15-537360	Campylobacter coli	226113	1358410
Campylobacter coli CVM N29710	Campylobacter coli	219322	1273173
Corynebacterium argentoratense DSM 44202	Corynebacterium argentoratense	217419	1348662
Enterococcus casseliflavus EC20	Enterococcus casseliflavus	55693	565655
Enterococcus mundtii QU 25	Enterococcus mundtii	229420	1300150
Erysipelothrix rhusiopathiae	Erysipelothrix rhusiopathiae	68021	1648
Erysipelothrix rhusiopathiae SY1027	Erysipelothrix rhusiopathiae	206518	1313290
Eubacterium siraeum	Eubacterium siraeum	197160	39492
Eubacterium siraeum V10Sc8a	Eubacterium siraeum	197178	717961
Faecalibacterium prausnitzii	Faecalibacterium prausnitzii	197157	853
Faecalibacterium prausnitzii L2-6	Faecalibacterium prausnitzii	197183	718252
Fretibacterium fastidiosum	Fretibacterium fastidiosum	197182	651822
Gordonibacter pamelaeae 7-10-1-b	Gordonibacter pamelaeae	197167	657308
Lactobacillus paracasei subsp. paracasei 8700:2	Lactobacillus paracasei	55295	537973
Listeria ivanovii	Listeria ivanovii	73473	1638
Mannheimia haemolytica D153	Mannheimia haemolytica	212303	1261126
Mannheimia haemolytica D171	Mannheimia haemolytica	212304	1311759
Mannheimia haemolytica D174	Mannheimia haemolytica	212305	1311760
Mannheimia haemolytica M42548	Mannheimia haemolytica	198769	1316932
Mannheimia haemolytica USDA-ARS-USMARC-183	Mannheimia haemolytica	195458	1249531
Mannheimia haemolytica USDA-ARS-USMARC-185	Mannheimia haemolytica	195457	1249526
Mannheimia haemolytica USMARC_2286	Mannheimia haemolytica	213228	1366053
Megasphaera elsdenii	Megasphaera elsdenii	71135	907

Mycoplasma hominis	Mycoplasma hominis	41875	2098
Nocardia brasiliensis ATCC 700358	Nocardia brasiliensis	86913	1133849
Nocardia cyriacigeorgica	Nocardia cyriacigeorgica	89395	135487
Pandoraea pnomenusa 3kgm	Pandoraea pnomenusa	229878	1416914
Photorhabdus asymbiotica subsp. asymbiotica ATCC 43949	Photorhabdus asymbiotica	59243	553480
Prevotella dentalis DSM 3688	Prevotella dentalis	184818	908937
Propionibacterium avidum 44067	Propionibacterium avidum	197361	1170318
Pseudomonas monteilii SB3078	Pseudomonas monteilii	232252	1435044
Pseudomonas monteilii SB3101	Pseudomonas monteilii	232253	1435058
Raoultella ornithinolytica B6	Raoultella ornithinolytica	198431	1286170
Roseburia intestinalis	Roseburia intestinalis	197164	166486
Roseburia intestinalis XB6B4	Roseburia intestinalis	197179	718255
Ruminococcus bromii	Ruminococcus bromii	197158	40518
Ruminococcus champanellensis 18P13	Ruminococcus champanellensis	197169	213810
Serratia marcescens FGI94	Serratia marcescens	185180	1249634
Serratia marcescens WW4	Serratia marcescens	188478	435998
Staphylococcus pasteuri SP1	Staphylococcus pasteuri	226267	1276282
Staphylococcus warneri SG1	Staphylococcus warneri	187059	1194526
Streptococcus anginosus C1051	Streptococcus anginosus	218003	862970
Streptococcus anginosus C238	Streptococcus anginosus	218004	862971
Streptococcus constellatus subsp. pharyngis C1050	Streptococcus constellatus	218002	862969
Streptococcus constellatus subsp. pharyngis C232	Streptococcus constellatus	217998	696216
Streptococcus constellatus subsp. pharyngis C818	Streptococcus constellatus	218001	862968
Streptococcus iniae SF1	Streptococcus iniae	206041	1318633
Streptococcus lutetiensis 033	Streptococcus lutetiensis	213397	1076934
Streptococcus oligofermentans AS 1.3089	Streptococcus oligofermentans	201429	1302863
Streptomyces albus J1074	Streptomyces albus	196849	457425
Vibrio alginolyticus NBRC 15630 = ATCC 17749	Vibrio alginolyticus	199933	1219076

(a) Strain designation

(b) Species name

(c) NCBI bioproject id of the sequencing project

(d) NCBI taxonomy strain id

Supplementary Table S5 Mapping of bacterial strains to 296 species described in Bergey's Manual of Systematic Bacteriology

Strain[(a)]	Species[(b)]	NCBI taxonomy id[(d)]
[Bacteroides] pectinophilus ATCC 43243	[Bacteroides] pectinophilus	483218.5
[Clostridium] mangenotii LM2	[Clostridium] mangenotii	1392497.3
[Clostridium] mangenotii TR	[Clostridium] mangenotii	1408823.3
[Eubacterium] cylindroides ATCC 27803	Faecalitalea cylindroides	649755.3
[Eubacterium] cylindroides T2-87	Faecalitalea cylindroides	717960.3
Abiotrophia defectiva ATCC 49176	Abiotrophia defectiva	592010.4
Acetobacterium woodii DSM 1030	Acetobacterium woodii	931626.3
Acidithiobacillus caldus ATCC 51756	Acidithiobacillus caldus	637389.3
Acidithiobacillus caldus SM-1	Acidithiobacillus caldus	990288.8
Acidithiobacillus ferrooxidans ATCC 23270	Acidithiobacillus ferrooxidans	243159.4
Acidithiobacillus ferrooxidans ATCC 53993	Acidithiobacillus ferrooxidans	380394.4
Acidithiobacillus thiooxidans	Acidithiobacillus thiooxidans	930.4
Acidithiobacillus thiooxidans A01	Acidithiobacillus thiooxidans	1432062.4
Acidithiobacillus thiooxidans ATCC 19377	Acidithiobacillus thiooxidans	637390.5
Aerococcus viridans ATCC 11563	Aerococcus viridans	655812.3
Aerococcus viridans LL1	Aerococcus viridans	1175629.3
Alkalibacillus haloalkaliphilus C5	Alkalibacillus haloalkaliphilus	1193119.3
Alkaliphilus transvaalensis ATCC 700919	Alkaliphilus transvaalensis	1408422.3
Amphibacillus xylanus NBRC 15112	Amphibacillus xylanus	698758.3
Anaerococcus hydrogenalis ACS-025-V-Sch4	Anaerococcus hydrogenalis	879306.3
Anaerococcus hydrogenalis DSM 7454	Anaerococcus hydrogenalis	561177.4
Anaerococcus lactolyticus ATCC 51172	Anaerococcus lactolyticus	525254.4
Anaerococcus lactolyticus S7-1-13	Anaerococcus lactolyticus	1284686.3
Anaerococcus tetradius ATCC 35098	Anaerococcus tetradius	525255.3
Anaerococcus vaginalis ATCC 51170	Anaerococcus vaginalis	655811.4
Anaerostipes hadrus DSM 3319	Anaerostipes hadrus	649757.3
Anaerotruncus colihominis DSM 17241	Anaerotruncus colihominis	445972.6
Aneurinibacillus aneurinilyticus ATCC 12856	Aneurinibacillus aneurinilyticus	649747.3
Aneurinibacillus terranovensis DSM 18919	Aneurinibacillus terranovensis	1121002.4
Anoxybacillus ayderensis AB04	Anoxybacillus ayderensis	265546.4
Anoxybacillus flavithermus	Anoxybacillus flavithermus	33934.3
Anoxybacillus flavithermus AK1	Anoxybacillus flavithermus	1297581.3
Anoxybacillus flavithermus NBRC 109594	Anoxybacillus flavithermus	1315967.3
Anoxybacillus flavithermus TNO-09.006	Anoxybacillus flavithermus	1267580.3
Anoxybacillus flavithermus WK1	Anoxybacillus flavithermus	491915.6
Anoxybacillus gonensis G2	Anoxybacillus gonensis	198467.4
Anoxybacillus kamchatkensis G10	Anoxybacillus kamchatkensis	1212546.3
Anoxybacillus tepidamans PS2	Anoxybacillus tepidamans	1382358.3
Azotobacter chroococcum NCIMB 8003	Azotobacter chroococcum	1328314.4
Azotobacter vinelandii CA	Azotobacter vinelandii	1283330.3
Azotobacter vinelandii CA6	Azotobacter vinelandii	1283331.3
Azotobacter vinelandii DJ	Azotobacter vinelandii	322710.5
Bacillus alcalophilus ATCC 27647	Bacillus alcalophilus	1218173.3
Bacillus halodurans C-125	Bacillus halodurans	272558.8
Bacillus methanolicus MGA3	Bacillus methanolicus	796606.3
Bacillus mycoides ATCC 6462	Bacillus mycoides	1405.1
Bacillus mycoides DSM 2048	Bacillus mycoides	526997.3

