Computational Methods for Omics Sequence Data with Focus on Non-Model Organisms

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Mathias Kuhring

Berlin
September 2017

Betreuer:	PD Dr. Bernhard Y. Renard
Erstgutachter:	PD Dr. Bernhard Y. Renard
Zweitgutachter:	Dr. Thomas D. Otto
Tag der Disputation:	13.02.2019

Abstract

Sequence data are the backbone for many biological research areas including but not limited to genomics, proteomics as well as proteogenomics. Sequence acquisition is facilitated by a wide selection of advanced technologies such as Next Generation Sequencing and Mass Spectrometry. These high-throughput methods produce substantial volumes of data with decreasing financial and time-based expenditures. These volumes of data render manual processing impossible and therefore require state-of-the-art computational methods for adequate analysis and interpretation. In proteogenomics the potential of combining omics methods to improve on sequence quality and availability is frequently emphasized, in particular for non-model organisms. In this thesis, we highlight and address several challenges in the "life cycle" of omics sequence data, from genome sequence acquisition through integrated evaluation to extensive utilization of comprehensive sequence collections.

We describe several methods with applications in different omics areas and emphasize means of potential integrative analysis. First, we introduce a method for *de novo* assembly contig quality ranking based on machine learning. Thereby, we demonstrate special potential for the application on metagenomic sequence data which usually feature a variety of previously sequenced as well as unsequenced, non-model organisms. Next, we elaborate on sequence availability of target sequences in databases considered for taxonomic classification of tandem MS spectra. Thereby, the effect of different sequence sources as well as different search strategies on taxonomic depth is taken in account. Finally, we introduce a novel approach for extensive taxonomic classification by iteratively processing recent and comprehensive protein sequence databases. We discuss diverse possibilities as well as the limits of our methods with respect to current public data basis. Thereby, we illustrate potential benefits of the presented methods for non-model organisms.

Acknowledgements

First of all, I want to thank Bernhard Renard for the opportunity to write my thesis under his supervision, for his profound support under every circumstances and his persistent patience with me.

Furthermore, I want to thank Thomas Otto for reviewing my thesis.

I want to thank my co-authors Piotr Wojtek Dabrowski, Vitor Piro, Andreas Nitsche, Thilo Muth and Bernhard Renard for their excellent contributions and collaboration.

In addition, I want to thank all my colleagues at the RKI for the constructive as well as entertaining time we spend together, on and off work.

Special thanks go to my girlfriend Bine for her continuous and loving support.

Contents

1. Introduction

1.1. Next Generation Sequencing Application in Genomics

High-throughput omics methods are the modern work horses in biological and medical science including but not limited to next generation sequencing (NGS) application in genomics as well as mass spectrometry (MS) application in proteomics.

Modern nucleotide sequencing technologies referred to as next generation sequencing enable analysis of genomes and transcriptomes in a cost-effective, fast and high-throughput manner. After the first generation sequencing driven by Sanger, second generation sequencing was defined by massive parallelization of short-reads followed by recent advancements in the third generation towards elongated and direct DNA sequencing referred to as single-molecule real-time sequencing (Goodwin et al., 2016; Heather and Chain, 2016). Sequencing methods are implemented in a plethora of instruments including but not limited to Illumina, Roche/454 and Ion Torrent as second generation, as well as Oxford Nanopore and Pacific BioSciences as third generation platforms, each featuring their own advantages and disadvantages (Glenn, 2011; Quail et al., 2012). While third generation sequencing improves upon previous technology, second generation sequencer such as the Illumina HiSeq platforms remain popular and prevalent in modern labs due to cost effectiveness as well as low error rates with established error profiles (Glenn, 2011; Goodwin et al., 2016).

Computational methods for genome or transcriptome reconstruction by use of NGS reads can be roughly classified either as alignment (often referred to as mapping) or *de novo* assembly procedures (Flicek and Birney, 2009; Horner et al., 2010). Alignment procedures make use of an corresponding reference genome or transcriptome and "map" the reads to likely positions of origin using for instance initial hash- or index-based heuristics followed by exact alignment verification (Reinert et al., 2015). In contrast, *de novo* assembly procedures allow sequence reconstruction completely without using a reference template by joining reads to contiguous sequences based on for instance overlap or de Bruijn graphs (Nagarajan and Pop, 2013; Sohn and Nam, 2016). In comparison to mapping, *de novo* assembly is often able to recover more individual or sample specific features such as genetic variations (Chaisson et al., 2015) and is an essential tool for sequencing genomes without close reference such as non-model organisms (Henson et al., 2012; Ekblom and Wolf, 2014).

Apart from whole genome sequencing (WGS), NGS supports several diverse ap-

plications such as transcriptome sequencing (RNA-Seq) and differential expression analysis (Wang et al., 2009; Ozsolak and Milos, 2011; Hrdlickova et al., 2017), (meta)genome-wide association studies (Luo et al., 2011; Chaitankar et al., 2016; Wang and Jia, 2016) as well as whole genome and 16S profiling and quantification of microbial communities (Chen and Pachter, 2005). Furthermore, genome sequencing provides the basis for major protein resources such as UniProtKB (The UniProt Consortium, 2017) or NCBI RefSeq (O'Leary et al., 2016) which apply automated annotation and manual curation procedures to populate their databases, thereby enabling a wide variety of proteomic studies.

1.2. Mass Spectrometry-based Proteomics

Mass spectrometry is the current *de facto* standard for protein identification since it is superior when it comes to providing high-throughput in combination with high accuracy (Käll and Vitek, 2011). Mass spectrometry technology and workflows vary widely and are commonly characterized by the type of ion source including electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) as well as mass analyser used such as ion traps, time-of-flight (TOF) or Fourier transform ion cyclotron (Aebersold and Mann, 2003).

Typical combinations include the MALDI-TOF for peptide mass fingerprinting (PMF) and ESI with two or more TOF sections or an ion trap for tandem MS peptide sequencing (Domon and Aebersold, 2006; McHugh and Arthur, 2008). For PMF, the peptide mass profile (fingerprint) of purified and digested proteins is matched and identified against spectral or protein databases. In addition, tandem MS enables to infer the full amino acids sequences of peptides. Here, proteins are digested (e.g. with trypsin), separated by liquid chromatography and then subject to MS analysis (LC-MS/MS). In addition to peptide mass acquisition, a few peptides per scan are selected by intensity, are randomly fragmented by for instance collision-induced dissociation and analysed in a second individual scan. The resulting peak pattern is used to infer peptide sequences, either by matching artificial spectra in a database search or by *de novo* sequencing (Nesvizhskii et al., 2007; McHugh and Arthur, 2008). This workflow is also referred to as shotgun bottom-up proteomics and is most frequently used for proteomic studies (Domon and Aebersold, 2006; Käll and Vitek, 2011).

Mass spectromic-based proteomics support a wide range of applications including but not limited to analysis of peptide sequences, proteome profiles, interactions and modifications (Aebersold and Mann, 2003), analysis and profiling of microbial mixtures and environmental samples (referred to as metaproteomics) (Hettich et al., 2013; Muth et al., 2016) as well as (differential) proteome quantification using label-free (e.g. tandem spectra counting and peak feature integration) or isotope labelling

methods (e.g. SILAC or iTraq) (Nesvizhskii et al., 2007; Käll and Vitek, 2011).

1.3. Integrative Applications of Omics Data

The field of proteogenomics takes advantage of different omics methods such as NGS and MS by integrative data acquisition, analyses and interpretation. Thereby, it combines the strength of different fields and technologies while simultaneously enabling a mutual gain in knowledge (Renuse et al., 2011; Armengaud et al., 2013). Originally, proteogenomics described the integration of omics methods to improve gene and genome (re-)annotations (also referred to as proteogenomics *sensus stricto*). However, a more wide application of the term (referred to as proteogenomics *sensus lato*) may also apply to studies focusing on improved protein identifications or visualization (Armengaud et al., 2014). Main applications such as gene annotation utilizes 6-frame translations or predicted open-reading frames (ORFs) of genome or transcriptome sequences as protein databases for tandem MS spectra searches in order to discover novel genes or protein-coding regions and to refine genome annotations in terms of e.g. exon/intron boundaries and alternative splicing (Nesvizhskii, 2014; Locard-Paulet et al., 2016). In addition, such customized databases aid peptide identification rates and the identification of novel peptides in cases where a comprehensive reference proteome is not available or enable comparative proteome profiling under different physiological conditions (Armengaud et al., 2014; Nesvizhskii, 2014). A third objective associated with proteogenomics is the mapping of previously identified spectra to genomic coordinates to enable integrative visualization (Sanders et al., 2011; Kuhring and Renard, 2012; Schlaffner et al., 2016).

Proteogenomics features many practical applications such as biomarker discovery for improved disease diagnosis, monitoring and therapy (Renuse et al., 2011; Armengaud et al., 2013), personalized medicine and monitoring (Locard-Paulet et al., 2016), analysis of cancer mechanisms and pathways (referred to as onco-proteogenomics) (Renuse et al., 2011; Menschaert and Fenyö, 2015; Locard-Paulet et al., 2016), improved characterisation of (human) pathogens and microbe-host interactions (Renuse et al., 2011; Locard-Paulet et al., 2016) as well as antibody sequencing and venomics (Menschaert and Fenyö, 2015). While proteogenomics facilitate general improvement in protein identification, non-model organisms with few or none reference sequence material benefit the most from combined omics analysis (Armengaud et al., 2014).

1.4. Non-Model Organisms

Experimental efforts in genomics and proteomics are often limited to established targets, i.e. model organisms. Therefore, reference proteomes of a great number

of organisms remain unavailable, incomplete or lack quality annotations. However, advancements in sequence technologies result in increased application of proteogenomic methods for unsequenced or partially sequenced organisms (Nesvizhskii, 2014). Non-model organisms play a key role in current and future research to acknowledge the full extent of biological diversity and thus benefit from continuous method development and transfer (Armengaud et al., 2014). The gain of popularity of non-model organisms is partially due to the constantly increasing amount of metagenomic and metaproteomic studies (Armengaud et al., 2014) since organisms in microbial communities are often poorly characterized (Nesvizhskii, 2014). The analysis of non-model organisms often relies on strategies including homologous organisms (Junqueira et al., 2008; Armengaud et al., 2014; Nesvizhskii, 2014) and is either limited to conserved proteins or needs to apply error-tolerant spectra identification (Renard et al., 2012). Thereby, microbes and unicellular organisms are particularly challenging for homology-based strategies due to their generally high diversity and occasional non-homologous coding-sequences (Armengaud et al., 2014). In contrast, microbial communities may additionally feature highly homologous groups of microbes which are difficult to distinguish (Nesvizhskii, 2014). Furthermore, homology-based search strategies are impractical for analysis with very meticulous objectives such as maximal taxonomic resolution or classification. In general, sample preparation somewhat follows "universal biochemical properties" and therefore adaptation of classic model organism methods for non-model organisms is relatively simple. In contrast, the computational analysis is considered the greater challenge with respect to unsequenced organisms with no close relatives available in reference databases (Armengaud et al., 2014) and therefore presents opportunities for methodological improvement.

1.5. Challenges in Omics Data Analysis

Next generation sequencing and mass spectrometry as well as genomics, proteomics and proteogenomics feature a highly diverse set of technologies, methodologies as well as applications. This results in a diverse set of unsolved problems and challenges, despite excellent preceding, recent and ongoing research in the respective fields. Nevertheless, the theme of sequence quality and availability is prevalent and common to all of these fields, in particular when methods are integratively used such as in proteogenomics.

Proteogenomic studies demonstrate the possibility of overcoming the lack of reference proteomes by simultaneously obtaining or consulting a draft genome. Acquiring a draft genome with the aid of NGS becomes continuously easier and more affordable. While mapping is most commonly used to create draft genomes based on NGS data for proteogenomic analysis (Menschaert and Fenyö, 2015), the application of *de*

novo assembly is necessary for many non-model organisms without a close reference genome (Henson et al., 2012; Ekblom and Wolf, 2014). However, low draft genome quality is a risk for applications in proteogenomics or gene annotation in general (Armengaud et al., 2013) since nucleotide uncertainties (in particular for transcriptome reads) and assembling errors increase the risk of errors in protein identification and sequencing based on, for instance, frame-shifts or incorrect ORF termination (Armengaud et al., 2014). In general, draft genomes as used in applications such as proteogenomics or gene annotation benefit from high quality genome sequencing and assembly procedures as it decreases the risk of misinterpreting subsequent analysis (Armengaud et al., 2013). Therefore, improved quality assessment and control of *de novo* assembled genomes can substantially support the analysis of non-model organisms, for instance in a proteogenomic context.

Apart from novel custom databases based on 6-frame translated draft genomes, tandem MS spectra analysis traditionally utilize existing and established protein sequences as provided by curated databases such as NCBI RefSeq and UniProtKB. However, independently of resources and application spectra database searches are based on two key assumptions: First, all genes of an organism are completely and thoroughly annotated and, second, their protein products are available in the reference database (Nesvizhskii, 2014). Completeness, integrity and overall high quality of protein databases are essential for identification performance but frequently not attainable, in particular for non-model organisms (Armengaud et al., 2014; Pible et al., 2014). Furthermore, the condition of the target database should not only be considered in final result interpretation but should support the choice of methodology applied. Therefore, databases need to be examined for their identification potential with respect to taxonomic resolution limits while considering different search conditions including exact and error-tolerant peptide matching as well as proteomic and proteogenomic databases.

Although current protein databases are still limited in content, for instance, with respect to non-model organisms, public curated as well as uncurated protein resources accumulate sequences day by day. Therefore, resources such as the NCBI Protein database (Wheeler et al., 2008) feature great potential for comprehensive tandem MS spectra database searches, not only for proteogenomics and non-model organisms but for detailed taxonomic classification in general, for instance for untargeted strain level identification. On the one hand, current MS biotyping methods are often limited to sets of common microbes (Singhal et al., 2015), specific species (Gekenidis et al., 2014; Pfrunder et al., 2016) or a restricted taxonomic depth (Alves et al., 2016; Boulund et al., 2017) and don't take full advantage of protein sequence data available, potentially excluding non-model organisms. On the other hand, proteogenomic studies, among others, regularly illustrate the impact of increasing search spaces and risks for accurate false discovery rate (FDR) estimation (Nesvizhskii, 2014). The likelihood of false-positive matches increases with

database size leading to a decrease in confidence and total number of identified peptides (Renuse et al., 2011; Jeong et al., 2012; Armengaud et al., 2014; Nesvizhskii, 2014; Menschaert and Fenyö, 2015). In addition, increasing numbers of ambiguous hits as well as contaminations (Pible et al., 2014) yield further challenges in the application of comprehensive database resources such as the NCBI Protein database. Therefore, viable taxonomic classification down to strain level needs to approach such database sufficiently and comprehensively enough to retain relevant taxa, but restricted enough to prevent false hits and to maintain confidence in identification.