Bacillus mycoides Rock1-4	Bacillus mycoides	526998.3
Bacillus mycoides Rock3-17	Bacillus mycoides	526999.3
Bacillus pseudofirmus OF4	Bacillus pseudofirmus	398511.4
Bacillus pseudomycoides DSM 12442	Bacillus pseudomycoides	527000.3
Bacillus selenitireducens MLS10	[Bacillus] selenitireducens	439292.5
Bacillus thermoamylovorans	Bacillus thermoamylovorans	35841.3
Bacillus weihenstephanensis	Bacillus weihenstephanensis	86662.6
Bacillus weihenstephanensis FSL H7-687	Bacillus weihenstephanensis	1227358.4
Bacillus weihenstephanensis FSL R5-860	Bacillus weihenstephanensis	1227359.4
Bacillus weihenstephanensis KBAB4	Bacillus weihenstephanensis	315730.11
Bacillus weihenstephanensis NBRC 101238 = DSM 11821	Bacillus weihenstephanensis	1220585.4
Blautia hansenii DSM 20583	Blautia hansenii	537007.6
Blautia producta ATCC 27340	Blautia producta	1121114.4
Blautia producta ER3	Blautia producta	33035.4
Blautia schinkii DSM 10518	Blautia schinkii	1410649.3
Brevibacillus agri 5-2	Brevibacillus agri	1444307.3
Brevibacillus agri BAB-2500	Brevibacillus agri	1246477.3
Brevibacillus borstelensis 3096-7	Brevibacillus borstelensis	1444309.3
Brevibacillus borstelensis AK1	Brevibacillus borstelensis	1300222.3
Brevibacillus borstelensis cifa_chp40	Brevibacillus borstelensis	1429889.3
Brevibacillus borstelensis LChuR05	Brevibacillus borstelensis	45462.4
Brevibacillus laterosporus	Brevibacillus laterosporus	1465.13
Brevibacillus laterosporus DSM 25	Brevibacillus laterosporus	1121121.3
Brevibacillus laterosporus GI-9	Brevibacillus laterosporus	1118154.3
Brevibacillus laterosporus LMG 15441	Brevibacillus laterosporus	1042163.3
Brevibacillus laterosporus PE36	Brevibacillus laterosporus	1399144.3
Brevibacillus thermoruber 423	Brevibacillus thermoruber	1346613.3
Brevibacillus thermoruber PM1	Brevibacillus thermoruber	1382302.3
Bulleidia extructa W1219	Bulleidia extructa	679192.3
Buttiauxella agrestis	Buttiauxella agrestis	82977.3
Buttiauxella agrestis ATCC 33320	Buttiauxella agrestis	1006004.4
Butyrivibrio fibrisolvens 16/4	Butyrivibrio fibrisolvens	657324.3
Butyrivibrio fibrisolvens AB2020	Butyrivibrio fibrisolvens	1280697.3
Butyrivibrio fibrisolvens FE2007	Butyrivibrio fibrisolvens	1280700.3
Butyrivibrio fibrisolvens MD2001	Butyrivibrio fibrisolvens	1280703.3
Butyrivibrio fibrisolvens ND3005	Butyrivibrio fibrisolvens	1280696.3
Butyrivibrio fibrisolvens WTE3004	Butyrivibrio fibrisolvens	1280699.3
Butyrivibrio fibrisolvens YRB2005	Butyrivibrio fibrisolvens	1280687.3
Caldibacillus debilis DSM 16016	Caldibacillus debilis	1121917.3
Cardiobacterium hominis ATCC 15826	Cardiobacterium hominis	638300.3
Carnobacterium maltaromaticum ATCC 35586	Carnobacterium maltaromaticum	1087479.3
Carnobacterium maltaromaticum LMA28	Carnobacterium maltaromaticum	1234679.3
Chromohalobacter israelensis 6768	Chromohalobacter israelensis	141390.3
Chromohalobacter salexigens DSM 3043	Chromohalobacter salexigens	290398.11
Citrobacter amalonaticus	Citrobacter amalonaticus	35703.8
Citrobacter braakii GTA-CB04	Citrobacter braakii	57706.1
Citrobacter farmeri GTC 1319	Citrobacter farmeri	1114922.3
Citrobacter freundii 4_7_47CFAA	Citrobacter freundii	742730.3
Citrobacter freundii ATCC 8090 = MTCC 1658	Citrobacter freundii	1006003.3
Citrobacter freundii CFNIH1	Citrobacter freundii	1333848.3

Citrobacter freundii GTC 09479	Citrobacter freundii	1288347.3
Citrobacter freundii GTC 09629	Citrobacter freundii	1297584.3
Citrobacter freundii MGH 56	Citrobacter freundii	1439318.3
Citrobacter freundii NBRC 12681	Citrobacter freundii	1114920.3
Citrobacter freundii RLS1	Citrobacter freundii	1454056.3
Citrobacter freundii str. ballerup 7851/39	Citrobacter freundii	670484.3
Citrobacter freundii UCI 31	Citrobacter freundii	1400136.3
Citrobacter freundii UCI 32	Citrobacter freundii	1400137.3
Citrobacter sedlakii NBRC 105722	Citrobacter sedlakii	1218086.3
Citrobacter werkmanii NBRC 105721	Citrobacter werkmanii	1114921.3
Citrobacter youngae ATCC 29220	Citrobacter youngae	500640.5
Clostridium acidurici 9a	Gottschalkia acidurici	1128398.3
Clostridium cellulosi CS-4-4	[Clostridium] cellulosi	1367212.3
Clostridium leptum DSM 753	[Clostridium] leptum	428125.8
Clostridium methylpentosum DSM 5476	[Clostridium] methylpentosum	537013.3
Clostridium orbiscindens 1_3_50AFAA	Flavonifractor plautii	742738.3
Clostridium sporosphaeroides DSM 1294	[Clostridium] sporosphaeroides	1121334.3
Clostridium sticklandii DSM 519	[Clostridium] sticklandii	499177.3
Clostridium thermocellum AD2	Ruminiclostridium thermocellum	1138384.3
Clostridium thermocellum ATCC 27405	Ruminiclostridium thermocellum	203119.11
Clostridium thermocellum BC1	Ruminiclostridium thermocellum	1349417.3
Clostridium thermocellum DSM 1313	Ruminiclostridium thermocellum	637887.3
Clostridium thermocellum DSM 2360	Ruminiclostridium thermocellum	572545.3
Clostridium thermocellum JW20	Ruminiclostridium thermocellum	492476.4
Clostridium thermocellum YS	Ruminiclostridium thermocellum	1094188.3
Colwellia psychrerythraea 34H	Colwellia psychrerythraea	167879.5
Colwellia psychrerythraea GAB14E	Colwellia psychrerythraea	28229.3
Desulfitobacterium dehalogenans ATCC 51507	Desulfitobacterium dehalogenans	756499.4
Desulfitobacterium hafniense DCB-2	Desulfitobacterium hafniense	272564.6
Desulfitobacterium hafniense DP7	Desulfitobacterium hafniense	537010.4
Desulfitobacterium hafniense PCP-1	Desulfitobacterium hafniense	1090321.3
Desulfitobacterium hafniense TCP-A	Desulfitobacterium hafniense	872024.4
Desulfitobacterium hafniense Y51	Desulfitobacterium hafniense	138119.41
Desulfitobacterium metallireducens DSM 15288	Desulfitobacterium metallireducens	871968.4
Dickeya chrysanthemi M074	Dickeya chrysanthemi	556.28
Dickeya chrysanthemi NCPPB 3533	Dickeya chrysanthemi	1224148.3
Dickeya chrysanthemi NCPPB 402	Dickeya chrysanthemi	1223569.3
Dickeya chrysanthemi NCPPB 516	Dickeya chrysanthemi	1223571.3
Dolosigranulum pigrum ATCC 51524	Dolosigranulum pigrum	883103.3
Dorea formicigenerans 4_6_53AFAA	Dorea formicigenerans	742765.5
Dorea formicigenerans ATCC 27755	Dorea formicigenerans	411461.4
Dorea longicatena AGR2136	Dorea longicatena	1280698.3
Dorea longicatena DSM 13814	Dorea longicatena	411462.6
Enterobacter cancerogenus ATCC 35316	Enterobacter cancerogenus	500639.8
Enterobacter cancerogenus M004	Enterobacter cancerogenus	69218.3
Enterobacter cloacae ATCC 13047	Enterobacter cloacae	550.124
Enterobacter cloacae BIDMC 66	Enterobacter cloacae	1439324.3
Enterobacter cloacae BIDMC 67	Enterobacter cloacae	1439325.3
Enterobacter cloacae BIDMC 8	Enterobacter cloacae	1329846.6
Enterobacter cloacae BWH 31	Enterobacter cloacae	1329845.3

Enterobacter cloacae BWH 43	Enterobacter cloacae	1439328.3
Enterobacter cloacae CHS 79	Enterobacter cloacae	1439326.3
Enterobacter cloacae EC_38VIM1	Enterobacter cloacae	1334630.3
Enterobacter cloacae ECNIH2	Enterobacter cloacae	1333850.3
Enterobacter cloacae ECNIH3	Enterobacter cloacae	1333851.3
Enterobacter cloacae ECR091	Enterobacter cloacae	1333849.3
Enterobacter cloacae EcWSU1	Enterobacter cloacae	1045856.3
Enterobacter cloacae IIT-BT 08	Enterobacter cloacae	1070842.3
Enterobacter cloacae ISC8	Enterobacter cloacae	1432556.3
Enterobacter cloacae JD6301	Enterobacter cloacae	1399774.3
Enterobacter cloacae JD8715	Enterobacter cloacae	1399775.3
Enterobacter cloacae MGH 53	Enterobacter cloacae	1439329.3
Enterobacter cloacae MGH 54	Enterobacter cloacae	1439330.3
Enterobacter cloacae MR2	Enterobacter cloacae	1312879.3
Enterobacter cloacae MRSN 11489	Enterobacter cloacae	1410032.3
Enterobacter cloacae P101	Enterobacter cloacae	1354030.3
Enterobacter cloacae S611	Enterobacter cloacae	1399146.3
Enterobacter cloacae str. Hanford	Enterobacter cloacae	1340854.3
Enterobacter cloacae subsp. cloacae 08XA1	Enterobacter cloacae	1203195.3
Enterobacter cloacae subsp. cloacae ATCC 13047	Enterobacter cloacae	716541.4
Enterobacter cloacae subsp. cloacae ENHKU01	Enterobacter cloacae	1211025.3
Enterobacter cloacae subsp. cloacae GS1	Enterobacter cloacae	1177927.3
Enterobacter cloacae subsp. cloacae NCTC 9394	Enterobacter cloacae	718254.4
Enterobacter cloacae subsp. cloacae SY-70	Enterobacter cloacae	1449089.4
Enterobacter cloacae subsp. dissolvens SDM	Enterobacter cloacae	1104326.3
Enterobacter cloacae UCI 23	Enterobacter cloacae	1400146.3
Enterobacter cloacae UCI 24	Enterobacter cloacae	1400147.3
Enterobacter cloacae UCI 29	Enterobacter cloacae	1400148.3
Enterobacter cloacae UCI 30	Enterobacter cloacae	1400149.3
Enterobacter cloacae UCI 35	Enterobacter cloacae	1400150.3
Enterobacter cloacae UCI 36	Enterobacter cloacae	1400151.3
Enterobacter cloacae UCI 39	Enterobacter cloacae	1400152.3
Enterobacter cloacae UCI 49	Enterobacter cloacae	1400154.3
Enterobacter cloacae UCI 50	Enterobacter cloacae	1400155.3
Enterobacter cloacae UCICRE 11	Enterobacter cloacae	1329855.3
Enterobacter cloacae UCICRE 12	Enterobacter cloacae	1329856.3
Enterobacter cloacae UCICRE 3	Enterobacter cloacae	1329852.3
Enterobacter cloacae UCICRE 5	Enterobacter cloacae	1329853.3
Enterobacter cloacae UCICRE 9	Enterobacter cloacae	1329854.3
Erwinia amylovora 01SFR-BO	Erwinia amylovora	1255306.3
Erwinia amylovora ACW56400	Erwinia amylovora	1027397.3
Erwinia amylovora ATCC 49946	Erwinia amylovora	716540.3
Erwinia amylovora CFBP 1232	Erwinia amylovora	1255307.3
Erwinia amylovora CFBP 2585	Erwinia amylovora	1255305.3
Erwinia amylovora CFBP1430	Erwinia amylovora	665029.3
Erwinia amylovora Ea266	Erwinia amylovora	1255304.3
Erwinia amylovora Ea356	Erwinia amylovora	1255303.3
Erwinia amylovora Ea644	Erwinia amylovora	1255309.3
Erwinia amylovora MR1	Erwinia amylovora	1255310.3
Erwinia amylovora NBRC 12687	Erwinia amylovora	1219359.3

Erwinia amylovora UPN527	Erwinia amylovora	1255308.3
Erwinia billingiae Eb661	Erwinia billingiae	634500.5
Erwinia mallotivora	Erwinia mallotivora	69222.5
Erwinia tracheiphila PSU-1	Erwinia tracheiphila	1044999.3
Erysipelothrix rhusiopathiae ATCC 19414	Erysipelothrix rhusiopathiae	525280.3
Erysipelothrix rhusiopathiae SY1027	Erysipelothrix rhusiopathiae	1313290.3
Escherichia hermannii NBRC 105704	Escherichia hermannii	1115512.3
Escherichia vulneris NBRC 102420	Escherichia vulneris	1115515.3
Eubacterium acidaminophilum DSM 3953	Eubacterium acidaminophilum	1286171.3
Eubacterium biforme DSM 3989	Holdemanella biformis	518637.5
Eubacterium brachy ATCC 33089	Eubacterium brachy	1321814.3
Eubacterium desmolans ATCC 43058	Eubacterium desmolans	1408437.3
Eubacterium dolichum DSM 3991	[Eubacterium] dolichum	428127.7
Eubacterium hallii DSM 3353	[Eubacterium] hallii	411469.3
Eubacterium infirmum F0142	[Eubacterium] infirmum	883109.3
Eubacterium nodatum ATCC 33099	[Eubacterium] nodatum	1161902.3
Eubacterium plexicaudatum ASF492	Eubacterium plexicaudatum	1235802.3
Eubacterium ramulus ATCC 29099	Eubacterium ramulus	1256908.3
Eubacterium saphenum ATCC 49989	Eubacterium saphenum	592031.3
Eubacterium siraeum 70/3	[Eubacterium] siraeum	657319.3
Eubacterium siraeum DSM 15702	[Eubacterium] siraeum	428128.7
Eubacterium siraeum V10Sc8a	[Eubacterium] siraeum	717961.3
Eubacterium sulci ATCC 35585	[Eubacterium] sulci	888727.3
Eubacterium ventriosum ATCC 27560	Eubacterium ventriosum	411463.4
Eubacterium xylanophilum ATCC 35991	Eubacterium xylanophilum	1336241.3
Exiguobacterium acetylicum DSM 20416	Exiguobacterium acetylicum	1397697.3
Exiguobacterium antarcticum B7	Exiguobacterium antarcticum	1087448.3
Exiguobacterium antarcticum DSM 14480	Exiguobacterium antarcticum	1397700.3
Exiguobacterium aurantiacum DSM 6208	Exiguobacterium aurantiacum	1397694.4
Facklamia hominis ACS-120-V-Sch10	Facklamia hominis	883110.3
Facklamia hominis CCUG 36813	Facklamia hominis	883111.3
Facklamia ignava CCUG 37419	Facklamia ignava	883112.3
Facklamia languida CCUG 37842	Facklamia languida	883113.3
Facklamia sourekii ATCC 700629	Facklamia sourekii	1408438.3
Faecalibacterium cf. prausnitzii KLE1255	Faecalibacterium prausnitzii	748224.3
Faecalibacterium prausnitzii A2-165	Faecalibacterium prausnitzii	411483.3
Faecalibacterium prausnitzii L2-6	Faecalibacterium prausnitzii	718252.3
Faecalibacterium prausnitzii M21/2	Faecalibacterium prausnitzii	411485.1
Faecalibacterium prausnitzii SL3/3	Faecalibacterium prausnitzii	657322.3
Flavonifractor plautii ATCC 29863	Flavonifractor plautii	411475.3
Gallibacterium anatis 10672-6	Gallibacterium anatis	1396515.3
Gallibacterium anatis 12656/12	Gallibacterium anatis	1195244.3
Gallibacterium anatis 23T10	Gallibacterium anatis	750.1
Gallibacterium anatis 4895	Gallibacterium anatis	1396510.3
Gallibacterium anatis 7990	Gallibacterium anatis	1396511.3
Gallibacterium anatis CCM5995	Gallibacterium anatis	1396513.3
Gallibacterium anatis DSM 16844 = F 149	Gallibacterium anatis	1121910.3
Gallibacterium anatis IPDH697-78	Gallibacterium anatis	1396514.3
Gallibacterium anatis str. Avicor	Gallibacterium anatis	1396512.3
Gallibacterium anatis UMN179	Gallibacterium anatis	1005058.3

Gemella bergeriae ATCC 700627	Gemella bergeri	1321820.3
Gemella cuniculi DSM 15828	Gemella cuniculi	1121914.3
Gemella haemolysans ATCC 10379	Gemella haemolysans	546270.5
Gemella haemolysans M341	Gemella haemolysans	562981.3
Gemella sanguinis ATCC 700632	Gemella sanguinis	1408440.3
Gemella sanguinis M325	Gemella sanguinis	562983.3
Geobacillus caldoxylosilyticus CIC9	Geobacillus caldoxylosilyticus	1234664.3
Geobacillus caldoxylosilyticus NBRC 107762	Geobacillus caldoxylosilyticus	1220594.3
Geobacillus kaustophilus GBlys	Geobacillus kaustophilus	1337888.4
Geobacillus kaustophilus HTA426	Geobacillus kaustophilus	235909.7
Geobacillus kaustophilus NBRC 102445	Geobacillus kaustophilus	1220595.3
Geobacillus stearothermophilus ATCC 7953	Geobacillus stearothermophilus	937593.4
Geobacillus subterraneus PSS2	Geobacillus subterraneus	1382357.3
Geobacillus thermocatenulatus GS-1	Geobacillus thermocatenulatus	1444308.3
Geobacillus thermodenitrificans DSM 465	Geobacillus thermodenitrificans	1413215.3
Geobacillus thermodenitrificans NG80-2	Geobacillus thermodenitrificans	420246.7
Geobacillus thermoglucosidans TNO-09.020	Geobacillus thermoglucosidasius	1136178.3
Geobacillus thermoglucosidasius C56-YS93	Geobacillus thermoglucosidasius	634956.3
Geobacillus thermoglucosidasius NBRC 107763	Geobacillus thermoglucosidasius	1223501.3
Geobacillus thermoleovorans B23	Geobacillus thermoleovorans	1406857.3
Geobacillus thermoleovorans CCB_US3_UF5	Geobacillus thermoleovorans	1111068.3
Geobacillus vulcani PSS1	Geobacillus vulcani	1382315.3
Gracilibacillus boraciitolerans JCM 21714	Gracilibacillus boraciitolerans	1298598.3
Granulicatella adiacens ATCC 49175	Granulicatella adiacens	638301.3
Granulicatella elegans ATCC 700633	Granulicatella elegans	626369.3
Halobacillus halophilus DSM 2266	Halobacillus halophilus	866895.3
Halorhodospira halochloris str. A	Halorhodospira halochloris	1354791.3
Halorhodospira halophila SL1	Halorhodospira halophila	349124.8
Halothermothrix orenii H 168	Halothermothrix orenii	373903.5
Halothiobacillus neapolitanus c2	Halothiobacillus neapolitanus	555778.5
Holdemania filiformis DSM 12042	Holdemania filiformis	545696.5
Hydrogenovibrio marinus	Hydrogenovibrio marinus	28885.3
Hydrogenovibrio marinus DSM 11271	Hydrogenovibrio marinus	1123513.3
Jeotgalibacillus campisalis SF-57	Jeotgalibacillus campisalis	220754.4
Jeotgalicoccus psychrophilus DSM 19085	Jeotgalicoccus psychrophilus	1122129.3
Kyrpidia tusciae DSM 2912	Kyrpidia tusciae	562970.4
Lachnoanaerobaculum saburreum DSM 3986	Lachnoanaerobaculum saburreum	887325.3
Lachnoanaerobaculum saburreum F0468	Lachnoanaerobaculum saburreum	1095750.3
Lelliottia amnigena CHS 78	Lelliottia amnigena	1439331.3
Leuconostoc argentinum KCTC 3773	Leuconostoc lactis	886872.3
Leuconostoc carnosum JB16	Leuconostoc carnosum	1229758.3
Leuconostoc carnosum KCTC 3525	Leuconostoc carnosum	1046593.3
Leuconostoc fallax KCTC 3537	Leuconostoc fallax	907931.3
Leuconostoc lactis	Leuconostoc lactis	1246.4
Leuconostoc lactis KCTC 3528 = DSM 20202	Leuconostoc lactis	935294.3
Leuconostoc pseudomesenteroides 1159	Leuconostoc pseudomesenteroides	1339246.3
Leuconostoc pseudomesenteroides 4882	Leuconostoc pseudomesenteroides	1154757.4
Leuconostoc pseudomesenteroides KCTC 3652	Leuconostoc pseudomesenteroides	935295.4
Leuconostoc pseudomesenteroides PS12	Leuconostoc pseudomesenteroides	1339247.3
Listeria grayi DSM 20601	Listeria grayi	525367.9