1.5.1. List of abbreviations

Abbreviation	Explanation
bp	Base pairs
FDR	False Discovery Rate
FPR	False Positive Rate
GC	Guanine-Cytosine content
LCA	Lowest Common Ancestor
mad	Median absolute deviation
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectometry
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
PSM	Peptide Spectrum Match
RefSeq	Reference Sequence database of the NCBI
ROC	Receiver Operating Characteristic
sd	Standard deviation
SRA	NCBI Sequence Read Archive
TPR	True Positive Rate

1.6. Thesis Outline

This thesis introduces several computational methods for NGS and tandem MS data analysis with applications in (meta-)genomics, (meta-)proteomics as well as potentially proteogenomics. Chapter 2, 3 and 4 address the three previously described challenges related to sequence quality and availability and present possible solutions each. Chapter 5 discusses and summarizes the three contributed methods and corresponding results. All contributions were developed under the guidance of Bernhard Renard who participated in overall software and experimental design as well as in drafting manuscripts for publication. Software and experimental design for Chapter 2 were supported by Piotr Wojtek Dabrowski, Vitor Piro and Andreas Nitsche. Vitor Piro additionally contributed extensively to the evaluation of metagenomic analysis. Furthermore, Thilo Muth contributed to software and experimental design of Chapter 4.

Chapter 2 addresses the quality control of *de novo* assembled draft genomes which provide the basis for proteogenomic studies in the short term and for annotated, curated as well as publicly available proteomes in the long term. SuRankCo, a novel machine learning-based tool for quality ranking of *de novo* assembled contigs is presented and discussed. Benchmarks on datasets with known ground truth feature are presented and illustrate potential benefit, in particular for the integrative application within metagenomic samples. The chapter is based on the publication:

> *SuRankCo: supervised ranking of contigs in de novo assemblies.* M. Kuhring, P. W. Dabrowski, V. C. Piro, A. Nitsche and B. Y. Renard. BMC Bioinformatics, 16, 240, 2015.

Furthermore, the presented contribution in Chapter 2 builds upon preceding work in the master thesis:

> *Estimation Of De Novo Assembly Contig Quality With Random Forests.* M. Kuhring. Master Thesis, Freie Universität Berlin, 2012.

In the master thesis, features and scores were established and the utility of random forest as classifier was evaluated using several *de novo* assemblies of single organism sequencing samples. The contribution to this thesis improves on the overall prediction performance by introducing binary classification of preliminary contig scores by means of quantiles of fitted exponential distributions. Furthermore, scores were aggregated to enable contig rankings whilst taking into account the classification probability. Experiments and evaluations were substantially extended with respect to datasets and assemblers used. In particular, the additional application for metagenomic samples was investigated at length. Additionally, the machine learning strategy was compared to related published work on *de novo* assembly quality

assessment. Finally, the developed method prospered from an initial prototype into a complete and published software tool.

Chapter 3 discusses a possible solution to assess sequence data availability for tandem MS based microbial identification. Several examples illustrate the taxonomic performance for model as well as non-model target organisms with respect to possible influences of error-tolerant peptide searches in comparison to exact strategies. Furthermore, the impact of database extension with translated genomic sequence data is highlighted. The chapter is based on the publication:

> *Estimating the computational limits of detection of microbial non-model organisms.* M. Kuhring and B. Y. Renard. Proteomics, 15, 3580–3584, 2015.

Chapter 4 concludes on the subject of sequence availability by addressing the application of comprehensive databases for untargeted strain level identification of tandem MS samples. With TaxIt, an iterative approach is presented that enables the selective differentiation of suitable strain proteomes by identifying a candidate species first. Strain identification is demonstrated on several viral and bacterial samples. The chapter is based on the publication:

> *TaxIt: An iterative and automated computational pipeline for untargeted strain level identification of microbial tandem MS spectra.* M. Kuhring, T. Muth and B. Y. Renard. Manuscript in preparation.

2. Supervised Ranking of Contigs in De Novo Assemblies

In contrast to mapping procedures, *de novo* assembled sequences lack the direct comparison to a reference genome and thus have no ground truth-based quality control readily available. Commonly, evaluation of *de novo* assemblies and their contigs is based on single metrics (such as the N50) and their individual interpretation (Bradnam et al., 2013) or on evaluations of accumulated metrics or mis-assembly features (Phillippy et al., 2008; Vezzi et al., 2012a,b; Gurevich et al., 2013). Several methods and tools were released lately that introduced a new degree of quality detail on a nucleotide level, such as ALE (Clark et al., 2013), CGAL (Rahman and Pachter, 2013), LAP (Ghodsi et al., 2013) or REAPR (Hunt et al., 2013). They provide log-likelihoods based on probabilistic assumptions to allow quality comparison between different assemblies.

In this contribution, we focus on the aspect of quality control within a *de novo* assembly. We introduce a machine learning based method to evaluate and rank contigs within a single *de novo* assembly, called SuRankCo (**Su**pervised **Rank**ing of **Co**ntigs). The method takes advantage of data already generated in related sequencing experiments. It allows the selection of a suitable subset of contigs for subsequent processing and analysis.

In general, not every contig can be assumed to be error-free and it may save time and resources to re- strict downstream analysis to reliable information. In doing so, for instance, conflicts in finishing procedures may be prevented (Salzberg and Yorke, 2005; Nielsen et al., 2009), expensive validation experiments can focus on contigs of sufficient quality (Hsu et al., 2012; Mascher et al., 2013) and ambiguities in derived gene annotations may be explained by contig quality (Vázquez-Castellanos et al., 2014).

Surankco ranks contigs by their quality and can help in identifying the error source by the various scores it produces. However, it is outside of the scope of this thesis to improve low-ranking contigs and repair their errors. There are other strategies and tools which are applicable, e.g. the integration of different assembler types with non-overlapping error profiles (Salzberg et al., 2012), the application of error correcting tools for the reads (Kelley et al., 2010), or the critical visual inspection and manual correction (Nielsen et al., 2009).

The main idea of SuRankCo is to rely on knowledge generated from contigs from

sequencing experiments of related organisms for which a genome reference is available. Aligning these contigs to the reference yields scores which can be used as targets for a machine learning approach. Contigs from a new assembly can then be examined and classified with respect to the learnt target scores based on different features.

In the following, we introduce the methodology and implementation of SuRankCo, evaluate it on bacterial *de novo* genome assemblies and compare to ALE as an existing and related method.

2.1. Training and Prediction of Contig Quality

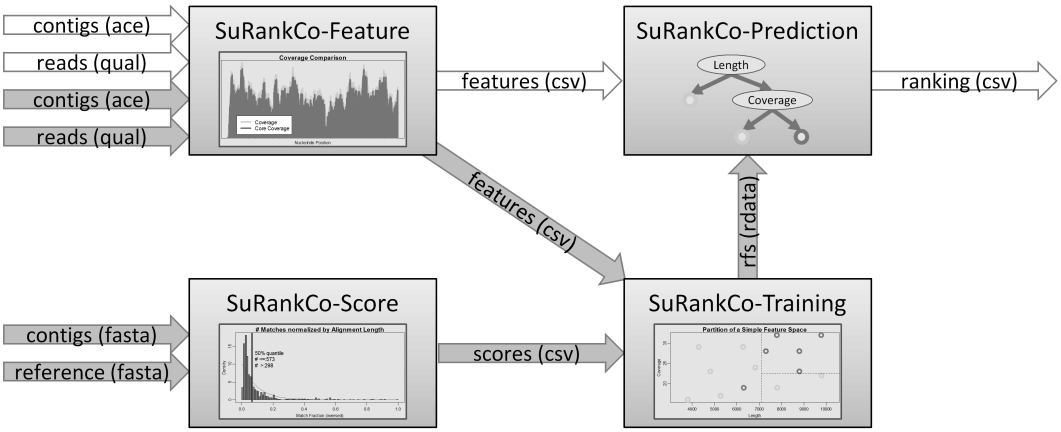

Figure 2.1.: Modularization and workflow of SuRankCo. The four modules of SuRankCo allow two workflows, training and prediction, indicated by grey and white arrows, respectively.

SuRankCo is divided into four modules (illustrated in Figure 2.1), including the extraction of contig features, the calculation of alignments and single scores, the training based on features and the prediction of single scores based on features to build the ranking. These modules can be combined to either perform training or prediction. In addition, intermediate data such as the features, single scores or trained classifiers can be examined or used within other applications.

Information on characteristics of contigs from a *de novo* assembly are extracted by the SuRankCo-Feature module. These features include common characteristics such as length (unpadded and padded), coverage, quality values, read counts, read lengths and read quality values. Additional features were developed, including core coverage, coverage confirmation and coverage drops. For a full list of features and

descriptions refer to Appendix A.1. SuRankCo-Feature accepts assemblies either as a pair of ace and fastq files or fasta and sam/bam files, respectively.

Training contigs are scored by comparison to a corresponding reference genome sequence. The SuRankCo-Score module utilizes BLAT (Kent, 2002) and accompanying tools to build alignments. Next, several single scores are calculated for each contig based on these alignments. Some scores are computed for each contig as a whole and some for certain critical areas such as the contig ends. Additionally, some scores are varied by introducing different normalizations, for instance based on contig or alignment length. A full list and descriptions of the single scores is given in Appendix A.1.

The classification of contigs in SuRankCo is performed using a random forest classifier (Breiman, 2001). Here, we rely on a random forest classifier as it adapts to different scenarios without the need for parameter tuning, can handle discrete and continuous input and can also uncover non-linear relationships. The training of the random forests is preceded by a separation of each single score into two classes to allow for binary classification using quantiles of fitted exponential distributions. Alternatively, a manual adjustment is possible based on histograms provided by the SuRankCo-Score module. A detailed description is given in Appendix A.1. Finally, the SuRankCo-Training module uses contig features and the transformed single scores to train a classification random forest for each score.

The SuRankCo-Prediction module estimates single score classes from contigs and their respective features by using the previously trained random forests. Different estimates are aggregated in a voting procedure to provide a final SuRankCo contig score. It is defined as $\sum_{i=1}^{|S|} S_i \times P_i$ where S_i is the $i^{th}s$ single score classification (0 or 1) and P_i denotes the probability of S_i being classified to that class, which is provided by the random forests. The SuRankCo contig score determines the final position in the ranking of the contigs.

2.2. Experiments

We evaluate the application and classification quality of SuRankCo by using various publicly available genome sequencing data sets. In the first experiment, we apply SuRankCo on the well-studied *Escherichia coli* strain K-12, substrain MG1655 (Blattner et al., 1997) and compare to ALE as an existing and related method. We constructed four *de novo* assemblies of Illumina Genome Analyzer II reads from the NCBI Sequence Read Archive (SRA), three for training and one for prediction and evaluation (accession numbers are provided in Appendix A.1). The training and the evaluation of the predictions make use of an established high quality reference [NCBI:NC_000913.3]. However, it should be noted that using the same organism for training and prediction is an artificial application as a proof-of-principle. Details on

the data preparation are given in Appendix A.1.

We calculated the classification quality for each single score by comparing predicted classes versus real classes. As additional validation with ground truth data, we compared the ranking based on the SuRankCo contig scores to the percentage identity (pIdent) of Blast hits in the current NCBI *E. coli* taxon [taxid:562], assuming that more reliable contigs should show better identity values.

Current methods for quality control in *de novo* assemblies do not score individual contigs, but rather focus on comparing complete assemblies. In order to still provide a meaningful comparison, we counted potential contig errors based on ALE subscores. Therefore, we manually evaluated the sub-scores and defined error thresholds (see Appendix A.1 - Figure A.1). Sub-scores below their corresponding thresholds are counted as error and errors are summed per contig over all positions. For the the *E. coli* prediction data set, these ALE contig scores were then compared to the Blast pIdent values in the same way as the SuRankCo contig scores. More details on the application of ALE are given in Appendix A.1.

To demonstrate the applicability for different organisms and assemblers, we applied SuRankCo on the staggered mock community of the Human Microbiome Project (Turnbaugh et al., 2007) and the bacteria assemblies of the GAGE study (Salzberg et al., 2012). We used three different settings for the mock community: (i) a metagenomics assembly, (ii) an organism specific assembly with different assemblers, and (iii) a combined training on assemblies by various assemblers. For (i), we constructed a meta-assembly of the complete community. We then assigned the resulting contigs to the respective organisms and then randomly divided the set of organisms in the community into a training and a prediction group. For (ii), we extracted all reads for each organism by a reference mapping procedure to have single organism sequencing data with identical technical origin. Each organism was then assembled separately using the assemblers Mira (Chevreux et al., 1999), SOAPdenovo (Luo et al., 2012) and Velvet (Zerbino and Birney, 2008). Training and prediction was performed for each assembler separately with a separation of organisms as in the metagenomics assembly experiment. For (iii), the assemblies of the different assemblers in (ii) were merged to provide a training and prediction data set across all organisms and assemblers. Details on the data preparation are given in Appendix A.1.

For the SuRankCo analyses of the GAGE bacteria, we made use of the assemblies, reads, and genomes provided for *Staphylococcus aureus* and *Rhodobacter sphaeroides*. In particular, we used the *S. aureus* assemblies for training and *R. sphaeroides* for prediction. We used two different settings for the GAGE assemblies: (i) an assembler specific training, and (ii) a combined training on assemblies by various assemblers. For (i), training and prediction was performed for each assembler used in the GAGE study separately. For (ii), the assemblies of the different assemblers were merged to provide a training data set across all assemblers. Details on the data preparation

are given in Appendix A.1.

To evaluate the mock and GAGE experiments, we compared the SuRankCo score rankings to Blast hits of contigs mapped against the corresponding known reference genomes. In particular, we calculate a contig evaluation score by forming the harmonic mean between the Blast pIdent and the Blast query coverage (qcovhsp). We then assigned the contigs based on the ground truth into a low-quality and a high-quality group and evaluated the performance of SuRankCo by ROC curves.