Listeria grayi FSL F6-1183	Listeria grayi	1265827.4
Listeria ivanovii FSL F6-596	Listeria ivanovii	702454.3
Listeria ivanovii subsp. ivanovii PAM 55	Listeria ivanovii	1638.4
Listeria ivanovii subsp. ivanovii WSLC 3010	Listeria ivanovii	202751.3
Listeria ivanovii subsp. londoniensis WSLC 30167	Listeria ivanovii	202752.6
Listeria ivanovii WSLC3009	Listeria ivanovii	1457190.3
Lonsdalea quercina subsp. quercina	Lonsdalea quercina	1082705.1
Luteimonas mephitis DSM 12574	Luteimonas mephitis	1122183.3
Lysinibacillus fusiformis	Lysinibacillus fusiformis	28031.4
Lysinibacillus fusiformis H1k	Lysinibacillus fusiformis	1416755.3
Lysinibacillus fusiformis ZB2	Lysinibacillus fusiformis	1231627.3
Lysinibacillus fusiformis ZC1	Lysinibacillus fusiformis	714961.3
Lysinibacillus odysseyi 34hs-1 = NBRC 100172	Lysinibacillus odysseyi	1220589.3
Lysobacter antibioticus	Lysobacter antibioticus	84531.4
Lysobacter antibioticus HS124	Lysobacter antibioticus	1286308.3
Macrococcus caseolyticus JCSC5402	Macrococcus caseolyticus	458233.11
Mannheimia granulomatis DSM 19156	Mannheimia granulomatis	1122190.3
Mannheimia haemolytica D153	Mannheimia haemolytica	1261126.6
Mannheimia haemolytica D171	Mannheimia haemolytica	1311759.4
Mannheimia haemolytica D174	Mannheimia haemolytica	1311760.4
Mannheimia haemolytica D193	Mannheimia haemolytica	1329904.3
Mannheimia haemolytica D35	Mannheimia haemolytica	1329905.3
Mannheimia haemolytica D38	Mannheimia haemolytica	1329906.3
Mannheimia haemolytica M42548	Mannheimia haemolytica	1316932.3
Mannheimia haemolytica MhBrain2012	Mannheimia haemolytica	1329902.3
Mannheimia haemolytica MhSwine2000	Mannheimia haemolytica	1329903.3
Mannheimia haemolytica PHL213	Mannheimia haemolytica	272629.3
Mannheimia haemolytica serotype 6 str. H23	Mannheimia haemolytica	1261125.3
Mannheimia haemolytica serotype A1/A6 str. PKL10	Mannheimia haemolytica	1450449.3
Mannheimia haemolytica serotype A2 str. BOVINE	Mannheimia haemolytica	669262.3
Mannheimia haemolytica serotype A2 str. OVINE	Mannheimia haemolytica	669261.3
Mannheimia haemolytica USDA-ARS-USMARC-183	Mannheimia haemolytica	1249531.3
Mannheimia haemolytica USDA-ARS-USMARC-185	Mannheimia haemolytica	1249526.3
Mannheimia haemolytica USMARC_2286	Mannheimia haemolytica	1366053.4
Mannheimia varigena USDA-ARS-USMARC-1261	Mannheimia varigena	1432056.3
Mannheimia varigena USDA-ARS-USMARC-1296	Mannheimia varigena	1433287.3
Mannheimia varigena USDA-ARS-USMARC-1312	Mannheimia varigena	1434214.3
Mannheimia varigena USDA-ARS-USMARC-1388	Mannheimia varigena	1434215.3
Marinococcus halotolerans DSM 16375	Marinococcus halotolerans	1122203.4
Marinomonas mediterranea MMB-1	Marinomonas mediterranea	717774.3
Megasphaera elsdenii 24-50	Megasphaera elsdenii	907.5
Megasphaera elsdenii DSM 20460	Megasphaera elsdenii	907.4
Megasphaera elsdenii T81	Megasphaera elsdenii	1410663.3
Methylobacter luteus IMV-B-3098	Methylobacter luteus	1095552.3
Methylobacter marinus A45	Methylobacter marinus	674036.3
Methylobacter whittenburyi	Methylobacter whittenburyi	39770.3
Methylococcus capsulatus str. Bath	Methylococcus capsulatus	243233.7
Methylococcus capsulatus str. Texas = ATCC 19069	Methylococcus capsulatus	1224744.3
Methylomicrobium agile	Methylomicrobium agile	39774.3
Methylomicrobium album BG8	Methylomicrobium album	686340.3

Methylomonas methanica MC09	Methylomonas methanica	857087.3
Mitsuokella jalaludinii DSM 13811	Mitsuokella jalaludinii	1410665.3
Mitsuokella multacida DSM 20544	Mitsuokella multacida	500635.8
Moorella thermoacetica ATCC 39073	Moorella thermoacetica	264732.11
Moorella thermoacetica Y72	Moorella thermoacetica	1325331.3
Oenococcus kitaharae DSM 17330	Oenococcus kitaharae	1045004.4
Oenococcus oeni ATCC BAA-1163	Oenococcus oeni	379360.3
Oenococcus oeni AWRIB202	Oenococcus oeni	1160703.3
Oenococcus oeni AWRIB304	Oenococcus oeni	1160702.3
Oenococcus oeni AWRIB318	Oenococcus oeni	1167631.3
Oenococcus oeni AWRIB418	Oenococcus oeni	1206769.3
Oenococcus oeni AWRIB419	Oenococcus oeni	1206770.3
Oenococcus oeni AWRIB422	Oenococcus oeni	1206771.3
Oenococcus oeni AWRIB429	Oenococcus oeni	655225.3
Oenococcus oeni AWRIB548	Oenococcus oeni	1206772.3
Oenococcus oeni AWRIB553	Oenococcus oeni	1206773.3
Oenococcus oeni AWRIB568	Oenococcus oeni	1206774.3
Oenococcus oeni AWRIB576	Oenococcus oeni	1206775.3
Oenococcus oeni DSM 20252 = AWRIB129	Oenococcus oeni	1122618.3
Oenococcus oeni PSU-1	Oenococcus oeni	203123.7
Oenococcus oeni X2L	Oenococcus oeni	1335618.3
Orenia marismortui DSM 5156	Orenia marismortui	926561.3
Paenibacillus panacisoli DSM 21345	Paenibacillus panacisoli	1122922.3
Paenibacillus pasadenensis DSM 19293	Paenibacillus pasadenensis	1122923.3
Paenibacillus peoriae KCTC 3763	Paenibacillus peoriae	1087481.3
Paenibacillus popilliae ATCC 14706	Paenibacillus popilliae	1212764.3
Paenibacillus sanguinis 2301083 = DSM 16941	Paenibacillus sanguinis	1122925.3
Paenibacillus stellifer DSM 14472	Paenibacillus stellifer	169760.4
Paenibacillus terrae HPL-003	Paenibacillus terrae	985665.3
Paenibacillus wynnii DSM 18334	Paenibacillus wynnii	268407.5
Parvimonas micra A293	Parvimonas micra	1408286.3
Parvimonas micra ATCC 33270	Parvimonas micra	411465.1
Parvimonas micra KCOM 1535; ChDC B708	Parvimonas micra	33033.4
Pasteurella dagmatis ATCC 43325	Pasteurella dagmatis	667128.3
Pectobacterium carotovorum M022	Pectobacterium carotovorum	554.6
Pectobacterium carotovorum subsp. brasiliense	Pectobacterium carotovorum	180957.1
Pectobacterium carotovorum subsp. brasiliensis PBR1692	Pectobacterium carotovorum	558269.5
Pectobacterium carotovorum subsp. carotovorum	Pectobacterium carotovorum	555.14
Pectobacterium carotovorum subsp. carotovorum PC1	Pectobacterium carotovorum	561230.3
Pectobacterium carotovorum subsp. carotovorum PCC21	Pectobacterium carotovorum	1218933.3
Pectobacterium carotovorum subsp. odoriferum	Pectobacterium carotovorum	78398.4
Pediococcus acidilactici 7_4	Pediococcus acidilactici	563194.3
Pediococcus acidilactici AGR20	Pediococcus acidilactici	1384067.3
Pediococcus acidilactici D3	Pediococcus acidilactici	1306952.3
Pediococcus acidilactici DSM 20284	Pediococcus acidilactici	862514.3
Pediococcus acidilactici MA18/5M	Pediococcus acidilactici	1080365.4
Pediococcus claussenii ATCC BAA-344	Pediococcus claussenii	701521.8
Peptoniphilus harei ACS-146-V-Sch2b	Peptoniphilus harei	908338.3
Peptoniphilus indolicus ATCC 29427	Peptoniphilus indolicus	997350.3
Peptoniphilus lacrimalis 315-B	Peptoniphilus lacrimalis	596330.3

Peptoniphilus lacrimalis DNF00528	Peptoniphilus lacrimalis	1401070.3
Peptoniphilus lacrimalis DSM 7455	Peptoniphilus lacrimalis	1122949.3
Peptostreptococcus anaerobius 653-L	Peptostreptococcus anaerobius	596329.3
Peptostreptococcus anaerobius VPI 4330	Peptostreptococcus anaerobius	1035196.3
Photobacterium angustum S14	Photobacterium angustum	314292.23
Photobacterium leiognathi lrivu.4.1	Photobacterium leiognathi	1248232.3
Photobacterium leiognathi subsp. mandapamensis svers.1.1.	Photobacterium leiognathi	1001530.3
Photobacterium phosphoreum ANT220	Photobacterium phosphoreum	1454202.3
Photobacterium profundum 3TCK	Photobacterium profundum	314280.5
Photobacterium profundum SS9	Photobacterium profundum	298386.8
Planococcus antarcticus DSM 14505	Planococcus antarcticus	1185653.3
Pluralibacter gergoviae FB2	Pluralibacter gergoviae	61647.1
Pontibacillus chungwhensis BH030062	Pontibacillus chungwhensis	1385513.3
Pontibacillus marinus BH030004 = DSM 16465	Pontibacillus marinus	1385511.3
Proteus penneri ATCC 35198	Proteus penneri	471881.3
Providencia alcalifaciens 205/92	Providencia alcalifaciens	1256988.3
Providencia alcalifaciens Ban1	Providencia alcalifaciens	663916.4
Providencia alcalifaciens Dmel2	Providencia alcalifaciens	1141661.3
Providencia alcalifaciens DSM 30120	Providencia alcalifaciens	520999.6
Providencia alcalifaciens F90-2004	Providencia alcalifaciens	1256987.3
Providencia alcalifaciens PAL-1	Providencia alcalifaciens	1256991.3
Providencia alcalifaciens PAL-2	Providencia alcalifaciens	1256992.3
Providencia alcalifaciens PAL-3	Providencia alcalifaciens	1256993.3
Providencia alcalifaciens R90-1475	Providencia alcalifaciens	1256989.3
Providencia alcalifaciens RIMD 1656011	Providencia alcalifaciens	1256990.3
Providencia rettgeri CCBH11880	Providencia rettgeri	587.17
Providencia rettgeri Dmel1	Providencia rettgeri	1141663.3
Providencia rettgeri DSM 1131	Providencia rettgeri	521000.6
Providencia rustigianii DSM 4541	Providencia rustigianii	500637.6
Pseudoalteromonas citrea	Pseudoalteromonas citrea	43655.3
Pseudoalteromonas citrea NCIMB 1889	Pseudoalteromonas citrea	1117314.3
Pseudoalteromonas luteoviolacea 2ta16	Pseudoalteromonas luteoviolacea	1353533.3
Pseudoalteromonas luteoviolacea B = ATCC 29581	Pseudoalteromonas luteoviolacea	1268239.3
Pseudoalteromonas luteoviolacea HI1	Pseudoalteromonas luteoviolacea	43657.9
Pseudoalteromonas piscicida ATCC 15057	Pseudoalteromonas piscicida	1279016.3
Pseudoalteromonas piscicida JCM 20779	Pseudoalteromonas piscicida	1117317.3
Pseudoalteromonas rubra ATCC 29570	Pseudoalteromonas rubra	1117318.14
Pseudobacteroides cellulosolvens ATCC 35603 = DSM 2933	Pseudobacteroides cellulosolvens	398512.4
Pseudobutyrivibrio ruminis AD2017	Pseudobutyrivibrio ruminis	1280694.3
Pseudobutyrivibrio ruminis CF1b	Pseudobutyrivibrio ruminis	1280688.3
Pseudobutyrivibrio ruminis HUN009	Pseudobutyrivibrio ruminis	1458469.3
Pseudoflavonifractor capillosus ATCC 29799	Pseudoflavonifractor capillosus	411467.6
Pseudomonas agarici NCPPB 2289	Pseudomonas agarici	690598.6
Pseudomonas alcaligenes MRY13-0052	Pseudomonas alcaligenes	1405803.3
Pseudomonas alcaligenes NBRC 14159	Pseudomonas alcaligenes	1215092.3
Pseudomonas alcaligenes OT 69	Pseudomonas alcaligenes	1333854.3
Pseudomonas cichorii JBC1	Pseudomonas cichorii	1441629.3
Pseudomonas corrugata CFBP 5454	Pseudomonas corrugata	1316927.4
Pseudomonas luteola XLDN4-9	Pseudomonas luteola	1207076.3
Pseudomonas oryzihabitans NBRC 102199	Pseudomonas oryzihabitans	1215113.3

Pseudomonas pseudoalcaligenes AD6	Pseudomonas pseudoalcaligenes	1453503.3
Pseudomonas pseudoalcaligenes CECT 5344	Pseudomonas pseudoalcaligenes	1182590.4
Pseudomonas pseudoalcaligenes KF707	Pseudomonas pseudoalcaligenes	1149133.6
Pseudomonas tolaasii 6264	Pseudomonas tolaasii	1161101.3
Pseudomonas tolaasii NCPPB 2192	Pseudomonas tolaasii	564423.7
Pseudomonas tolaasii PMS117	Pseudomonas tolaasii	1030145.6
Psychromonas arctica DSM 14288	Psychromonas arctica	1123036.3
Ruminococcus albus 7	Ruminococcus albus	697329.1
Ruminococcus albus 8	Ruminococcus albus	246199.4
Ruminococcus albus AD2013	Ruminococcus albus	1384065.3
Ruminococcus albus SY3	Ruminococcus albus	1341156.4
Ruminococcus bromii L2-63	Ruminococcus bromii	657321.5
Ruminococcus callidus ATCC 27760	Ruminococcus callidus	411473.3
Ruminococcus flavefaciens 007c	Ruminococcus flavefaciens	1341157.4
Ruminococcus flavefaciens 17	Ruminococcus flavefaciens	1030842.4
Ruminococcus flavefaciens AE3010	Ruminococcus flavefaciens	1384066.3
Ruminococcus flavefaciens ATCC 19208	Ruminococcus flavefaciens	1336236.3
Ruminococcus flavefaciens FD-1	Ruminococcus flavefaciens	641112.4
Ruminococcus flavefaciens MA2007	Ruminococcus flavefaciens	1410670.3
Ruminococcus flavefaciens MC2020	Ruminococcus flavefaciens	1410671.3
Ruminococcus flavefaciens ND2009	Ruminococcus flavefaciens	1410672.3
Ruminococcus gnavus AGR2154	[Ruminococcus] gnavus	1384063.4
Ruminococcus gnavus ATCC 29149	[Ruminococcus] gnavus	411470.6
Ruminococcus gnavus CC55_001C	[Ruminococcus] gnavus	1073375.3
Ruminococcus lactaris ATCC 29176	Ruminococcus lactaris	471875.6
Ruminococcus lactaris CC59_002D	Ruminococcus lactaris	1073376.3
Ruminococcus obeum A2-162	[Ruminococcus] obeum	657314.3
Ruminococcus obeum ATCC 29174	[Ruminococcus] obeum	411459.7
Ruminococcus torques ATCC 27756	[Ruminococcus] torques	411460.6
Ruminococcus torques L2-14	[Ruminococcus] torques	657313.3
Selenomonas artemidis DSM 19719	Selenomonas artemidis	1123249.3
Selenomonas artemidis F0399	Selenomonas artemidis	749551.3
Selenomonas flueggei ATCC 43531	Selenomonas flueggei	638302.3
Selenomonas infelix ATCC 43532	Selenomonas infelix	679201.3
Selenomonas noxia ATCC 43541	Selenomonas noxia	585503.3
Selenomonas noxia F0398	Selenomonas noxia	702437.3
Selenomonas ruminantium AB3002	Selenomonas ruminantium	1392502.3
Selenomonas ruminantium AC2024	Selenomonas ruminantium	1392501.3
Selenomonas ruminantium subsp. ruminantium ATCC 12561	Selenomonas ruminantium	1280706.4
Serratia fonticola AU-AP2C	Serratia fonticola	1332071.4
Serratia fonticola AU-P3(3)	Serratia fonticola	1332070.3
Serratia fonticola LMG 7882	Serratia fonticola	1378072.3
Serratia fonticola RB-25 [PRJNA232952]	Serratia fonticola	1441930.3
Serratia fonticola UTAD54	Serratia fonticola	1379259.4
Shewanella algae JCM 21037	Shewanella algae	1236544.3
Shewanella amazonensis SB2B	Shewanella amazonensis	326297.1
Shewanella baltica BA175	Shewanella baltica	693974.3
Shewanella baltica OS117	Shewanella baltica	693970.3
Shewanella baltica OS155	Shewanella baltica	325240.15
Shewanella baltica OS183	Shewanella baltica	693971.4

Shewanella baltica OS185	Shewanella baltica	402882.13
Shewanella baltica OS195	Shewanella baltica	399599.8
Shewanella baltica OS223	Shewanella baltica	407976.7
Shewanella baltica OS625	Shewanella baltica	693972.3
Shewanella baltica OS678	Shewanella baltica	693973.6
Shewanella colwelliana ATCC 39565	Shewanella colwelliana	1336240.3
Shewanella frigidimarina NCIMB 400	Shewanella frigidimarina	318167.14
Shewanella putrefaciens 200	Shewanella putrefaciens	399804.5
Shewanella putrefaciens CN-32	Shewanella putrefaciens	319224.16
Shewanella putrefaciens HRCR-6	Shewanella putrefaciens	1305841.3
Shewanella putrefaciens JCM 20190	Shewanella putrefaciens	1236543.3
Shewanella woodyi ATCC 51908	Shewanella woodyi	392500.6
Shimwellia blattae DSM 4481 = NBRC 105725	Shimwellia blattae	630626.3
Shuttleworthia satelles DSM 14600	Shuttleworthia satelles	626523.3
Solibacillus silvestris StLB046	Solibacillus silvestris	1002809.3
Solobacterium moorei DSM 22971	Solobacterium moorei	1123263.3
Solobacterium moorei F0204	Solobacterium moorei	706433.3
Sporolactobacillus inulinus CASD	Sporolactobacillus inulinus	1069536.3
Sporolactobacillus laevolacticus DSM 442	Sporolactobacillus laevolacticus	1395513.3
Sporolactobacillus terrae DSM 11697	Sporolactobacillus terrae	1444306.3
Sporolactobacillus terrae HKM-1	Sporolactobacillus terrae	1449983.3
Sporomusa ovata DSM 2662	Sporomusa ovata	1123288.3
Staphylococcus arlettae CVD059	Staphylococcus arlettae	1212545.3
Staphylococcus caprae C87	Staphylococcus capitis	435838.3
Staphylococcus chromogenes MU 970	Staphylococcus chromogenes	1220551.3
Staphylococcus delphini 8086	Staphylococcus delphini	1141105.7
Staphylococcus epidermidis M23864:W1	Staphylococcus caprae	525378.3
Staphylococcus hyicus ATCC 11249	Staphylococcus hyicus	1284.6
Staphylococcus intermedius NCTC 11048	Staphylococcus intermedius	1141106.7
Streptococcus anginosus 1_2_62CV	Streptococcus anginosus	742820.3
Streptococcus anginosus 1505	Streptococcus anginosus	1163301.3
Streptococcus anginosus C1051	Streptococcus anginosus	862970.3
Streptococcus anginosus C238	Streptococcus anginosus	862971.3
Streptococcus anginosus DORA_7	Streptococcus anginosus	1403946.3
Streptococcus anginosus F0211	Streptococcus anginosus	706437.3
Streptococcus anginosus SA1	Streptococcus anginosus	1328.12
Streptococcus anginosus SK1138	Streptococcus anginosus	1161422.3
Streptococcus anginosus SK52 = DSM 20563	Streptococcus anginosus	1000570.3
Streptococcus anginosus subsp. whileyi CCUG 39159	Streptococcus anginosus	1095729.3
Streptococcus anginosus subsp. whileyi MAS624	Streptococcus anginosus	1353243.3
Streptococcus anginosus T5	Streptococcus anginosus	1163302.3
Streptococcus bovis ATCC 700338	Streptococcus equinus	864569.5
Streptococcus bovis B315	Streptococcus equinus	1280690.3
Streptococcus bovis SN033	Streptococcus equinus	1280704.3
Streptococcus canis FSL Z3-227	Streptococcus canis	482234.3
Streptococcus criceti HS-6	Streptococcus criceti	873449.3
Streptococcus devriesei DSM 19639	Streptococcus devriesei	1123300.3
Streptococcus didelphis DSM 15616	Streptococcus didelphis	1123301.3
Streptococcus entericus DSM 14446	Streptococcus entericus	1123302.3
Streptococcus equinus	Streptococcus equinus	1335.4

Streptococcus equinus 2B	Streptococcus equinus	1410675.5
Streptococcus equinus ATCC 33317	Streptococcus equinus	1210006.5
Streptococcus equinus ATCC 9812	Streptococcus equinus	525379.3
Streptococcus equinus JB1	Streptococcus equinus	1294274.5
Streptococcus ferus DSM 20646	Streptococcus ferus	1123303.3
Streptococcus hyovaginalis DSM 12219	Streptococcus hyovaginalis	1123305.3
Streptococcus infantis ATCC 700779	Streptococcus infantis	889204.3
Streptococcus infantis SK1076	Streptococcus infantis	1005705.3
Streptococcus infantis SK1302	Streptococcus infantis	871237.3
Streptococcus infantis SK970	Streptococcus infantis	1035189.4
Streptococcus infantis X	Streptococcus infantis	997830.4
Streptococcus iniae	Streptococcus iniae	1346.13
Streptococcus iniae 9117	Streptococcus iniae	386894.6
Streptococcus iniae IUSA1	Streptococcus iniae	1273539.3
Streptococcus iniae KCTC 11634BP	Streptococcus iniae	1260129.3
Streptococcus iniae SF1	Streptococcus iniae	1318633.3
Streptococcus lutetiensis 033	Streptococcus lutetiensis	1076934.5
Streptococcus macacae NCTC 11558	Streptococcus macacae	764298.3
Streptococcus minor DSM 17118	Streptococcus minor	1123309.3
Streptococcus mutans 11A1	Streptococcus mutans	857155.3
Streptococcus mutans 11SSST2	Streptococcus mutans	857147.3
Streptococcus mutans 11VS1	Streptococcus mutans	857143.3
Streptococcus mutans 14D	Streptococcus mutans	857113.3
Streptococcus mutans 15JP3	Streptococcus mutans	857152.3
Streptococcus mutans 15VF2	Streptococcus mutans	857145.3
Streptococcus mutans 1ID3	Streptococcus mutans	857154.3
Streptococcus mutans 1SM1	Streptococcus mutans	857151.3
Streptococcus mutans 21	Streptococcus mutans	857112.3
Streptococcus mutans 24	Streptococcus mutans	857107.3
Streptococcus mutans 2ST1	Streptococcus mutans	857148.3
Streptococcus mutans 2VS1	Streptococcus mutans	857144.3
Streptococcus mutans 3SN1	Streptococcus mutans	857149.3
Streptococcus mutans 4SM1	Streptococcus mutans	857150.3
Streptococcus mutans 4VF1	Streptococcus mutans	857146.3
Streptococcus mutans 5DC8	Streptococcus mutans	1257037.3
Streptococcus mutans 5SM3	Streptococcus mutans	857142.3
Streptococcus mutans 66-2A	Streptococcus mutans	857111.3
Streptococcus mutans 8ID3	Streptococcus mutans	857153.3
Streptococcus mutans A19	Streptococcus mutans	857136.3
Streptococcus mutans A9	Streptococcus mutans	857139.3
Streptococcus mutans AC4446	Streptococcus mutans	1257040.3
Streptococcus mutans ATCC 25175	Streptococcus mutans	1257041.3
Streptococcus mutans B	Streptococcus mutans	857110.3
Streptococcus mutans B04Sm5	Streptococcus mutans	1225197.3
Streptococcus mutans B05Sm11	Streptococcus mutans	1225187.3
Streptococcus mutans B06Sm2	Streptococcus mutans	1225199.3
Streptococcus mutans B07Sm2	Streptococcus mutans	1225192.3
Streptococcus mutans B082SM-A	Streptococcus mutans	1225198.3
Streptococcus mutans B084SM-A	Streptococcus mutans	1225190.3
Streptococcus mutans B09Sm1	Streptococcus mutans	1225193.3