In addition, we compared the SuRankCo results of the GAGE assemblies to the corresponding GAGE evaluation metrics including contig number, errors, N50, and corrected N50. We calculated mean values of final SuRankCo contig scores per assembler in order to enable ranking based comparisons assuming a correlation between SuRankCo score distribution order of the different assemblies and their corresponding GAGE evaluation metrics.

2.3. Results and Discussion

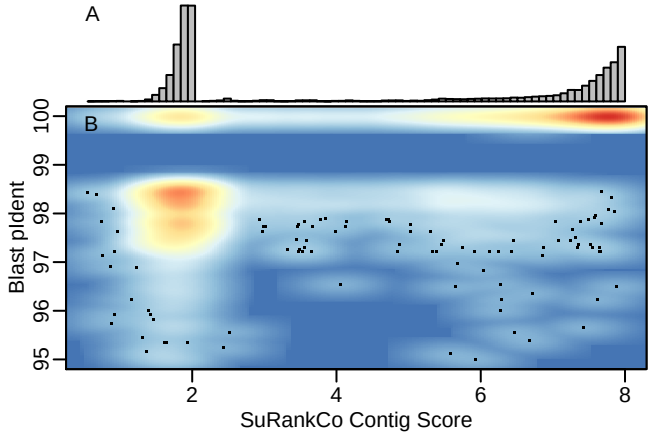

Figure 2.2.: Evaluation of the SuRankCo rankings on the *E. coli* test data. (A) shows the distribution of SuRankCo contig scores. They form two clusters based on the high correlation of target scores in this data set. Clusters are skewed due to classification probabilities incorporated into the SuRankCo contig scores. (B) shows a scatterplot comparison of the ranking and the pIdent of Blast matches against the *E. coli* taxon. High and low density areas are indicated in red and blue, respectively. Data points below 95 % pIdent are not shown to improve the scaling (25 of 11336).

The *E. coli* experiment illustrates three key characteristics of the single scores. First, the contigs used in training show good quality in their alignments to the reference sequence. Thus, they feature low variance in the single score distributions.

Second, these variances are still sufficient to allow an automated separation into two classes (see Appendix A.1 - Figure A.2). Third, a successful prediction can be made with a low number of false positives and false negatives in the test data (Appendix A.1 - Figure A.3). Further, the validity of the SuRankCo contig score is supported by a comparison to the percentage identity of the corresponding Blast hits (Figure 2.2 B) with Pearson and Spearman correlation coefficients of 0.77 and 0.72, respectively.

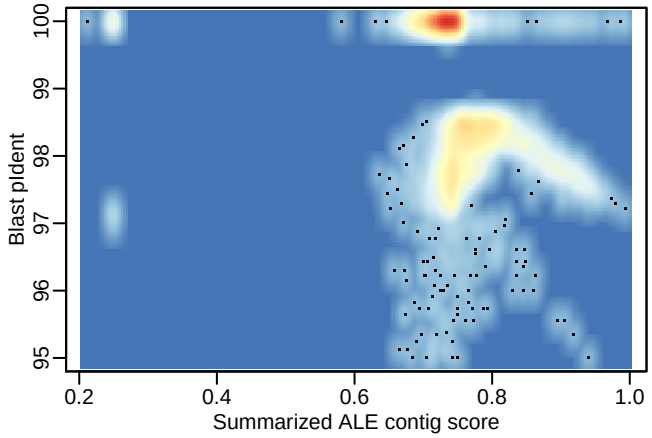

Figure 2.3.: Evaluation of ALE contig score of E. coli test data. The figure shows a scatterplot comparison of the ALE contig scores and the pIdent of Blast matches against the *E. coli* taxon. High and low density areas are indicated in red and blue, respectively. The ALE contig scores are shown in reversed order to allow a simpler comparison to Figure 2.2. Data points below 95 % pIdent are not shown to improve the scaling (25 of 11336).

Figure 2.3 shows a comparison of ALE contigs scores and Blast pIdent values. In addition, the comparative evaluation results for SuRankCo and ALE on contigs of varying length is shown in Table 2.1. Correlations between Blast pIdent values and SuRankCo contig scores are generally higher than correlations between Blast pIdent values and ALE contig scores, independent of whether Spearman or Pearson correlation is used and how long contigs are. However, it should be noted that ALE was applied here outside its regular scope and results should by no means be interpreted as general criticism of the tool. To the contrary, differences in the performance between SuRankCo scores and ALE scores only emphasize the differences regarding their approaches and objectives. The fact that ALE does not provide contigs scores directly further supports this observation.

For SuRankCo, a high correlation between the single scores is notable in the *E. coli* experiment (as shown in Appendix A.1 - Figure A.6). However, correlated scores do

Table 2.1.: Comparative evaluation of SuRankCo and ALE. The table shows the Spearman and Pearson correlations of SuRankCo and ALE contig scores to the percentage identity of corresponding Blast hits. The correlations are calculated for all contigs as well as separately for short contigs (with lengths below the 10% quantile) and long contigs (with lengths above the 90% quantile).

Score	Contig Length	$\text{Cor}_{Pearson}$	$\text{Cor}_{Spearman}$
SuRankCo	all	0.77	0.72
ALE	all	0.35	0.49
SuRankCo	$\leq Q_{0.1}$	0.58	0.55
ALE	$\leq Q_{0.1}$	0.16	0.37
SuRankCo	$\geq Q_{0.9}$	0.75	0.68
ALE	$\geq Q_{0.9}$	0.19	0.12

not corrupt the predictions, but favor clustering of contigs within the ranking rather than a more uniform distribution (compare Figure 2.2 A). In general, contig scores may be less correlated and thus provide a wider distribution of SuRankCo contig scores in the ranking as shown for the data of the metagenomics mock community experiment in Appendix A.1 - Figure A.7. In addition, the variety of SuRankCo contig scores enables a broader integration and indentification of common assembly error types (see Appendix A.1 - Table A.5 and A.6).

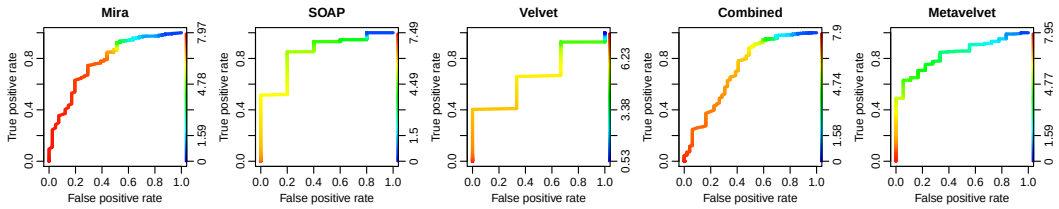

Figure 2.4.: Evaluation of the SuRankCo predictions of the mock community test data. Each plot illustrates a ROC curve of the contig evaluation score grouping in contrast to a varying grouping of the SuRankCo scores. Thereby, the changing color of the graph represents the changing threshold for the SuRankCo score grouping. Here, the predictions for the different organisms in the test group are combined to feature ROC curves of specific, combined and meta-assemblies.

The mock experiments allow a detailed view on parameters influencing SuRankCo results. Altogether, results indicate good prediction with regard to true positive rates (TPR) and false positive rates (FPR) (see Figure 2.4). However, some exceptions can be observed on the organism and on assembler level as exemplified in Appendix A.1 - Figure A.4. In general, merging the training data from various assemblers

does not improve on individual assembler results, but rather has negative effects. This indicates that there are assembler specific error types that can be learnt with SuRankCo. Comparing assembler results, the evaluation of Velvet assemblies performs poorly in contrast to the other assemblers. However, for Velvet we observed the lowest number of contigs with low quality based on the blast generated ground truth. This indicates that the performance of SuRankCo decreases for assemblies of very high quality since there is only few variance left for proper training or prediction. For organisms, we note that comparatively poor results are obtained for *S. epidermidis*, in particular for Mira, Metavelvet and the combined assemblers, although an apparently closely related organism (*S. aureus*) is present in the training data. However, examining the relation of mock organisms based on sequencing data reveals low similarities in general (as shown in Appendix A.1 - Table A.7).

Table 2.2.: Contig metric values of *R. sphaeroides* assemblies as provided by the GAGE study. Note, ABySS2 metric values were not available.

Assembler	Num	N50	Errors	N50corr
ABySS	1915	5.9	76	4.2
ALLPATHS-LG	204	42.5	49	34.4
Bambus2	177	93.2	373	12.8
MSR-CA	395	22.1	52	19.1
SGA	3067	4.5	12	2.9
SOAPdenovo	204	131.7	422	14.3
Velvet	583	15.7	43	14.5

Table 2.3.: Comparative evaluation of SuRankCo and GAGE. The table shows the Spearman correlations between assembly ranks based on SuRankCo score means and GAGE metrics for *R. sphaeroides* assemblies. Correlations are calculated for SuRankCo ranks based on assembler specific trained classifier as well as combined trained classifier.

	Num	N50	Errors	N50corr
Specific Training	0.7208	-0.7857	-0.6071	-0.6071
Combined Training	0.6847	-0.6786	-0.1786	-0.8571

Similar to the mock experiments, the GAGE experiments result in overall accurate predictions as illustrated by the ROC curves in Figure 2.5. However, few assemblies yield low prediction power including MSR-CA and SGA. The comparably low error rate in these two assemblies (as shown in Table 2.2) supports the conclusion that the performance of SuRankCo decreases for assemblies with very few errors. Since SuRankCo is a learning based approach, it requires also negative

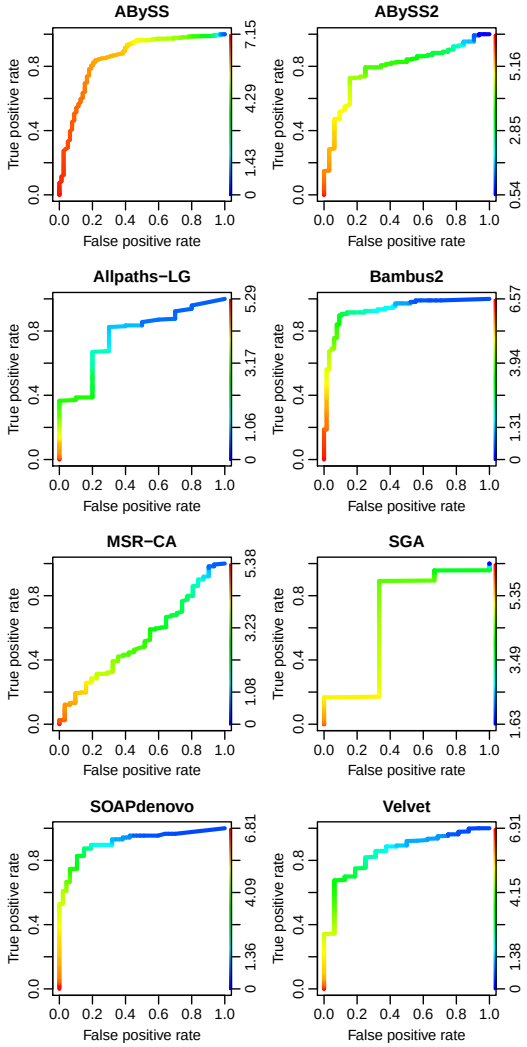

Figure 2.5.: Evaluation of the SuRankCo predictions of the GAGE assemblies. Each plot illustrates a ROC curve of the contig evaluation score grouping in contrast to a varying grouping of the SuRankCo scores. Thereby, the changing color of the graph represents the changing threshold for the SuRankCo score grouping. Here, one ROC curve represents the evaluation of *R. sphaeroidis* assemblies classified by the combined training classifier.

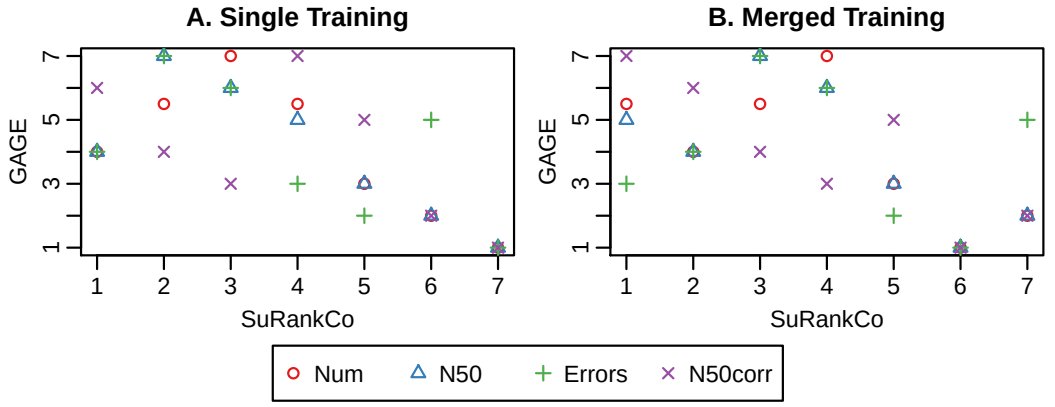

Figure 2.6.: Scatterplot of the SuRankCo score mean ranks and GAGE metric ranks. The figure shows scattorplots of ranks for GAGE assemblies of *R. sphaeroidis* based on the SuRankCo score means vs. each GAGE metric including contig number, errors, N50, and corrected N50. A. features SuRankCo score mean ranks based on assembler specific trained classifier and B. based on the combined trained classifier, respectively. To improve visualization, the contig number ranks have been inverted since they are the only ones yielding positive correlation.

examples containing errors in the assemblies. If these are missing, artifacts may arise more frequently. In summary, assemblies which provide a few, but potentially error-prone contigs may benefit more from SuRankCo than assemblers with a high number of short, but error-free contigs. In contrast to the mock experiment, on average there is no significant difference between predictions based on assembler specific (Appendix A.1 - Figure A.5) or combined training (Figure 2.5). However, the correlation of SuRankCo score means with the GAGE error metric shows a significant decrease from assembler specific to combined training based predictions (Table 2.3). Again, this indicates that there are assembler specific characteristics that can only be learnt and discriminated by separate training. Apart from that, the comparison of SuRankCo and GAGE yields good rank correlations with values of up to 0.85 as shown in Table 2.3 and Figure 2.6 for both, assembler specific and combined training and prediction. Therefore, based on independent ground truth data, the correlations indicate that SuRankCo infers the relationship of different assemblies in terms of quality, even if trained separately. Nonetheless, as also indicated by the diversity of the metrics in the GAGE study itself, it is difficult to perfectly represent the quality of assemblies in few scores. Thus, it cannot be expected to observe a direct one-to-one correspondence of SuRankCo scores with single GAGE metrics. At the same time, it should be noted that SuRankCo was developed to score individual contigs and that the overall ranking of assemblies by their mean ranking score – while well

correlated with the metrics in the GAGE study – is not its standard usage.

In classic assembly metrics such as the N50, a high value is placed on obtaining longer contig scores. However, it has been frequently noted that longer contigs scores do not necessarily coincide with higher contig quality (Narzisi and Mishra, 2011). SuRankCo scores are evaluated with regard to the identity and query coverage of the reference genome. Increasing values in these metrics may correlate with longer contigs, but are by no means ensured and rather focus on the number of matches and mismatches.

Overall, several factors may influence the assembly of contigs significantly and thereby also influence the performance of SuRankCo. These include for instance sequencing parameters such as coverage and read length, sequencer error profiles, organism relationships, biases such as GC content and characteristics of read processing algorithms such as these used for *de novo* assembly. Thus, SuRankCo is mainly designed with a focus on stable workflows applied within a lab. SuRankCo has been mainly developed for and tested on microbial genomes, however, there is no theoretical limitation which should restrict the application to other genomes.

2.3.1. Conclusions

We introduced SuRankCo as a tool for a learning-based quality prediction and ranking of contigs within a *de novo* assembly. To take full advantage of the machine learning approach and for optimal performance, training and test data have to be similar in their key characteristics. In our benchmark, we observe promising results in terms of sensitivity and specificity and favorable comparison to existing methodology. We foresee practical application in ranking contigs for downstream analyses.

SuRankCo is available for download under open-source license at `http://sourceforge.net/projects/surankco/`.

3. Limits of Detection of Microbial Non-Model Organisms

In recent years, there has been an increasing interest in using mass spectrometry for studying microbial non-model organisms (Armengaud et al., 2014). In particular, this holds true in metaproteomic studies which allow studying the composition of more complex and heterogeneous microbial communities (Muth et al., 2013; Penzlin et al., 2014). Non-model organisms are often only inaccurately covered or even entirely missing in available protein reference databases. As a consequence, the identification of spectra from these experiments is not completely achievable via common protein database searches, leaving an unknown number of proteins or even organisms undetected. Due to constant evolution of microbial organisms and the resulting vast diversity, even ongoing sequencing efforts will not overcome this challenge. Depending on database size and coverage, spectra of novel organisms often match to sequences of more or less related organisms resulting in peptide identifications in different taxonomic levels such as genus, family or class. The quantity and ratio of matches to these taxonomic levels is, however, not directly assessable for organisms with uncertain origin. Thus, a final taxonomic classification of such organisms is often inaccurate and unreliable.

Some research has been carried out on improving the identification of MS/MS spectra in general by the use of error-tolerant hybrid approaches integrating *de novo* sequencing and database searches (Renard et al., 2012) or by metaproteogenomic approaches (Seifert et al., 2013). However, the effect of extended search strategies on taxonomic classification of non-model organisms remains unknown prior to experiments. In particular, whether an available database is sufficient or should be supported by additional methods and data is inherently difficult to assess. Thus, when designing experiments often only a rough estimation can be performed to decide on the subsequent search strategy, e.g. whether parallel sequencing experiments are undertaken to allow proteogenomic approaches. In Figure 3.1 A, we illustrate which benefits extended search strategies can provide to databases with varying coverage: With a complete reference database, both standard and error-tolerant search will identify the correct organism for an MS/MS spectrum. In contrast, when there is no perfect match, only error-tolerant strategies may still identify the closest matching organism.

Within this contribution, we introduce the LiDSiM (LImits of Detection SImula-

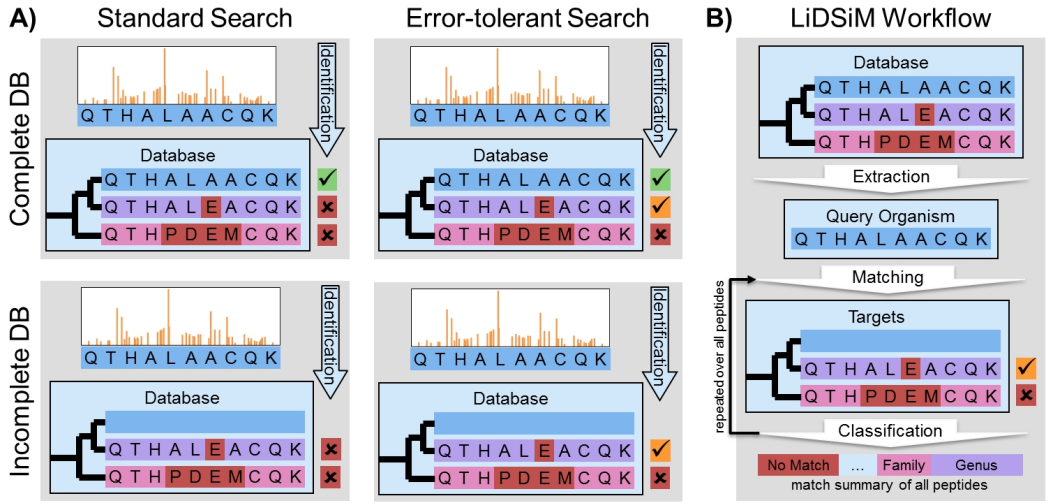

Figure 3.1.: Comparison of search strategies vs. database content and schematic LiDSiM workflow. A) illustrates the identification process for an MS/MS spectrum depending on the search strategy used and database completeness. A single query peptide (spectrum) is used as an example in all cases. Error-tolerant matches are indicated by orange tick marks. B) illustrates how LiDSiM estimates the number of expected matches per taxonomic level for an error-tolerant search against a database of interest. This comprises the extraction of an organism, the iterative matching of its peptides against the remaining database and the classification of the taxonomic distances of the matches to infer taxonomic level ratios.

tion for Microbes) method. The focus of LiDSiM is to evaluate whether a database is suitable for a specific identification task. Therefore, LiDSiM systematically estimates the ratio of MS/MS spectra that are identifiable at various taxonomic levels via exact and error-tolerant search strategies, based on available database information as well as next generation sequencing reads. We do not regard spectral information as numerous tools are available for this (Sturm et al., 2008; Muntel et al., 2014), but rather exclusively focus on estimating whether the database will contain the corresponding sequences.

A schematic workflow of LiDSiM is illustrated in Figure 3.1 B. The workflow indicates how to estimate the number of expected matches per taxonomic level for an error-tolerant search. For a given organism, all proteins are extracted from the database and digested into peptides. These are then searched against the remainder of the database. Thereby, the absence of an organism from a target database is simulated and peptides need to be identified at higher taxonomic distance. Peptide identification and taxonomic classification are repeated over all peptides to obtain taxonomic level ratios. The error-tolerant search enables the identification of more closely related peptides (as depicted by the example peptide) resulting in higher ratios of lower taxonomic levels. In contrast, a standard search would yield, for instance, less genus hits and more unidentified peptides.

3.1. Simulation and Identification of Related Organisms

The described extraction procedure is extended to all organisms in the database under analysis. This corresponds to a cross-validation-inspired strategy. Similar to leave-one-out cross-validation, individual organisms or taxonomic branches (representative for the species of interest in the experiment, e.g. the closest known relatives) are removed from the protein database of interest one by one. Additionally, the database can be partially extended by a proteogenomic approach, by including (simulated) genome sequencing data to examine its impact on the taxonomic classification. For instance, six-frame translations of annotated contigs may introduce an even higher, homology-based error-tolerance and therefore yield more identifications. Finally, the given origin of the artificial spectra allows the classification of the taxonomic distance of the search results. Thereby, we estimate to which extent the sequences of related organisms can be used to identify an organism not contained in the database.

The implementation of LiDSiM comprises several stages including database and taxonomy parsing, subset extraction, peptide generation, the database search and the taxonomic classification of the matches. The protein database is expected to be annotated with NCBI GIs to allow the classification within the NCBI taxonomy. An annotation with GI identifiers may be obtained by using, for instance, the UniProt ID

mapping service (The UniProt Consortium, 2014). A taxonomic tree representation is constructed from local files including GI-to-taxid mappings and tree nodes with parent and rank information. Given the taxonomy, one or more query subsets are extracted from the database. These subsets represent the proteins measured in an MS/MS experiment. A subset could contain, for instance, a single species or even major taxonomic segments such as genera, families etc. For larger subsets and databases, representative organisms can be specified or sampled. For each subset, the remaining proteins represent the target database used for spectra identification. Query protein sequences are in silico digested at the common trypsin cleavage site (after K and R, if not followed by P). All "I"s are substituted by "L" since they are not distinguishable due to their equal molecular mass. In addition, peptides are filtered for lengths of 8 to 35 and can randomly be downsampled in number to speed up the computational process. Since the focus of LiDSiM is to evaluate the database, we do not simulate the peptide spectrum match itself, but rather evaluate the matching of sequences across species in a database and the benefit of error-tolerant searches and proteogenomic approaches. For this aim, we rely on established string matching algorithms for sequence comparisons. These include the Wu-Manber algorithm (Wu and Manber, 1994) for exact parallel pattern matching and the approximate Boyer-Moore algorithm (Tarhio and Ukkonen, 1993) to allow matches with a definite maximal hamming distance. Wu-Manber and Boyer-Moore allow the deterministic search of pattern strings (spectra respective peptides) in target strings (protein database) without missing a possible hit. Thereby, the later algorithm allows the simulation of error-tolerant spectra identifications. Finally, for each match the lowest common ancestor (LCA) is calculated between query peptide and target protein. For each query, the match and rank of the closest LCA is reported. Counting these ranks results in the number of queries matching to certain taxonomic levels. Thus, we refer to the proportion of all ranks as taxonomic level ratio.

For particular organisms in the database, results can benefit from integrating genome sequencing data in a proteogenomic approach. However, it requires a *de novo* assembly for those organisms, either from experimental data or, at least, constructed from simulated reads. A simulated assembly can be generated, for instance, by using the Mason read simulator (Holtgrewe, 2010) and a *de novo* assembler such as Mira (Chevreux et al., 1999). The contigs of an assembly are six-frame translated using EMBOSS transeq (Rice et al., 2000) and a basic annotation is conducted by using BLAST+ (Camacho et al., 2009) against the database excluding the query proteins and selecting the best hit (i.e. the first hit with the smallest e-value). Besides the provided basic annotation procedure, the LiDSiM analysis can also be combined with more sophisticated gene annotations of the user's choice.

3.2. Comprehensive and Targed Database Evaluation

To illustrate the application of LiDSiM, we applied the simulation on databases composed of NCBI RefSeq bacteria proteins (Release 66) of selected phyla, in particular the *Proteobacteria* phylum (taxid 1224). We provide a cross section view of detection levels by extracting each genus one at a time. For each genus, a representative organism is selected randomly and sampled down to 1000 peptides. Each resulting subset is searched in the corresponding genus-free phylum database, once with an exact peptide search and once with error-tolerance allowing one amino acid substitution.

In addition to the phyla cross validation, we demonstrate the effect of additional search strategies on particular organisms with varying degree of exploration. We selected two bacteria with high contrast in the number of relatives in their corresponding phylum databases. The first bacterium is *Escherichia coli* which is highly common in terms of scientific exploration and taxonomic density of relatives in the RefSeq database. And second, we selected the by contrast relatively uncommon bacterium *Deinococcus deserti*. Both bacteria were extracted to varying extent from their corresponding phylum databases (*Proteobacteria* phylum, taxid 1224, 4192793 proteins and *Deinococcus-Thermus* phylum, taxid 1297, 57694 proteins, resp.). This includes extracting the species, genera and families each (taxids 562, 561 and 543 for *E. coli* and taxids 310783, 1298 and 183710 for *D. deserti*, resp.). Representatives were selected manually to enable validation with experimental data and include the *E. coli O157:H7* strain Sakai (taxid 386585) and the *D. deserti* strain VCD115 (taxid 546414). In the simulation, query peptides were sampled down to 1000 peptides each and searched in the corresponding databases with and without error-tolerance.

3.3. Experimental Validation

We evaluate the simulation by comparing against real data. Therefore, we selected publicly available MS/MS spectra from the PRIDE archive including an *E. coli* O157:H7 strain Sakai experiment (PXD000583) (Kocharunchitt et al., 2014) and a *D. deserti* strain VCD115 experiment (PRD000139) (Baudet et al., 2010). Spectra were selected and prepared as described in Appendix A.2. We applied MS-GF+ (v9979) (Kim and Pevzner, 2014) to identify the spectra with an exact database search. For details on parameters refer to Appendix A.2. We sampled 1000 spectra from each dataset to improve runtime and to be consistent with the simulation. The sampled spectra were searched in the corresponding species-free phylum database (taxid 1224 and 1297, resp.). Results were filtered via a decoy-based false discovery rate (FDR) cutoff of 0.01. Given the origin of the spectra, we calculated and reported the matches with the closest LCA just as in the simulation. Since the focus of

LiDSiM is on protein identification rather than quantification, we only regarded the first occurrence of a spectrum for a protein rather than counting every occurrence. Thereby, we account for protein abundance in the experimental data which is not present in the simulation. The simulation may still contain a minor bias since unexpressed proteins cannot be predicted and removed. However, the impact is negligible in contrast to the excessive number of peptides resulting from highly expressed proteins. Finally, to investigate spectra sampling variance we repeated the complete procedure with 10 sampling replicates per bacterium each.

We applied BICEPS (v1.0) (Renard et al., 2012) to increase the error-tolerance of spectra identification by allowing one amino acid substitution per spectrum. BICEPS is an error-tolerant search approach that allows overcoming 1-2 amino acid substitutions in a peptide sequence within a database search. Thereby, by using BICEPS, a higher number of matches to related species – which are likely to show some amino acid substitutions - can be identified. BICEPS results of identified substitutions can then be used as input to a standard MS-GF+ search to have the same scoring scheme as in exact searches without error-tolerance.