Streptococcus mutans B102SM-B	Streptococcus mutans	1225195.3
Streptococcus mutans B107SM-B	Streptococcus mutans	1225191.3
Streptococcus mutans B111SM-A	Streptococcus mutans	1225203.3
Streptococcus mutans B112SM-A	Streptococcus mutans	1225196.3
Streptococcus mutans B114SM-A	Streptococcus mutans	1225204.3
Streptococcus mutans B115SM-A	Streptococcus mutans	1225205.3
Streptococcus mutans B12Sm1	Streptococcus mutans	1225189.3
Streptococcus mutans B13Sm1	Streptococcus mutans	1225188.3
Streptococcus mutans B23Sm1	Streptococcus mutans	1225202.3
Streptococcus mutans B24Sm2	Streptococcus mutans	1225194.3
Streptococcus mutans B85SM-B	Streptococcus mutans	1225200.3
Streptococcus mutans B88SM-A	Streptococcus mutans	1225201.3
Streptococcus mutans DSM 20523	Streptococcus mutans	1123310.3
Streptococcus mutans G123	Streptococcus mutans	857134.3
Streptococcus mutans GS-5	Streptococcus mutans	1198676.3
Streptococcus mutans KK21	Streptococcus mutans	1257038.3
Streptococcus mutans KK23	Streptococcus mutans	1257039.3
Streptococcus mutans M21	Streptococcus mutans	857133.3
Streptococcus mutans M230	Streptococcus mutans	857100.3
Streptococcus mutans M2A	Streptococcus mutans	857126.3
Streptococcus mutans N29	Streptococcus mutans	857138.3
Streptococcus mutans N3209	Streptococcus mutans	857125.3
Streptococcus mutans N34	Streptococcus mutans	857131.3
Streptococcus mutans N66	Streptococcus mutans	857124.3
Streptococcus mutans NCTC 11060	Streptococcus mutans	1257042.3
Streptococcus mutans NFSM1	Streptococcus mutans	857130.3
Streptococcus mutans NFSM2	Streptococcus mutans	857141.3
Streptococcus mutans NLML1	Streptococcus mutans	857114.3
Streptococcus mutans NLML4	Streptococcus mutans	857129.3
Streptococcus mutans NLML5	Streptococcus mutans	857128.3
Streptococcus mutans NLML8	Streptococcus mutans	857115.3
Streptococcus mutans NLML9	Streptococcus mutans	857127.3
Streptococcus mutans NMT4863	Streptococcus mutans	857137.3
Streptococcus mutans NN2025	Streptococcus mutans	511691.3
Streptococcus mutans NV1996	Streptococcus mutans	857123.3
Streptococcus mutans NVAB	Streptococcus mutans	857140.4
Streptococcus mutans OMZ175	Streptococcus mutans	857099.3
Streptococcus mutans PKUSS-HG01	Streptococcus mutans	1403829.3
Streptococcus mutans PKUSS-LG01	Streptococcus mutans	1404260.3
Streptococcus mutans R221	Streptococcus mutans	857101.3
Streptococcus mutans S1B	Streptococcus mutans	857105.3
Streptococcus mutans SA38	Streptococcus mutans	857104.3
Streptococcus mutans SA41	Streptococcus mutans	857103.3
Streptococcus mutans SF1	Streptococcus mutans	857121.3
Streptococcus mutans SF12	Streptococcus mutans	857102.3
Streptococcus mutans SF14	Streptococcus mutans	857120.3
Streptococcus mutans SM1	Streptococcus mutans	857108.3
Streptococcus mutans SM4	Streptococcus mutans	857109.4
Streptococcus mutans SM6	Streptococcus mutans	857119.3
Streptococcus mutans ST1	Streptococcus mutans	857118.3

Streptococcus mutans ST6	Streptococcus mutans	857117.3
Streptococcus mutans str. B16 P Sm1	Streptococcus mutans	1225186.3
Streptococcus mutans T4	Streptococcus mutans	857132.3
Streptococcus mutans TCI-101	Streptococcus mutans	1074113.3
Streptococcus mutans TCI-109	Streptococcus mutans	1074114.3
Streptococcus mutans TCI-11	Streptococcus mutans	1074095.3
Streptococcus mutans TCI-110	Streptococcus mutans	1074115.3
Streptococcus mutans TCI-116	Streptococcus mutans	1074116.3
Streptococcus mutans TCI-120	Streptococcus mutans	1074118.3
Streptococcus mutans TCI-123	Streptococcus mutans	1074119.3
Streptococcus mutans TCI-125	Streptococcus mutans	1074120.3
Streptococcus mutans TCI-138	Streptococcus mutans	1074121.3
Streptococcus mutans TCI-143	Streptococcus mutans	1074122.3
Streptococcus mutans TCI-145	Streptococcus mutans	1074123.3
Streptococcus mutans TCI-146	Streptococcus mutans	1074124.3
Streptococcus mutans TCI-148	Streptococcus mutans	1074125.3
Streptococcus mutans TCI-149	Streptococcus mutans	1074126.3
Streptococcus mutans TCI-151	Streptococcus mutans	1074127.3
Streptococcus mutans TCI-152	Streptococcus mutans	1074128.3
Streptococcus mutans TCI-153	Streptococcus mutans	1074129.3
Streptococcus mutans TCI-154	Streptococcus mutans	1074130.3
Streptococcus mutans TCI-162	Streptococcus mutans	1074134.3
Streptococcus mutans TCI-163	Streptococcus mutans	1074135.3
Streptococcus mutans TCI-164	Streptococcus mutans	1074136.3
Streptococcus mutans TCI-169	Streptococcus mutans	1074137.3
Streptococcus mutans TCI-170	Streptococcus mutans	1074138.3
Streptococcus mutans TCI-173	Streptococcus mutans	1074140.3
Streptococcus mutans TCI-176	Streptococcus mutans	1074143.3
Streptococcus mutans TCI-177	Streptococcus mutans	1074144.3
Streptococcus mutans TCI-179	Streptococcus mutans	1074146.3
Streptococcus mutans TCI-187	Streptococcus mutans	1074148.3
Streptococcus mutans TCI-191	Streptococcus mutans	1074149.3
Streptococcus mutans TCI-196	Streptococcus mutans	1074151.3
Streptococcus mutans TCI-201	Streptococcus mutans	1074153.3
Streptococcus mutans TCI-202	Streptococcus mutans	1074154.3
Streptococcus mutans TCI-204	Streptococcus mutans	1074155.3
Streptococcus mutans TCI-210	Streptococcus mutans	1074156.3
Streptococcus mutans TCI-212	Streptococcus mutans	1074157.3
Streptococcus mutans TCI-218	Streptococcus mutans	1074159.3
Streptococcus mutans TCI-219	Streptococcus mutans	1074160.3
Streptococcus mutans TCI-220	Streptococcus mutans	1074161.3
Streptococcus mutans TCI-222	Streptococcus mutans	1074162.3
Streptococcus mutans TCI-223	Streptococcus mutans	1074163.3
Streptococcus mutans TCI-224	Streptococcus mutans	1074164.3
Streptococcus mutans TCI-227	Streptococcus mutans	1074165.3
Streptococcus mutans TCI-228	Streptococcus mutans	1074166.3
Streptococcus mutans TCI-234	Streptococcus mutans	1074167.3
Streptococcus mutans TCI-239	Streptococcus mutans	1074168.3
Streptococcus mutans TCI-242	Streptococcus mutans	1074169.3
Streptococcus mutans TCI-243	Streptococcus mutans	1074170.3