To analyze the impact of complementary genome sequencing data on the simulation, we constructed simulated *de novo* assemblies for *E. coli* O157:H7 strain Sakai and *D. deserti* strain VCD115 and, in addition, a *de novo* assembly for *E. coli* based on experimental reads (SRR587217) from the NCBI Sequence Read Archive (Leinonen et al., 2011). The sampling of artificial Illumina reads and the assembly procedures are described in Appendix A.2. The assemblies have been annotated as described above, added to the corresponding RefSeq protein target databases and searched with error-tolerance.

3.4. Results and Discussion

The iterative extraction of genera from the *Proteobacteria* phylum database resulted in 267 taxonomic level ratio estimations. The simulation results are shown in Figure 3.2 A and Figure 3.2 B for searches with and without error-tolerance. The exact search yields many extractions with a high number of unidentified peptides. On average, over 80% of peptides have not been identified at all. In contrast, only very few samples possess such high ratios for identified peptides. In comparison, the simulation with error-tolerance of one amino acid shows a substantial increase of identifications. The number of unidentified peptides is reduced to 60% on average yielding considerably more identified peptides, in particular within the family rank. In general, the majority of identified peptides are classified within the family rank, i.e. the peptides matched to proteins of rather close relatives.

The taxonomic level ratio estimations of the selected bacteria *E. coli* and *D. deserti* and the corresponding real spectra identifications are illustrated in Figure

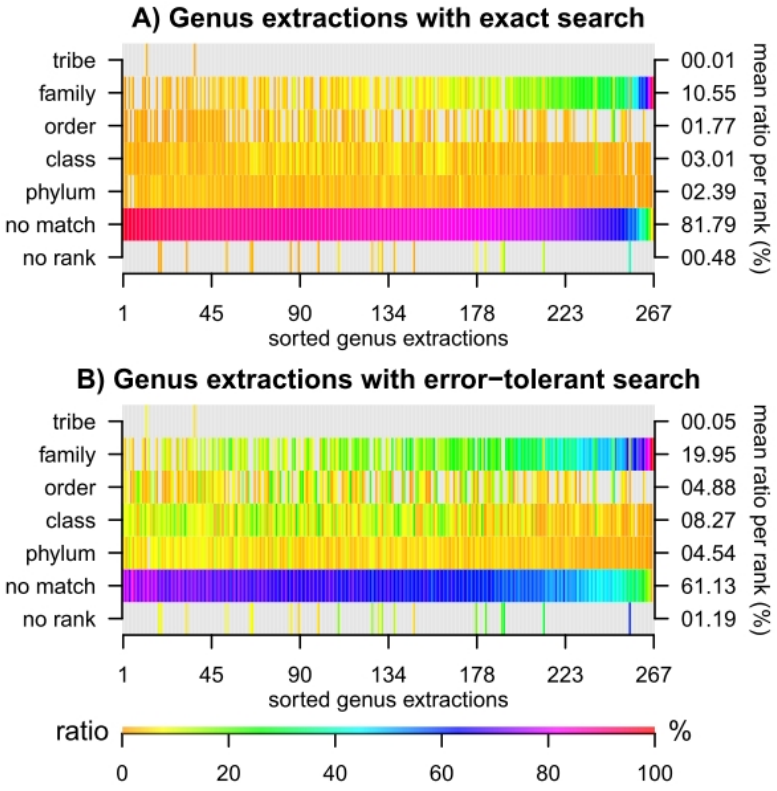

Figure 3.2.: Taxonomic level ratios of genus extractions. A and B show the simulations of peptide searches with an exact search and error-tolerant search, respectively. Each genus extraction is represented by a column in the heat map, with the same position in A and B. Extractions are sorted by the number of unmatched peptides (no match), phylum, class, order and family in the exact search. Minor taxonomic ranks (incl. sub, super, infra and parv ranks) are pooled with their corresponding major rank.

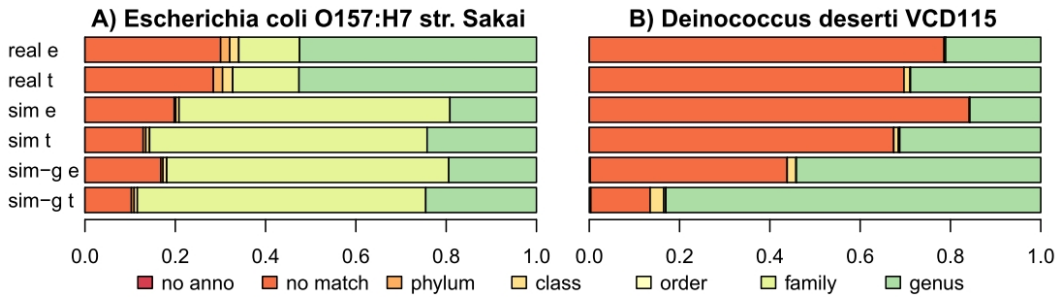

Figure 3.3.: Taxonomic level ratios of selected bacteria. For both, E. coli O157:H7 strain Sakai (A) and D. deserti strain VCD115 (B) the plots show the taxonomic level ratios of the first real spectra identification sample (real), the standard simulation (sim) and the proteogenomic simulation (sim-g). Results are shown for searches against databases with the corresponding species removed in an exact (e) and error-tolerant (t) search, respectively.

3.3. The results show favorable similarities between estimations and real spectra identifications, in particular for the ratios of identified and unidentified spectra. However, for *E. coli* real spectra are more often classified as genus matches than estimated. Additionally, the benefits introduced by error-tolerant and proteogenomic approaches are rather small but still apparent. In contrast, the number of *D. deserti* identifications significantly increases when error-tolerant and proteogenomic approaches are applied.

The results reveal a discrepancy between simulation and real spectra identifications for the E. coli experiments. This may be explained by several variables the simulation does not account for. While the simulation is a basic deterministic procedure, real spectra identifications are subjected to several additional effects of different origin, including, for instance, variations and errors in peak mass and intensity, different resolutions and noise. In addition, as previously mentioned the simulation might have sampled proteins from the database which are not expressed in the experimental data. However, the results of multiply sampled spectra for the real spectra identifications (see Appendix A.2 - Figure A.10) show a notable variance for unidentified spectra and thereby account for the ratio discrepancy of identified and unidentified peptides between simulation and real spectra identifications.

Both, the iterative genus extractions and the selective species extractions demonstrate the high variance in taxonomic level ratios between different samples or organisms. This variance originates from the varying density of relative organisms and sequences in the target databases. However, the analysis highlights the impact of error-tolerant and proteogenomic searches. In particular, a great benefit can be expected for comparably underexplored organisms such as the *D. deserti* strain

VCD115 in contrast to well explored organisms such as *E. coli* O157:H7 strain Sakai which only features minor improvements in the taxonomic level ratios.

In this contribution, we presented a method to estimate the taxonomic level ratios of MS/MS spectra identifications and, in particular, the amount of unidentified spectra with respect to a target database. The simulation evaluates the detection potential and limits of a specific database when applied for MS/MS spectra identifications. Furthermore, the simulation evaluates how these limits are affected by error-tolerant spectra identifications provided by capable search methods or proteogenomic approaches. While providing a comprehensive overview across organisms, the presented results are by no means intended for a direct application on other organisms and databases. To the contrary, we designed LiDSiM to support experiments by estimating the effect of error-tolerant and proteogenomic searches on particular databases as needed. Thereby, we provide a tool for experimental design, allowing researchers to decide in the planning stage of an experiment which benefit to expect from various strategies.

LiDSiM is developed in Java and R and is available for Linux and (with minor restrictions) for Windows at `https://sourceforge.net/projects/lidsim/`.

4. Iterative and Untargeted Strain Level Identification

LC-MS/MS driven strain identification is a crucial yet challenging task. Many microbial strains feature significant phenotypic differences within a species including differences in pathogenicity, zoonotic potential, cell attachment and entry, host-virus interaction and clinical symptoms (Bengali et al., 2009, 2012; Doellinger et al., 2015). Strain level knowledge is important to infer virulence (Choi et al., 2002; Genersch et al., 2005) and drug resistance (Boulund et al., 2017) for appropriate therapy. However, inferring strain information from proteomic samples remains challenging, in particular when the taxonomic status of a sample is unknown.

In the recent years, MALDI-TOF mass spectrometry gained in popularity as a fast, sensitive and economical method for microbial biotyping. However, identifying strains via peptide mass fingerprints requires curated and generally proprietary spectral databases (Singhal et al., 2015). Several commercial platforms for microbial biotyping down to species or strain level are available such as the Bruker MALDI Biotyper Systems (Bruker MALDI), the Bruker Strain typing with IR Biotyper (Bruker IR) and the Ibis T5000 Universal Biosensor (Ecker et al., 2006).

Several studies report on insufficient performance of MALDI-TOF biotyping for strain level identifications and advocate advancements towards tandem MS marker peptide detection. However, database searches are often already targeted or restricted to particular species or limited sets (Gekenidis et al., 2014; Pfrunder et al., 2016). In contrast, untargeted tandem MS typing approaches are limited to species level identification (Alves et al., 2016; Boulund et al., 2017). However, in general tandem MS is preferred for the analysis of complex unpurified peptide mixtures as it is considered to provide more distinct and unambiguous peptide and protein identifications (Aebersold and Mann, 2003) and thus increased proteome resolution (Domon and Aebersold, 2006) as well as higher statistical confidence (McHugh and Arthur, 2008). In particular, unknown organisms benefit from peptide sequence-based identification (Liu et al., 2007). Furthermore, advances in instrumentation including higher resolutions, mass accuracy and dynamic range increasingly allow for identification of the majority of all fragmented peptides (Mann and Kelleher, 2008) resulting in higher sensitivity, higher coverage of target proteomes and thus higher availability of distinctive features.

Taking advantage of the vast amount of available protein sequences for tandem

MS strain level identification is challenging. On one hand, constraining the search space may result in unidentified strains or incorrectly assigned taxa, in particular for non-model organisms (Kuhring and Renard, 2015). On the other hand, applying large databases is not recommended either since it decreases peptide identification rates (Jeong et al., 2012) and thus eventually impedes taxonomic inference. Furthermore, with increasing database size sequence quality often decreases (e.g. when using the complete NCBI Protein in comparison to the NCBI RefSeq database) and contaminations may occur more often (Pible et al., 2014). Therefore, extended databases should only be used when necessary. However, strain level identification of tandem MS spectra from samples with unclear taxonomic status requires an untargeted search against comprehensive databases holding as many strains as possible.

A common and popular concept to handle increased search spaces is the application of multiple identification steps in general, independently of target application such as strain level identification. These tandem MS search strategies are described by several different terms such as multi-step (Craig and Beavis, 2003), iterative (Nesvizhskii et al., 2006; Rooijers et al., 2011), multi-stage (Ning et al., 2010), two-step (Jagtap et al., 2012, 2013, 2014) as well as cascade search (Kertesz-Farkas et al., 2015) and they find application in proteomics, metaproteomics (Rooijers et al., 2011; Jagtap et al., 2012) and proteogenomics (Chapman and Bellgard, 2014; Jagtap et al., 2014). Most of these strategies do not only overlap in their objective of increasing the identification rate or identification confidence but share methodological principles as well. This includes the concept of identifying primarily unassigned spectra with databases of increasing complexity (for instance, by employing altered digestion parameters, additional post-translational modifications or additional spectral and genomic databases) (Craig and Beavis, 2003; Nesvizhskii et al., 2006; Ning et al., 2010; Rooijers et al., 2011; Kertesz-Farkas et al., 2015) as well as the recurring theme of database size reduction (Craig and Beavis, 2003; Rooijers et al., 2011; Jagtap et al., 2012, 2013; Chapman and Bellgard, 2014; Jagtap et al., 2014). In addition, some methods rely on spectral quality assessment to enhance subsequent identification steps (Nesvizhskii et al., 2006; Chapman and Bellgard, 2014) or exhibit a focus on algorithmic runtime reduction (Craig and Beavis, 2003).

These multi-step procedures illustrate the effect of database size on identification confidence and the advantage of applying concise, prefiltered or specialized databases. Thus, we transfer the general concept of multi-step procedures to approach the increased search space necessary for untargeted and detailed taxonomic classification. We present TaxIt, an iterative workflow for untargeted strain level identification of microbial protein samples. By applying two separate identification steps for species and strain level classification, we circumvent the immediate need for a comprehensive strain sequence database. Thereby, a first untargeted search allows the selection of a relevant species and enables to focus a second search on a highly

reduced but adequate choice of strain proteomes, resulting in increased identification confidence and reduced taxonomic ambiguity. Moreover, the workflow takes advantage of a free, publicly available and continuously growing protein sequence resource (NCBI Protein) and is thus suitable for most common established tandem MS instrumentation and workflows.

4.1. Traversing the Comprehensive Search Space

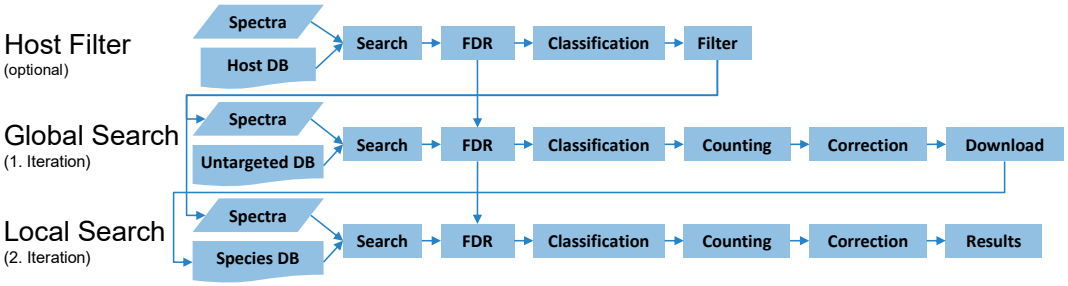

Figure 4.1.: Overview of the TaxIt workflow. In up to three stages, spectra are searched against a host proteome (optional) to pre-filter host proteins, against an untargeted database for global species identification (primary iteration) and against a targeted, species-based and automatically fetched database for strain level identification (secondary iteration).