Streptococcus mutans TCI-244	Streptococcus mutans	1074171.3
Streptococcus mutans TCI-249	Streptococcus mutans	1074173.3
Streptococcus mutans TCI-256	Streptococcus mutans	1074175.3
Streptococcus mutans TCI-260	Streptococcus mutans	1074176.3
Streptococcus mutans TCI-264	Streptococcus mutans	1074177.3
Streptococcus mutans TCI-267	Streptococcus mutans	1074178.3
Streptococcus mutans TCI-268	Streptococcus mutans	1074179.3
Streptococcus mutans TCI-278	Streptococcus mutans	1074180.3
Streptococcus mutans TCI-279	Streptococcus mutans	1074181.3
Streptococcus mutans TCI-280	Streptococcus mutans	1074182.3
Streptococcus mutans TCI-289	Streptococcus mutans	1074183.3
Streptococcus mutans TCI-292	Streptococcus mutans	1074184.3
Streptococcus mutans TCI-294	Streptococcus mutans	1074185.3
Streptococcus mutans TCI-298	Streptococcus mutans	1074186.3
Streptococcus mutans TCI-30	Streptococcus mutans	1074190.3
Streptococcus mutans TCI-399	Streptococcus mutans	1074092.3
Streptococcus mutans TCI-400	Streptococcus mutans	1074093.3
Streptococcus mutans TCI-51	Streptococcus mutans	1074100.3
Streptococcus mutans TCI-62	Streptococcus mutans	1074101.3
Streptococcus mutans TCI-70	Streptococcus mutans	1074102.3
Streptococcus mutans TCI-75	Streptococcus mutans	1074104.3
Streptococcus mutans TCI-78	Streptococcus mutans	1074105.3
Streptococcus mutans TCI-82	Streptococcus mutans	1074106.3
Streptococcus mutans TCI-85	Streptococcus mutans	1074107.3
Streptococcus mutans TCI-86	Streptococcus mutans	1074108.3
Streptococcus mutans TCI-92	Streptococcus mutans	1074109.3
Streptococcus mutans TCI-96	Streptococcus mutans	1074111.3
Streptococcus mutans TCI-99	Streptococcus mutans	1074112.3
Streptococcus mutans U138	Streptococcus mutans	857135.3
Streptococcus mutans U2A	Streptococcus mutans	857116.3
Streptococcus mutans U2B	Streptococcus mutans	857106.3
Streptococcus mutans UA159	Streptococcus mutans	210007.7
Streptococcus mutans UA159-FR	Streptococcus mutans	1437447.3
Streptococcus mutans W6	Streptococcus mutans	857122.3
Streptococcus oligofermentans AS 1.3089	Streptococcus oligofermentans	1302863.3
Streptococcus orisratti DSM 15617	Streptococcus orisratti	1123311.3
Streptococcus ovis DSM 16829	Streptococcus ovis	1123312.3
Streptococcus parasanguinis ATCC 15912	Streptococcus parasanguinis	760570.3
Streptococcus parasanguinis ATCC 903	Streptococcus parasanguinis	888048.3
Streptococcus parasanguinis CC87K	Streptococcus parasanguinis	1073372.3
Streptococcus parasanguinis F0405	Streptococcus parasanguinis	905067.3
Streptococcus parasanguinis F0449	Streptococcus parasanguinis	1095733.3
Streptococcus parasanguinis FW213	Streptococcus parasanguinis	1114965.3
Streptococcus parasanguinis SK236	Streptococcus parasanguinis	1035185.3
Streptococcus parauberis	Streptococcus parauberis	1348.3
Streptococcus parauberis KCTC 11537	Streptococcus parauberis	936154.3
Streptococcus parauberis KCTC 11980BP	Streptococcus parauberis	1260132.3
Streptococcus parauberis KRS-02083	Streptococcus parauberis	1207545.3
Streptococcus parauberis KRS-02109	Streptococcus parauberis	1207544.3
Streptococcus parauberis NCFD 2020	Streptococcus parauberis	873447.3

Streptococcus phocae C-4	Streptococcus phocae	1000562.3
Streptococcus porcinus str. Jelinkova 176	Streptococcus porcinus	873448.3
Streptococcus ratti FA-1 = DSM 20564	Streptococcus ratti	699248.3
Streptococcus sanguinis ATCC 29667	Streptococcus sanguinis	997356.4
Streptococcus sanguinis CC94A	Streptococcus sanguinis	1073373.3
Streptococcus sanguinis SK1	Streptococcus sanguinis	888807.3
Streptococcus sanguinis SK1056	Streptococcus sanguinis	888820.3
Streptococcus sanguinis SK1057	Streptococcus sanguinis	888821.3
Streptococcus sanguinis SK1058	Streptococcus sanguinis	888822.3
Streptococcus sanguinis SK1059	Streptococcus sanguinis	888823.3
Streptococcus sanguinis SK1087	Streptococcus sanguinis	888824.3
Streptococcus sanguinis SK115	Streptococcus sanguinis	888810.3
Streptococcus sanguinis SK150	Streptococcus sanguinis	888811.3
Streptococcus sanguinis SK160	Streptococcus sanguinis	888812.3
Streptococcus sanguinis SK330	Streptococcus sanguinis	888813.3
Streptococcus sanguinis SK340	Streptococcus sanguinis	888814.4
Streptococcus sanguinis SK353	Streptococcus sanguinis	888815.3
Streptococcus sanguinis SK355	Streptococcus sanguinis	888816.3
Streptococcus sanguinis SK36	Streptococcus sanguinis	388919.9
Streptococcus sanguinis SK405	Streptococcus sanguinis	888817.3
Streptococcus sanguinis SK408	Streptococcus sanguinis	888818.3
Streptococcus sanguinis SK49	Streptococcus sanguinis	888808.3
Streptococcus sanguinis SK678	Streptococcus sanguinis	888819.3
Streptococcus sanguinis SK72	Streptococcus sanguinis	888809.3
Streptococcus sanguinis VMC66	Streptococcus sanguinis	888825.3
Streptococcus sinensis HKU4	Streptococcus sinensis	176090.4
Streptococcus sobrinus DSM 20742	Streptococcus sobrinus	1123317.3
Streptococcus sobrinus TCI-107	Streptococcus sobrinus	1074066.3
Streptococcus sobrinus TCI-118	Streptococcus sobrinus	1074117.3
Streptococcus sobrinus TCI-119	Streptococcus sobrinus	1074067.3
Streptococcus sobrinus TCI-121	Streptococcus sobrinus	1074068.3
Streptococcus sobrinus TCI-124	Streptococcus sobrinus	1074069.3
Streptococcus sobrinus TCI-13	Streptococcus sobrinus	1074053.3
Streptococcus sobrinus TCI-157	Streptococcus sobrinus	1074132.3
Streptococcus sobrinus TCI-16	Streptococcus sobrinus	1074054.3
Streptococcus sobrinus TCI-160	Streptococcus sobrinus	1074133.3
Streptococcus sobrinus TCI-172	Streptococcus sobrinus	1074139.3
Streptococcus sobrinus TCI-175	Streptococcus sobrinus	1074142.3
Streptococcus sobrinus TCI-194	Streptococcus sobrinus	1074150.3
Streptococcus sobrinus TCI-2	Streptococcus sobrinus	1074094.3
Streptococcus sobrinus TCI-200	Streptococcus sobrinus	1074152.3
Streptococcus sobrinus TCI-215	Streptococcus sobrinus	1074158.3
Streptococcus sobrinus TCI-28	Streptococcus sobrinus	1074055.3
Streptococcus sobrinus TCI-336	Streptococcus sobrinus	1074189.3
Streptococcus sobrinus TCI-342	Streptococcus sobrinus	1074070.3
Streptococcus sobrinus TCI-345	Streptococcus sobrinus	1074071.4
Streptococcus sobrinus TCI-348	Streptococcus sobrinus	1074072.3
Streptococcus sobrinus TCI-349	Streptococcus sobrinus	1074073.3
Streptococcus sobrinus TCI-352	Streptococcus sobrinus	1074074.3
Streptococcus sobrinus TCI-355	Streptococcus sobrinus	1074076.3

Streptococcus sobrinus TCI-357	Streptococcus sobrinus	1074077.3
Streptococcus sobrinus TCI-363	Streptococcus sobrinus	1074078.3
Streptococcus sobrinus TCI-366	Streptococcus sobrinus	1074079.3
Streptococcus sobrinus TCI-367	Streptococcus sobrinus	1074080.3
Streptococcus sobrinus TCI-373	Streptococcus sobrinus	1074081.3
Streptococcus sobrinus TCI-374	Streptococcus sobrinus	1074082.3
Streptococcus sobrinus TCI-376	Streptococcus sobrinus	1074083.3
Streptococcus sobrinus TCI-377	Streptococcus sobrinus	1074084.3
Streptococcus sobrinus TCI-381	Streptococcus sobrinus	1074085.3
Streptococcus sobrinus TCI-384	Streptococcus sobrinus	1074086.3
Streptococcus sobrinus TCI-392	Streptococcus sobrinus	1074088.3
Streptococcus sobrinus TCI-395	Streptococcus sobrinus	1074089.3
Streptococcus sobrinus TCI-396	Streptococcus sobrinus	1074090.3
Streptococcus sobrinus TCI-50	Streptococcus sobrinus	1074056.3
Streptococcus sobrinus TCI-53	Streptococcus sobrinus	1074057.3
Streptococcus sobrinus TCI-54	Streptococcus sobrinus	1074058.3
Streptococcus sobrinus TCI-56	Streptococcus sobrinus	1074059.3
Streptococcus sobrinus TCI-61	Streptococcus sobrinus	1074060.3
Streptococcus sobrinus TCI-77	Streptococcus sobrinus	1074061.3
Streptococcus sobrinus TCI-79	Streptococcus sobrinus	1074062.3
Streptococcus sobrinus TCI-80	Streptococcus sobrinus	1074063.3
Streptococcus sobrinus TCI-89	Streptococcus sobrinus	1074064.3
Streptococcus sobrinus TCI-9	Streptococcus sobrinus	1074052.3
Streptococcus sobrinus TCI-98	Streptococcus sobrinus	1074065.3
Streptococcus sobrinus W1703	Streptococcus sobrinus	1227275.3
Streptococcus suis 05HAS68	Streptococcus suis	672190.3
Streptococcus suis 05ZYH33	Streptococcus suis	391295.8
Streptococcus suis 07SC3	Streptococcus suis	1214149.3
Streptococcus suis 10581	Streptococcus suis	1214158.3
Streptococcus suis 11538	Streptococcus suis	1214180.3
Streptococcus suis 11611	Streptococcus suis	1214148.3
Streptococcus suis 12814	Streptococcus suis	1214156.3
Streptococcus suis 13730	Streptococcus suis	1214159.3
Streptococcus suis 14636	Streptococcus suis	1214183.3
Streptococcus suis 14A	Streptococcus suis	1214167.3
Streptococcus suis 161_00P5	Streptococcus suis	1214150.3
Streptococcus suis 22083	Streptococcus suis	1214184.3
Streptococcus suis 2524	Streptococcus suis	1214181.3
Streptococcus suis 2651	Streptococcus suis	1214154.3
Streptococcus suis 2726	Streptococcus suis	1214161.3
Streptococcus suis 4417	Streptococcus suis	1214155.3
Streptococcus suis 4961	Streptococcus suis	1214176.3
Streptococcus suis 6407	Streptococcus suis	1214179.3
Streptococcus suis 8074	Streptococcus suis	1214182.3
Streptococcus suis 86-5192	Streptococcus suis	1214166.3
Streptococcus suis 8830	Streptococcus suis	1214157.3
Streptococcus suis 89/1591	Streptococcus suis	286604.5
Streptococcus suis 89-1591	Streptococcus suis	1214151.3
Streptococcus suis 89-2479	Streptococcus suis	1214169.3
Streptococcus suis 89-3576-3	Streptococcus suis	1214171.3