The TaxIt workflow for strain identification consists of several recurring modules interconnected, controlled (in terms of input and output) and automatically executed by the Snakemake workflow management system (Köster and Rahmann, 2012). The designed workflow executes up to three stages including an optional host filter, species identification (primary iteration) and strain identification (secondary iteration). A concise overview of the workflow is illustrated in Figure 4.1. The main iterations comprise the execution of a peptide search engine, false discovery rate (FDR) control, taxonomic classification, taxa counting and adjustment as well as candidate selection and visualization. The download of strain proteomes bridges the primary and secondary iteration. The procedures of the main modules are described in detail below.

4.1.1. Peptide Search

The central step in tandem MS spectra analysis is an efficient peptide search. Therefore, we rely on established and reliable open-source search engines such as X!Tandem (Craig and Beavis, 2004) in combination with the XTandem Parser (Muth

et al., 2010) or MS-GF+ (Kim and Pevzner, 2014). However, any command-line search engine including proprietary ones could be implemented via additional Snakemake rules. We apply a classic target-decoy approach for false discovery rate (FDR) control. Decoy sequences are created upfront and independently of the search engine with fasta-decoy.pl (Masselot) by reversing the target sequences (including contaminant proteins, for instance cRAP (The Global Proteome Machine)) and both target and decoy databases are concatenated as suggested by Jeong et al. (Jeong et al., 2012). Peptide spectrum matches (PSMs) are subjected to an FDR cutoff based on a per-match FDR calculated as N_{decoy}/N_{target}, with N_{decoy} being the number of decoys in between targets (N_{target}) in a list of matches sorted by e-value (Jeong et al., 2012; Nesvizhskii, 2014). To acknowledge established false positive hits of previous iterations, decoy sequences left after FDR control are passed on and concatenated to the database of the next iteration.

4.1.2. Taxonomic Classification

We make use of the NCBI Taxonomy to assign PSMs to corresponding taxa and redistribute shared hits introduced by proteins associated with higher taxonomic level such as genus. First, NCBI protein accessions are mapped to NCBI taxids using the NCBI protein id mapping file (`ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz`). Next, taxonomic relations are inferred from the NCBI Taxonomy nodes dump file (`ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz`) and PSMs are reassigned to a taxonomic level based on the objective of the current iteration. In general, PSMs assigned to higher taxa such as genus are propagated as shared hits to corresponding leaf taxa as long as the target leaves already exhibit matches on their own. In addition, in the first iteration, PSMs are summarized at species level.

4.1.3. Count Adjustment

For each candidate taxa (i.e. species in primary and strains in secondary iteration, respectively), raw counts are calculated by summing over all assigned PSMs including non-unique matches. To account for taxonomic biases due to shared matches, we integrate a simple weighting scheme based on the level of uniqueness using the global frequency of a PSM. Here, a PSM count per taxa is adjusted by the number of occurrences in all candidate taxa. Thereby, unique PSMs gain in value for the final taxa selection without fully neglecting the importance of high numbers of non-unique matches which are often highly present within closely related taxa such as strains.

4.1.4. Selection and Downloads

After count adjustment, the most dominant candidate taxon is selected as the most likely species or strain, respectively. In the final step of the first iteration, the selected species is utilized to query strain level material for the strain identification in the second iteration. Once more, we rely on the NCBI Taxonomy and infer all available strains for the candidate species via the nodes dump file. Next, strain proteins are automatically downloaded from NCBI Protein using the NCBI Entrez API (Wheeler et al., 2008) in combination with the jsoup: Java HTML Parser (Hedley). This includes all available RefSeq as well as non-RefSeq sequences since the availability of curated strain material is often limited. Finally, the obtained protein sequences are merged into one database and redundant entries are removed using seqkit rmdup (Shen et al., 2016).

4.2. Experiments

We compare TaxIt against classic comprehensive search strategies based on straight non-iterative taxonomic identification supported by unique PSMs or abundance similarity correction as provided by Pipasic (Penzlin et al., 2014).

TaxIt will utilize NCBI RefSeq proteins of selected kingdoms as reference databases for initial species identification followed by automated and selective strain protein incorporation. Uniques- and Pipasic-based strategies however, will apply comprehensive databases integrating as much strain level sequences as possible at once including all protein sequences from the NCBI Protein database for selected kingdoms. In general, a preselection of kingdoms may be supported by clinical findings based on, for instance, symptoms or microscopic examination (Laue, 2010). Both, uniques- and Pipasic-based strategies utilize the same procedures for peptide search, FDR control and taxonomic classification as described in the iterative workflow. However, PSMs are not summarized at species level and counts are directly inferred at the lowest possible taxonomic level. For the uniques-based strategy, adjusted counts are based on PSMs which occur only once. Pipasic's abundance similarity correction utilizes the similarity of expressed proteomes between taxa to account for attribution biases. Originally intended for metaproteomic abundance correction, it is here applied to highlight the most likely strain. Since Pipasic is sensitive to a high amount of taxa, we limit the input to taxa with a minimum of two hits as well as to the most 100 abundant taxa. Expressed proteins per taxa are extracted according to the taxonomic classification and digested peptides with a minimum length of six amino acids are prepared using trypsin digestion (yafeng, 2017). PSMs and tryptic peptides are then passed to Pipasic to obtain corrected relative counts.

We perform all three strategies on several viral and bacterial tandem MS spectra samples with available strain level knowledge. This includes a *Cowpox virus*

(Brighton Red) strain (in-house sample), an *Avian infectious bronchitis virus* (strain Beaudette CK) (Dent et al., 2015) and a *Bacillus subtilis* BSN238 (Trappe et al., 2017) which we will refer to as cowpox, bronchitis and bacillus sample, respectively. *B. subtilis* BSN238 is a transgenic organism resulting from horizontal gene transfer (HGT) of the DivIVA protein from *Listeria monocytogenes* strain EGD-e to *Bacillus subtilis* subsp. *subtilis* str. 168. Since *B. subtilis* BSN238 is not yet present in the NCBI Taxonomy nor NCBI Protein database and only one protein is modified, we expect *B. subtilis* subsp. *subtilis* str. 168 to be selected as final strain candidate. Bacillus samples are examined twice, once complete with 28902 spectra (bacillus all) and once randomly reduced to 1000 spectra (bacillus 1k) to improve performance with respect to the vast bacterial search space. A detailed description of sample acquisition and the search parameters for all samples is provided in Appendix A.3.

For the viral samples we used all viral NCBI RefSeq proteins (via Entrez on July 10, 2017) in the iterative workflow and additionally all viral non-RefSeq proteins (via Entrez on July 11, 2017) for the uniques- and Pipasic-based strategies. Viral spectra were filtered beforehand using corresponding host proteomes (all isoforms) including the UniProt *Homo sapiens* reference proteome (UP000005640, May 23, 2017) for the cowpox sample and *Gallus gallus Red jungle fowl* reference proteome (UP000000539, May 16, 2017) for the bronchitis sample.

For the iterative examination of the bacillus samples, all bacterial NCBI RefSeq proteins were downloaded via FTP (`ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/`, release 82). Since downloading all NCBI Protein entries for bacteria (taxid 2) is impractical and requires too much time using the Entrez API, we utilized the NCBI Blast NR database as most comprehensive, common and readily available protein resource. This database can be obtained in fasta format via FTP (`ftp.ncbi.nlm.nih.gov/blast/db/FASTA/`, release 19.07.2017) and the bacteria subset was extracted with in-house scripts to create the databases for the uniques- and Pipasic-based strategies.

4.3. Results

The final selections of the top taxa candidates for all samples and all three compared identification strategies are summarized in Table 4.1. For the cowpox sample, identification results agree the most. TaxIt (Figure 4.2) and the Pipasic strategy (Appendix A.3 - Figure A.12) are both able to identify the expected *Cowpox virus* (Brighton Red) strain. However, unique PSMs are limited to the parent *Cowpox virus* species and not available at strain level, thereby giving place to an incorrect identification of *Bat astrovirus* Hil GX bszt12 (Appendix A.3 - Figure A.12).

In comparison, bronchitis and bacillus samples feature notable variability in proposed taxa candidates. For bronchitis, TaxIt is able to identify the expected *Avian*

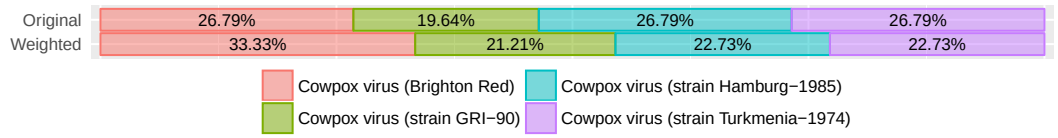

Figure 4.2.: Relative counts of TaxIt cowpox sample analysis. Relative counts are illustrated as result of the TaxIt workflow. Original and weighted relative counts are summarized by means of one vertical stacked bar each. Candidate strains are labeled and color-coded and ratios highlighted as percentages within bars. The original counts do not allow to distinguish the expected strain from competing candidates. In contrast, the weight-based correction method implemented in TaxIt resolves the present tie and emphasizes the correct strain, Brighton Red.

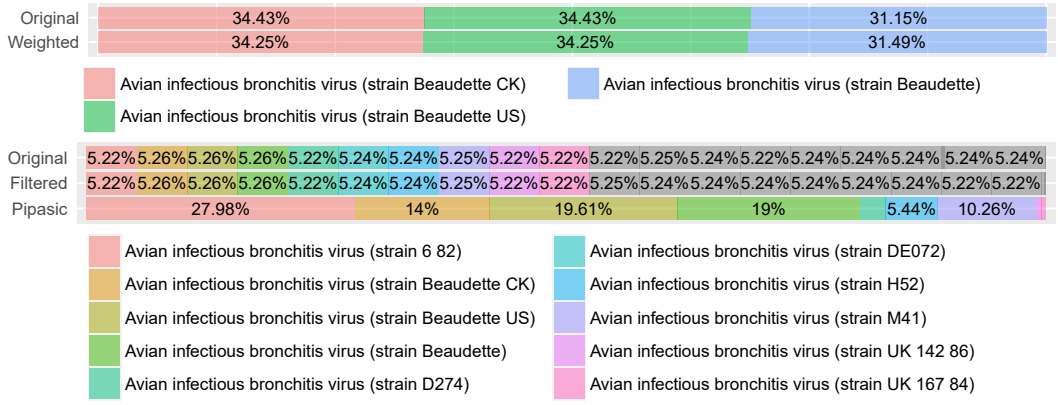

Figure 4.3.: Relative counts of bronchitis sample analysis. Relative counts are illustrated as result of the TaxIt (top) or Pipasic approach (bottom), respectively. TaxIt's original and weighted relative counts as well as Pipasic's original, filtered (a minimum of two hits and only the most 100 abundant taxa) and corrected relative counts are summarized by means of one vertical stacked bar each. Candidate strains are labeled and color-coded and ratios highlighted as percentages within bars. Despite applying a small correction based on weighting, TaxIt is not able to fully resolve the expected strain Beaudette CK but features a highly constrained selection of candidates in the first place. In contrast, the Pipasic-based strategy results in considerably more initial candidates and eventually promotes the incorrect strain 6/82.

infectious bronchitis virus (strain Beaudette CK) strain. However, final candidate selection is not limited to one strain but additionally includes the closely related *Avian infectious bronchitis virus* (strain Beaudette US) (Figure 3). Pipasic on the other hand supports the incorrect *Avian infectious bronchitis virus* (strain 6/82) strain (Figure 4.3). No unique PSMs are available for the bronchitis sample, rendering the identification of a strain impossible with this strategy.

Using TaxIt, the expected *B. subtilis* str. 168 strain is identified correctly in the reduced as well as in the complete sample. Pipasic consistently rejects the correct strain and species in favor of *Bacillus cereus* strains such as *Bacillus cereus* SJ1 and *Bacillus cereus* B4264. The uniques-based strategy is able to include the true strain *B. subtilis* str. 168 into the final candidate list of the bacillus 1k sample. However, it fails to separate the strain from several distinct species due to equal amounts of uniques as illustrated in Appendix A.3 - Figure A.14. Even more, analysing the complete bacillus sample results in the species *Paenisporosarcina quisquiliarum* being predominantly present in terms of uniquely assigned PSMs.

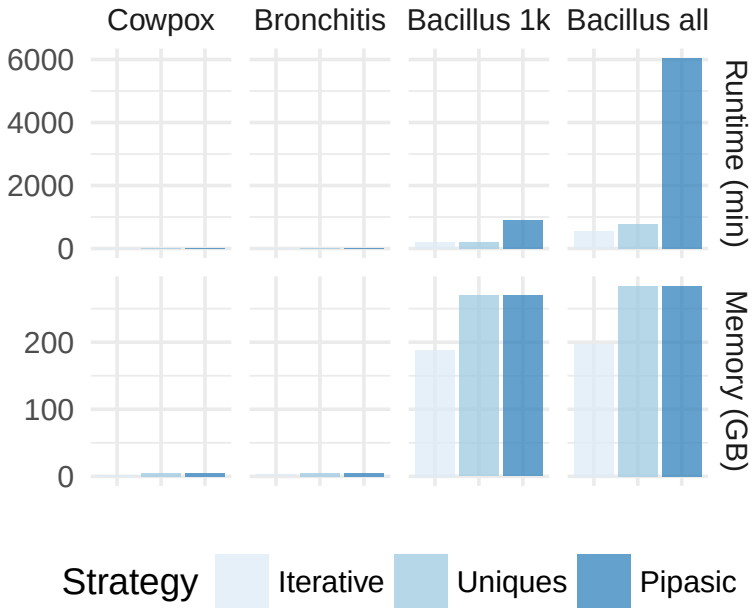

Figure 4.4.: Runtime and memory benchmarks. Total runtime in minutes and maximum memory usage in gigabyte is represented per sample and color-coded identification strategy.

Analysis was performed with X!Tandem as database search engine and limited to

Table 4.1.: Expected and identified taxa per sample and strategy

		Species/Strain	NCBI taxid
Cowpox	**Expected**	*Cowpox virus* (Brighton Red)	265872
	TaxIt	*Cowpox virus* (Brighton Red)	265872
	Uniques	*Bat astrovirus* Hil GX bszt12	1748291
	Pipasic	*Cowpox virus* (Brighton Red)	265872
Bronchitis	**Expected**	*Avian infectious bronchitis virus* (strain Beaudette CK)	160235
	TaxIt	*Avian infectious bronchitis virus* (strain Beaudette CK)	160235
		Avian infectious bronchitis virus (strain Beaudette US)	
	Uniques	no unique PSMs available	
	Pipasic	*Avian infectious bronchitis virus* (strain 6/82)	11121
Bacillus 1k	**Expected**	*Bacillus subtilis* subsp. *subtilis* str. 168	224308
	TaxIt	*Bacillus subtilis* subsp. *subtilis* str. 168	224308
	Uniques	*Acidobacteria bacterium* RIFCSPLOWO2_12_FULL_59_11	1797187
		Anaerobacillus alkalilacustris	393763
		Bacillus lentus	1467
		Bacillus niacin	86668
		Bacillus sp. Marseille-P2384	1805475
		Bacillus subtilis subsp. *subtilis* str. 168	224308
		Bacteroides luti	1297750
		Candidatus Glassbacteria bacterium RIFCSPLOWO2_12_FULL_58_11	1817867
		cyanobacterium TDX16	1503470
		Sporolactobacillus laevolacticus	33018
		Streptomyces griseus	1911
	Pipasic	*Bacillus cereus* SJ1	699184
Bacillus all	**Expected**	*Bacillus subtilis* subsp. *subtilis* str. 168	224308
	TaxIt	*Bacillus subtilis* subsp. *subtilis* str. 168	224308
	Uniques	*Paenisporosarcina quisquiliarum*	365346
	Pipasic	*Bacillus cereus* B4264	405532

Table 4.2.: Runtime and maximal memory consumption (resident set size) per identification strategy and sample

	Runtime (h:m:s)			Memory (max RSS MB)		
	TaxIt	**Uniques**	**Pipasic**	**TaxIt**	**Uniques**	**Pipasic**
Cowpox	0:09:53	0:09:59	0:12:16	1638.28	4949.69	4951.09
Bronchitis	0:12:20	0:09:41	0:16:45	2946.19	4579.61	4579.84
Bacillus 1k	3:15:59	3:12:41	15:01:04	193357.94	276134.38	276134.28
Bacillus all	9:29:06	13:00:55	125:53:07	202574.46	290719.44	290487.57

24 threads on a server with Debian GNU/Linux 8.9 (jessie), 64 cores (128 threads) of type Intel(R) Xeon(R) CPU E5-4667 v4 @ 2.20 GHz, 512 GB of RAM and SSD storage. Runtime and memory consumption for all samples and strategies are illustrated in Table 4.2 and Figure 4.4. Applying the iterative approach reduces memory usage down to one third for viral strain identification and two third for bacterial strain identification. While Pipasic's runtime is comparably high, TaxIt shows no substantial change in runtime in comparison to the uniques-based strategy for small databases such as the collective viral sequences or small sample sizes. However, analyzing the full bacillus sample reveals a gain in runtime when utilizing NCBI RefSeq proteins plus selected strain proteomes instead of the extensive NCBI Blast NR database.

4.4. Discussion

In summary, the TaxIt approach is able to unambiguously identify organisms of all samples down to a low taxonomic level, with the minor exception of a tie for the bronchitis sample. However, the uniques-based identification strategy is repeatedly deficient in strain level PSMs or features highly ambivalent results. Furthermore, Pipasic frequently favors an incorrect strain or even incorrect species. In general, for some samples correct strains are observed as a top candidate even in the original counts, independently of database choice and prior to count adjustments. Nevertheless, all samples benefit from either the iterative and focused database usage, the count adjustment procedure or reduced resource consumption while the outcome remains legitimate.

TaxIt improves on the count based ranking independently of count adjustments by limiting candidates to strains of one species early on. For instance, the search of bronchitis samples against the NCBI Blast NR bacteria subspace results in a rather uniform distribution of original counts for *Infectious bronchitis virus* strains with the strains Beaudette, Beaudette CK and Beaudette UK being only slightly increased in comparisons to other strains (Figure 4.3). In comparison, the iterative approach re-

sults in less strain candidates of the same species focusing solely on Beaudette strains when selectively searching against *Avian coronavirus* strains in the secondary iteration (Figure 4.3). This is partly a consequence of how and whether parental or multispecies proteins are associated with strains or taxa in general within the NCBI database. In case of the iterative approach, fewer mutual species or genus proteins are consulted in the strain identification iteration. However, the extend of this effect varies between samples or taxonomic kingdoms, respectively. For instance, bacterial strain proteomes such as the *Bacillus subtilis* strains feature numerous directly assigned mutual protein sequences and thus result in an extended range of candidate strains of the same species (Appendix A.3 - Figure A.14-A.15). Nevertheless, the restriction to a specific set of strain proteomes prevents manifold primary misassignments to distant species, genera or even phyla as can be observed for Pipasic- and uniques-based original counts, respectively, which cannot be sufficiently resolved after correction (Appendix A.3 - Figure A.14-A.15). In general, the iterative and selective database usage ensures that final strain selection is limited to strain candidates of an appropriate species, thus prevents false positive hits on distinct strains of other taxa including species, genera and phyla and allows for a more confident final strain candidate selection.

Furthermore, uniform distributions and even consensus in original counts of strain candidates demonstrate the need and benefit of count adjustment methods. The implemented weighting procedure is able to resolve ties between strains such as in the TaxIt cowpox sample analysis (Figure 4.2) or at least amplifies the correct strain and increases the distance to competing candidates.

TaxIt is able to infer exactly one strain for the presented samples each with the exception of the bronchitis sample where the final differentiation between the strains Beaudette CK and Beaudette UK fails. We observed that the corresponding PSMs are fully shared between the two proteomes. Although different proteins are available for each strain in general, peptide hits are either assigned to shared proteins of the parent strain *Infectious bronchitis virus* or to homologous proteins which differ only in identifier but not in sequence. Though a more granular taxonomic relation cannot be ascertained from the NCBI taxonomy, we expect the Beaudette strains to feature a considerably closer relationship as compared to other *Infectious bronchitis virus* strains. Therefore, we consider the draw between Beaudette strains as sufficiently appropriate strain identification.

As for the uniques-based strategy, we observed a poor availability of uniques PSMs on strain level. While the exploitation of purely unique features is a common theme for species level identification, the low amount of unique PSMs in strains is insufficient for strain level inference. However, in general the frequency of spectra matching to distinct proteins and proteomes remains a valuable parameter for strain differentiation when considering and weighing both, unique and the plethora of non-unique matches.

TaxIt has a comparable runtime for small samples and databases (such as the viral data) despite utilizing a constrained search space. This is primarily a result of the additional strain proteome downloads since the NCBI Entrez API is not designed and optimized for large scale downloads and proteins need to be fetched in numerous iterations of small chunks. However, the download overhead fades into the background when considering full bacterial samples such as bacillus all (Table 4.2 and Figure 4.4) and gives place to a runtime improvement of three quarters when compared to NCBI Blast NR database searches. In contrast, Pipasic's runtime is afflicted with additional sequence comparisons necessary for constructing the similarity matrix which, in addition, is highly influenced by increasing numbers of PSMs and taxa to compare. Finally, the memory footprint of TaxIt in comparison to the uniques- and Pipasic-based strategies remains constantly less for all samples, as would be expected when utilizing substantially less proteins in the search databases.

4.4.1. Conclusion

Untargeted strain level identification via tandem MS spectra is a challenging task with respect to the excessive quantity of strains which need to be considered competitively. To this end, we present an iterative approach focusing on species identification first and thus limiting strain identification to concise selected target databases. Both iterations take advantage of publicly available data from the NCBI Taxonomy and Protein databases. In general, strain level identification performance is limited by the availability or integrity of taxa and proteomes in these databases. However, constantly increasing quality and quantity of the NCBI Taxonomy and Protein databases will induce constant improvement of strain level identification strategies such as the presented iterative workflow.

TaxIt is available for download under open-source license at `https://gitlab.com/rki_bioinformatics`.

5. Summary and Conclusion

Proteogenomics (among others) illustrates the importance of sequence quality and availability in the "life cycle" of omics sequence data. From *de novo* assembled draft genomes, through 6-frame translated protein sequences and protein databases in general to eventual application in gene annotation or peptide identification, each step benefits from improved quality and evaluation in a successive manner. In this thesis, we provide methods that focus on sequence and database quality and comprehensiveness. We contribute to initial acquisition of quality sequences and evaluation of resulting or available sequence databases and finally take full advantage of these sequences database in terms of comprehensiveness and potential taxonomic range and depth. Thereby, we support applications in (meta-)genomics, (meta-)proteomics and eventually proteogenomics. While all methods were designed for universal application in genomic or proteomic sequence data analysis respectively, they possess special potential for the analysis of non-model organisms.

The quality of genome sequences is vital for countless applications in (meta-) genomics, (meta-)proteomics and proteogenomics, among others. Quality control for *de novo* assembled sequences is particularly challenging since reference-based ground truth is not available. Nevertheless, *de novo* assembly is an essential instrument, especially for the study of unsequenced organisms. In Chapter 2, we presented SuRankCo as a novel approach for *de novo* assembly contig quality assessment based on machine learning. We observed a high dependency of the prediction performance on the similarity of training and test data with respect to sequencing, organism and assembly parameters and characteristics. However, as demonstrated this results in promising performance for metagenomics samples. In particular, the training can be based on already known and sequenced organisms to enable the scoring and thus quality control of the remaining unsequenced organisms in the same sample. This procedure takes advantage of the uniformity of characteristics within one sample. Furthermore, the metagenomic analysis illustrated the special benefit for non-model organisms as sequence quality knowledge may be transferred to several unsequenced organisms of one sample at once. Thereby, SuRankCo supports the preparation of quality controlled draft genomes of non-model organisms (among others) for potential applications in, for instance, proteogenomics or general contribution to genome and thus eventually protein sequence databases.

The suitability of protein sequence databases for tandem MS spectra identification with respect to taxonomic depth is influenced by two main factors: The

number of related organisms and their similarity among each other. Both determine whether a suitable close related proteome is available for detailed taxonomic classification. Thereby, the knowledge of the database' suitability can support the choice of methodology applied. Non-model organisms may especially benefit from extended search strategies if their respective neighbourhood in the target database is sparse. With LiDSiM we presented in Chapter 3 a method to estimate the suitability for the database of interest. Thereby, the impact of alternative search strategies such as error-tolerant searches as well as of database extension strategies are taken into consideration. In particular, we illustrated the improvements in taxonomic classification of *D. deserti* strain VCD115 by extended search strategies and thereby demonstrated the application and benefit for non-model organisms.

While evaluations, as provided by LiDSiM, reveal potential limits of protein databases in terms of taxonomic classification, the accumulation and collective application of more and more protein sequences may improve the range and depth for tandem MS-based taxon identification. Thereby, public resources such as the NCBI Protein database provide the most comprehensive collection readily available. However, the negative impact of database size has been repeatedly reported and therefore needs to be taken into consideration. In Chapter 4 we presented TaxIt for untargeted strain level identification of tandem MS spectra. By applying an iterative procedure, the method can take advantage of the full taxonomic potential offered by the NCBI Protein database while still limiting the number of proteomes actually used. Strain level identification was successfully demonstrated for viral as well as bacterial samples.

Although methods such as TaxIt are able to approach comprehensive search spaces such as provided by the NCBI Protein database and enable the identification of a large group of microbial organisms, they are nevertheless limited by proteome availability - in particular with respect to non-model organisms. Therefore, in-depth applications such as strain level identification benefit substantially from improvements at various levels. Including for instance increased draft genome quality or reformed sequence database evaluation, which will eventually result in yet more comprehensive resources. On the one hand, the underrepresentation of non-model organisms illustrates the need of constant sequence acquisition and evaluation. On the other hand, it emphasizes the apparent positive synergy effects of methods from initially distinct areas of omics research. Advances in genomic and proteomic technology and methodology will reveal more and more potential links and overlaps, jointly fill existing gaps and holes in sequence databases step by step and therefore push the limits of omics data analysis on and on.

Future research While the presented methods in this thesis contribute to the overall improvement of omics data processing and utilization with respect to quality,

availability and level of detail, many problems and challenges remain and admit of improvement. In particular, advancements and changes in omics technologies concurrently allow and demand for continuous adaption and fundamental redevelopment of appropriate methodology.

The contig ranking in *de novo* assemblies provided by SuRankCo is highly depending on preceding technology used. Consequently, changes in sequencing, sequence characteristics and error-profiles as well as assembly procedures will affect the ranking performance and demand the adaption and complementation of appropriate features and scores. Since the prediction accuracy is amendable in general, an adaption with respect to more recent advances in machine learning may improve overall results. For instance, deep learning has been frequently and successfully applied on different omics sequence data (Angermueller et al., 2016). Furthermore, to improve on general usability of ranked contigs clear cutoffs should be identified to completely separate supposedly superior contigs from poor quality contigs. Here, possible approaches range from mixture model fitting to the incorporation of marker contigs to evaluate their placement in the ranked contig set similar to the target-decoy approach for tandem MS database searches (Nesvizhskii, 2010; Jeong et al., 2012).

The simulated evaluation of protein databases in terms of integrity and utility for taxonomic classification possess one major disadvantage. The simulation as currently implemented in LiDSiM repeatedly removes a proteome which would have been available in actual applications. Thereby, the procedure overestimates the limit of taxonomic classification of the assessed database. While this has minor effects on the estimations for highly represented organisms such as *E. coli*, it could underrate the potential of taxa which feature a very sparse neighborhood in the target database including non-model organisms. Therefore, further progression of the method should include a non-removal based estimation. One possibility would be the simulation of close related artificial surrogate proteomes, for instance by altering several proteins using probabilistic profile hidden Markov models as made possible by the HMMER software (Eddy, 1998; HMMER). Furthermore, the simulation is a deterministic procedure based on efficient string matching algorithms and therefore doesn't account for factors which influence real spectral analysis such as variations and errors in peak mass and intensity. An approach based on simulated spectra for instance based on MSSimulator (Bielow et al., 2011) in combination with actual database search engines such as MS-GF+ would be possible in general, but is impractical due to high amount of iterations when considering several or all organisms of comprehensive databases.