Streptococcus suis 89-4109-1	Streptococcus suis	1214172.3
Streptococcus suis 89-5259	Streptococcus suis	1214173.3
Streptococcus suis 92-1191	Streptococcus suis	1214175.3
Streptococcus suis 92-1400	Streptococcus suis	1214177.3
Streptococcus suis 92-4172	Streptococcus suis	1214178.3
Streptococcus suis 93A	Streptococcus suis	1214162.3
Streptococcus suis 98HAH33	Streptococcus suis	391296.8
Streptococcus suis A7	Streptococcus suis	993512.3
Streptococcus suis BM407	Streptococcus suis	568814.3
Streptococcus suis D12	Streptococcus suis	1004952.3
Streptococcus suis D9	Streptococcus suis	1005042.3
Streptococcus suis EA1832.92	Streptococcus suis	1321372.3
Streptococcus suis GZ1	Streptococcus suis	423211.3
Streptococcus suis JS14	Streptococcus suis	945704.3
Streptococcus suis NT77	Streptococcus suis	1214163.3
Streptococcus suis P1/7	Streptococcus suis	218494.6
Streptococcus suis R61	Streptococcus suis	996306.3
Streptococcus suis R735	Streptococcus suis	1214165.3
Streptococcus suis RC1	Streptococcus suis	1214152.3
Streptococcus suis S15	Streptococcus suis	1214160.3
Streptococcus suis S19	Streptococcus suis	1214164.3
Streptococcus suis S22	Streptococcus suis	1214168.3
Streptococcus suis S24	Streptococcus suis	1214170.3
Streptococcus suis S28	Streptococcus suis	1214174.3
Streptococcus suis S428	Streptococcus suis	1214153.3
Streptococcus suis S735	Streptococcus suis	1184252.3
Streptococcus suis SC070731	Streptococcus suis	1246365.4
Streptococcus suis SC84	Streptococcus suis	568813.3
Streptococcus suis SS12	Streptococcus suis	1005041.3
Streptococcus suis ST1	Streptococcus suis	1004951.3
Streptococcus suis ST3	Streptococcus suis	1007064.3
Streptococcus suis T15	Streptococcus suis	1340847.3
Streptococcus suis TL13	Streptococcus suis	1276647.3
Streptococcus suis YB51	Streptococcus suis	1380773.3
Streptococcus suis YS1	Streptococcus suis	1214185.3
Streptococcus suis YS10-2	Streptococcus suis	1214186.3
Streptococcus suis YS12	Streptococcus suis	1214187.3
Streptococcus suis YS14	Streptococcus suis	1214188.3
Streptococcus suis YS16	Streptococcus suis	1214189.3
Streptococcus suis YS17-2	Streptococcus suis	1214190.3
Streptococcus suis YS19-3	Streptococcus suis	1214191.3
Streptococcus suis YS21	Streptococcus suis	1214192.3
Streptococcus suis YS23-2	Streptococcus suis	1214193.3
Streptococcus suis YS24	Streptococcus suis	1214194.3
Streptococcus suis YS27-2	Streptococcus suis	1214195.3
Streptococcus suis YS3	Streptococcus suis	1214196.3
Streptococcus suis YS31	Streptococcus suis	1214197.3
Streptococcus suis YS34	Streptococcus suis	1214198.3
Streptococcus suis YS35	Streptococcus suis	1214199.3
Streptococcus suis YS39	Streptococcus suis	1214200.3

Streptococcus suis YS4	Streptococcus suis	1214201.3
Streptococcus suis YS43	Streptococcus suis	1214202.3
Streptococcus suis YS44	Streptococcus suis	1214203.3
Streptococcus suis YS46	Streptococcus suis	1214204.3
Streptococcus suis YS49	Streptococcus suis	1214205.3
Streptococcus suis YS50	Streptococcus suis	1214206.3
Streptococcus suis YS53	Streptococcus suis	1214207.3
Streptococcus suis YS54-2	Streptococcus suis	1214208.3
Streptococcus suis YS56	Streptococcus suis	1214209.3
Streptococcus suis YS57	Streptococcus suis	1214210.3
Streptococcus suis YS59	Streptococcus suis	1214211.3
Streptococcus suis YS6	Streptococcus suis	1214212.3
Streptococcus suis YS64	Streptococcus suis	1214213.3
Streptococcus suis YS66	Streptococcus suis	1214214.3
Streptococcus suis YS67	Streptococcus suis	1214215.3
Streptococcus suis YS7	Streptococcus suis	1214216.3
Streptococcus suis YS72	Streptococcus suis	1214217.3
Streptococcus suis YS74	Streptococcus suis	1214218.3
Streptococcus suis YS77	Streptococcus suis	1214219.3
Streptococcus thermophilus ASCC 1275	Streptococcus thermophilus	1408178.4
Streptococcus thermophilus CNCM I-1630	Streptococcus thermophilus	1042404.3
Streptococcus thermophilus CNRZ1066	Streptococcus thermophilus	299768.6
Streptococcus thermophilus DGCC7710	Streptococcus thermophilus	1268061.3
Streptococcus thermophilus JIM 8232	Streptococcus thermophilus	1051074.3
Streptococcus thermophilus LMD-9	Streptococcus thermophilus	322159.8
Streptococcus thermophilus LMG 18311	Streptococcus thermophilus	264199.4
Streptococcus thermophilus MN-ZLW-002	Streptococcus thermophilus	1187956.3
Streptococcus thermophilus MTCC 5460	Streptococcus thermophilus	1073569.3
Streptococcus thermophilus MTCC 5461	Streptococcus thermophilus	1073570.5
Streptococcus thermophilus ND03	Streptococcus thermophilus	767463.3
Streptococcus thoraltensis DSM 12221	Streptococcus thoraltensis	1123318.3
Streptococcus urinalis 2285-97	Streptococcus urinalis	764291.3
Streptococcus urinalis FB127-CNA-2	Streptococcus urinalis	883168.3
Syntrophothermus lipocalidus DSM 12680	Syntrophothermus lipocalidus	643648.3
Tetragenococcus muriaticus 3MR10-3	Tetragenococcus muriaticus	1302648.3
Tetragenococcus muriaticus DSM 15685	Tetragenococcus muriaticus	1123359.3
Tetragenococcus muriaticus PMC-11-5	Tetragenococcus muriaticus	1302649.3
Thiomicrospira chilensis DSM 12352	Thiomicrospira chilensis	1123515.3
Thiomicrospira crunogena XCL-2	Thiomicrospira crunogena	317025.9
Thiomicrospira kuenenii DSM 12350	Thiomicrospira kuenenii	1168067.3
Thiomicrospira pelophila DSM 1534	Thiomicrospira pelophila	1123517.3
Virgibacillus halodenitrificans	Virgibacillus halodenitrificans	1482.4
Virgibacillus halodenitrificans 1806	Virgibacillus halodenitrificans	1196028.3
Weissella cibaria KACC 11862	Weissella cibaria	911104.3
Weissella confusa LBAE C39-2	Weissella confusa	1127131.3
Weissella halotolerans DSM 20190	Weissella halotolerans	1123500.3
Weissella hellenica	Weissella hellenica	46256.5
Weissella koreensis KACC 15510	Weissella koreensis	1045854.4
Weissella koreensis KCTC 3621	Weissella koreensis	1123721.3
Weissella paramesenteroides ATCC 33313	Weissella paramesenteroides	585506.3

Xanthomonas albilineans	Xanthomonas albilineans	29447.3
Xanthomonas campestris JX	Xanthomonas campestris	1182783.3
Xanthomonas campestris LMCP11	Xanthomonas campestris	339.49
Xanthomonas campestris pv. arecae NCPPB 2649	Xanthomonas campestris	487849.3
Xanthomonas campestris pv. campestris str. 8004	Xanthomonas campestris	314565.5
Xanthomonas campestris pv. campestris str. ATCC 33913	Xanthomonas campestris	190485.4
Xanthomonas campestris pv. campestris str. B100	Xanthomonas campestris	509169.4
Xanthomonas campestris pv. campestris str. Xca5	Xanthomonas campestris	1211707.3
Xanthomonas campestris pv. cannabis NCPPB 2877	Xanthomonas campestris	92824.15
Xanthomonas campestris pv. musacearum 'Kenyan'	Xanthomonas campestris	1075759.3
Xanthomonas campestris pv. musacearum NCPPB 2005	Xanthomonas campestris	1094183.3
Xanthomonas campestris pv. musacearum NCPPB 4379	Xanthomonas campestris	1094184.3
Xanthomonas campestris pv. musacearum NCPPB 4380	Xanthomonas campestris	1094185.3
Xanthomonas campestris pv. musacearum NCPPB 4381	Xanthomonas campestris	559737.3
Xanthomonas campestris pv. musacearum NCPPB 4384	Xanthomonas campestris	1094186.4
Xanthomonas campestris pv. musacearum NCPPB 4392	Xanthomonas campestris	1184263.3
Xanthomonas campestris pv. musacearum NCPPB 4394	Xanthomonas campestris	1094187.3
Xanthomonas campestris pv. raphani 756C	Xanthomonas campestris	990315.4
Xanthomonas campestris pv. viticola LMG 965	Xanthomonas campestris	487899.3
Xanthomonas fragariae LMG 25863	Xanthomonas fragariae	1131451.6
Xanthomonas oryzae ATCC 35933	Xanthomonas oryzae	1313303.3
Xanthomonas oryzae pv. oryzae KACC 10331	Xanthomonas oryzae	291331.8
Xanthomonas oryzae pv. oryzae MAFF 311018	Xanthomonas oryzae	342109.8
Xanthomonas oryzae pv. oryzae NAI8	Xanthomonas oryzae	1423889.3
Xanthomonas oryzae pv. oryzae PXO99A	Xanthomonas oryzae	360094.4
Xanthomonas oryzae pv. oryzicola BLS256	Xanthomonas oryzae	383407.3
Xanthomonas oryzae pv. oryzicola MAI10	Xanthomonas oryzae	1423890.3
Xanthomonas oryzae X11-5A	Xanthomonas oryzae	1009853.4
Xanthomonas oryzae X8-1A	Xanthomonas oryzae	1009854.4
Xenorhabdus bovienii SS-2004	Xenorhabdus bovienii	406818.4
Xenorhabdus nematophila C2-3	Xenorhabdus nematophila	628.3
Xenorhabdus nematophila F1	Xenorhabdus nematophila	1306162.3
Xylella fastidiosa 32	Xylella fastidiosa	1214121.5
Xylella fastidiosa 6c	Xylella fastidiosa	1211847.5
Xylella fastidiosa 9a5c	Xylella fastidiosa	160492.11
Xylella fastidiosa ATCC 35879	Xylella fastidiosa	2371.35
Xylella fastidiosa Dixon	Xylella fastidiosa	155919.4
Xylella fastidiosa EB92.1	Xylella fastidiosa	945689.3
Xylella fastidiosa M12	Xylella fastidiosa	405440.5
Xylella fastidiosa M23	Xylella fastidiosa	405441.5
Xylella fastidiosa MUL0034	Xylella fastidiosa	1401256.4
Xylella fastidiosa Mul-MD	Xylella fastidiosa	1403344.3
Xylella fastidiosa PLS229	Xylella fastidiosa	1444770.3
Xylella fastidiosa subsp. fastidiosa GB514	Xylella fastidiosa	788929.3
Xylella fastidiosa subsp. multiplex ATCC 35871	Xylella fastidiosa	1267006.3
Xylella fastidiosa subsp. multiplex str. Red Oak 1	Xylella fastidiosa	1343737.3
Xylella fastidiosa subsp. sandyi Ann-1	Xylella fastidiosa	155920.4
Xylella fastidiosa Temecula1	Xylella fastidiosa	183190.5

(a) Strain designation

(b) Species name

(c) NCBI taxonomy strain id