As a pipeline relying on several substeps and modules, TaxIt is highly qualified for continues improvement and development. Taxonomic range or resolution of strains might be improved by relying on additional sequence resources. For instance, the NCBI Genome database in combination with 6-frame translation or ORF predic-

tion would be a valuable and comprehensive extension. Local specialized custom sequence databases could extend public resources but would require proper custom taxonomic mappings to enable adequate species and strain proteome extraction and inference. In general, the integration of any additional strain databases would be possible as long as compatibility with the NCBI taxonomy is ensured and, in case of public databases, a proper interface for automated access is available. Identification performance could be improved by utilizing the consensus of different search engines (McHugh and Arthur, 2008) as demonstrated by MSblender (Kwon et al., 2011) or the PeptideShaker platform (Vaudel et al., 2015) but should be carefully considered with respect to required identification speed. Furthermore, strain inference confidence could be supported by additional metrics such as unified e-values (Alves et al., 2016) in comparison to sole spectral count rankings. Improvements on runtime could be achieved by introducing novel high-speed search engines such as MSFragger (Kong et al., 2017), spectral clustering and peak filtering methods such as MaRaCluster (The and Käll, 2016) and MS-REDUCE (Awan and Saeed, 2016), respectively, or by dismissing protein decoys (which represent half of the database to be searched) in favor of mixture-model approaches such as PeptideProphet (Ma et al., 2012). Additionally, early on strain identification may be achieved by step-by-step analysis of spectra subsamples with increasing confidence over time. Finally, with increasing interest in environmental samples the extension of TaxIt with metaproteomic profiling capabilities on strain level is an encouraging objective. Here, the adaption of metagenomic strain discrimination methods as demonstrated by Pipasic is a promising approach and should be continued by taking latest advances in metagenomic abundance correction such as DiTASiC (Fischer et al., 2017) into account.

Although all three presented methods were independently developed and not actually linked yet, they possess the potential for integrative application as for instance in a proteogenomic setting. This integrative concept is hypothetical so far and has yet to be implemented and evaluated. However, a general scenario of joint and consecutive application may look like this: An unsequenced microbial non-model strain sample is analysed via NGS sequencing, a draft genome is created by *de novo* assembly with assistance of SuRankCo. Peptide sequences for tandem MS analysis could be created either by 6-frame translation for immediate proteogenomic applications or, in the long term, by submitting the draft genome to a public genome database and relying on automated annotation and integration processes of public protein database as provided by NCBI or UniProt. The overall effect on taxonomic classification for related taxa by the latest addition of the non-model organism can then be examined with LiDSiM by comparing databases with and without the novel peptide sequences. Finally, TaxIt can take advantage of the additional strain data in public databases (if adequately assigned in the NCBI Taxonomy) and potentially improve on prospective untargeted strain identifications. Overall performance of this scenario may be validated by simultaneous tandem MS analysis of the same

microbial sample followed by TaxIt identification under varying conditions including an unmodified database as well as database extended by the draft genome, once with and once without quality control.

A. Appendix

A.1. Additional Material for Chapter 2

A.1.1. Contig Features

The features applied by SuRankCo are listed in Table A.1 with name, description and variants.

Some features are based on single values and thus have no variants. Other features consisting of several values per contig (such as coverage, read count etc.) are summarized by six common variants to describe their statistical distribution. This includes the mean, the standard deviation (sd), the median, the median absolute deviation (mad), the minimum and the maximum.

Further variants are noted in the description as *additional variants*, if appropriate.

Table A.1.: Contig Features

Name	Description	Variants
Length	Unpadded length, i.e. length of the final contig.	
Base Count	Padded length, i.e. length of the consensus of a read alignment, including potential gaps introduced by inserts in reads which are not consistent with the majority.	
Base Segment Count	Number of continuous gapless segments in the padded contig.	
Read Count	Number of reads contributing to the contig.	
Read Complement Fraction	The fraction of complement reads in the total number of reads contributing to the contig.	
N50 Relation	Relation between contig length and N50.	

Estimated Genome Size	The expected genome size (EGS), either indicated by a parameter or estimated as the sum of contig lengths.	
Genome Relation	Relation between the contig length and the estimated genome size.	
Contig Qualities	Pooled base-wise quality values of the contig as provided from assemblers (e.g. in Ace files).	✓
Read Qualities	Pooled base-wise quality values of the reads as provided from base-callers (e.g. in Sff files).	✓
Read Length	Pooled original lengths of the reads.	✓
Read Length Padded	Pooled padded lengths, i.e. lengths of the reads in the alignment, including potential gaps introduced by deletions in other reads which are not consistent with the majority (compare Base Count).	✓
Read Length Quotient	Relation between original read length and padded read length (for all six variants).	
Read Length Clipped	Pooled lengths of clipped reads, i.e. the lengths of the padded read parts which are actually used and thus contribute to the contig (for instance, some read ends do not).	✓
GC-Content	The fraction of GC-content in the contig.	
Coverage	Pooled number of reads contributing to each position in the contig. *Additional variants*: Contig ends coverage is reported in addition, with end size equal to read length mean.	✓

Core Coverage	Pooled number of reads contributing to each position in the contig with the same nucleotide as the one selected for the consensus. *Additional variants*: Contig ends core coverage is reported in addition, with end size equal to read length mean.	✓
Base Confirmation	Significance of the core coverage in contrast to the coverage per position, tested with a binomial test with $k = core\ coverage$, $n = coverage$ and $p = 0.98$. $p = 1 - error\ rate$, where error rate denotes the average sequencing error. With an error rate of 2% the expectation of reads contributing the same correct nucleotide to each position is therefor 98%. *Additional variants*: Contig ends base confirmation is reported in addition, with end size equal to read length mean.	✓
Coverage Comparison	Coverage comparison within an assembly represented by the relation of the contig coverage to the mean coverage of all contigs in the assembly. *Additional variants*: Contig ends coverage comparisons are reported in addition, with end size equal to read length mean.	

Coverage Curve Drops	Coverage curve drops indicate local minima in the coverage of a contig with a value of less than 25% and 50% in contrast to their adjacent maxima within a fixed window size w. The coverage is preprocessed with a sliding window smoothing with window size w which is chosen as the mean read length of a contig. The number of drops is reported normalized by the contig length as well as the biggest drop, i.e. maximal difference between a minima and its smaller adjacent maxima.
K-mer Uniqueness Global	Number of K-mers unique in a contig in contrast to other contigs within the assembly normalized by the contig length (since longer contigs comprise more unique K-mers by chance). K-mers are extracted with a size of 8, and only K-mers containing standard nucleotide symbols (i.e. A,C,G and T) are considered.
K-mer Uniqueness Ends	Number of K-mers unique in a contig end in contrast to both ends of other contigs within the assembly. Reported as minimal and maximal K-mer uniqueness to avoid implicit orientation of the contig. The read length mean of a contig is chosen as end size.

A.1.2. Contig Scores

Table A.2 provides an overview of the single contig scores calculated by SuRankCo. The scores are either based on match counts or error counts (edit distance, including mismatches and gaps) of contig-reference alignments.

Table A.2.: Contig Scores

Name	Description & Motivation
General Scores	Account for mismatches/errors in general as well as insertions and deletions to the contig (Normed Match Count 1 resp. Normed Match Count 2). These scores provide a basic penalization for small errors in general.
Normed Match Count 1	Number of alignment matches normalized by the contig length.
Normed Match Count 2	Number of alignment matches normalized by the alignment length.
Normed Error Count 1	The edit distance respectively error count normalized by the contig length.
Large Error Scores	Account for very large errors or unstable regions which might originate from mis-joins or badly sequenced/covered regions, resp. While small errors are only considered by the General Scores, critical large errors are additionally penalized hereby.
Max. Contiguous Error	The largest contiguous stretch of alignment errors normalized by the contig length.
Max. Region Error	The largest number of alignment errors in a fixed region size (100 bp).
End Scores	Similar to Large Error Scores but applied to contig ends only. Errors in this region are rather critical for subsequent applications (e.g. for scaffolding) are therefore additionally penalized.
Max. End Error Stretch	Largest stretch of errors right at the contig ends (unfixed length) normalized by the contig length.
Max. End Error Count	The largest number of alignment errors in the ends (fixed length of 100 bp).
Other Scores	Additionally account for insertions and critical mis-joins.
Normed Contig Length	Relation of contig length to alignment length.

A.1.3. Training Class Definitions

Each contig score applied by SuRankCo is separated into two classes to allow for binary classification. The separation into the two classes can be either set manually or automatically by fitting exponential distributions. A threshold selection may be supported by histograms provided by the SuRankCo-Score module (as shown in Figure A.2).

The automatic exponential fitting makes use of the MASS R package (Venables and Ripley, 2007). It fits an exponential distribution to each single score distribution of the training contigs. Finally, a certain quantile of a fit is considered as the threshold for the score class separation. The selection of a quantile should be based on the concrete training data and be adjusted accordingly. However, 25% yielded a good separation for the *E. coli* contigs. See Figure A.2 for examples.

A.1.4. Experiment Preparation

A.1.5. E. Coli

To demonstrate the usage of SuRankCo, we chose four next-generation sequencing experiments available in the NCBI Sequence Read Archive (SRA) (Wheeler et al., 2008). The reads were all sequenced with an Illumina Genome Analyzer II. Additional properties are listed in Table A.3.

Table A.3.: SRA Experiments

Experiment ID	# of Spots	# of Bases
SRR400617	14,299,251	514.8M
SRR400618	13,539,459	487.4M
SRR400619	16,720,568	601.9M
SRR400620	16,359,717	588.9M

The reads were assembled using Mira (Chevreux et al., 1999) with basic settings. A sample configuration for one experiment is provided in listing A.1. Finally, the four resulting assemblies were randomly divided into three training assemblies (SRR400617, SRR400618 and SRR400619) and one test assembly (SRR400620).

Listing A.1: Mira Configuration Manuscript

```
project = SRR400617
job = genome,denovo,accurate

readgroup = IlluminaReads
data = SRR400617.fastq
technology = solexa
```

```
parameters = --GE:not=10 --NW:cmrnl=warn --OUT:ora=on
```

A.1.6. ALE

To compare the SuRankCo results of the E. coli experiment to ALE, the reads of the prediction data (SRR400620) were mapped against the corresponding contigs using Bowtie2 with default settings. Thereby, ambiguous reads where assigned according to the best alignment. The resulting sam file were sorted and, together with the contigs provided to ALE.

Since ALE does not provide a score per contig, ALE sub-scores were transformed to error counts and summed up per contig. For each ALE sub score, a histogram over all contigs and positions was created to manually choose a threshold (as shown in Figure A.1). Each contig position below the threshold of a sub-score is counted as a potential error. The counts were summed for each contigs and normalized by the contigs length and the total number of sub-scores.

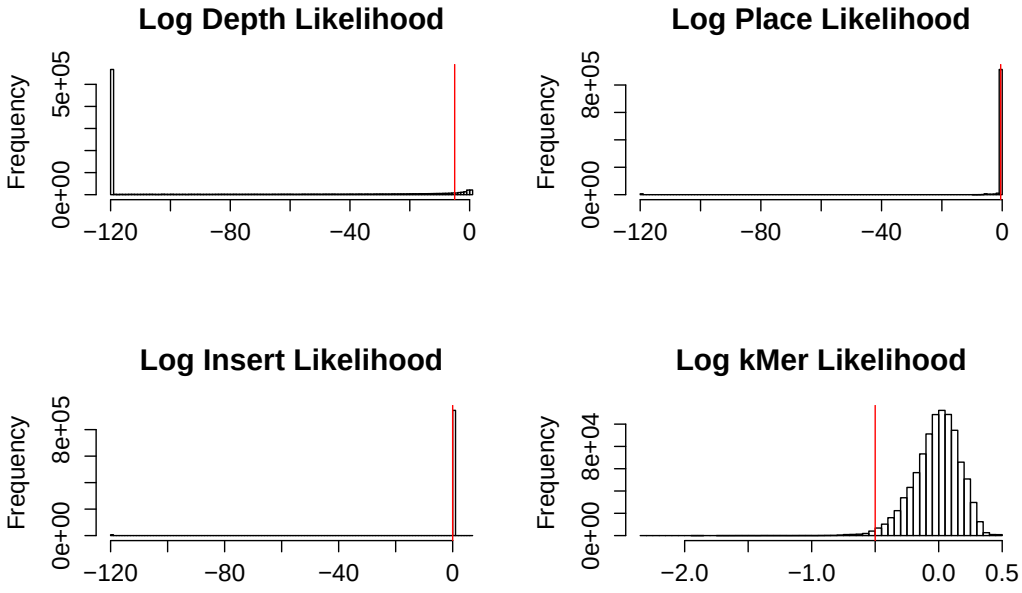

Figure A.1.: Histogram of ALE sub-scores. For each sub-score, histogram were constructed over all contigs and position. Thresholds were set manually to -5, -0.5, 0 and -0.5 for the ALE sub-scores Log Depth Likelihood, Log Place Likelihood, Log Insert Likelihood and Log kMer Likelihood, respectivly.

A.1.7. Mock Community

To demonstrate the usage of SuRankCo in conjunction with several organisms and assemblers, we make use of the staggered mock community of the Human Microbiome Project. The data set is available in the NCBI Sequence Read Archive [SRA:SRR172903] with a total number of 7,932,819 reads and 595M bases. Organisms in the mock community are represented in Table A.4 as well as the reference sequences used to classify reads and contigs and for evaluation. The mock data is evaluated in three different settings, a meta-assembly, single organism assemblies and a merged evaluation of single organism assemblies of different assemblers.

Mock Meta-Assembly

The meta-assembly of the mock community is constructed using MetaVelvet (Namiki et al., 2012) with kmer size 31 and no scaffolding. The resulting contigs are assigned to organisms by using Blast (Altschul et al., 1990) against all reference sequences and selecting the best hits according to the e-value. For training and prediction of SuRankCo scores, the organisms are randomly divided into two equal groups. The grouping and number of assigned contigs is depicted in Table A.4. The MetaVelvet output is converted to ace files using AMOS (Treangen et al., 2011).

Mock Single Assembly

For the single organism approach, the mock reads are mapped against all references and thereby assigned to organisms by using Bowtie2 (Langmead and Salzberg, 2012) with default settings. We selected all organisms with sufficient coverage for the following assemblies, including *E. coli* ($\sim$ 9x), *M. smithii* ($\sim$ 11x), *R. sphaeroides* ($\sim$ 30x), *S. aureus* ($\sim$ 38x), *S. epidermidis* ($\sim$ 35x) and *S. mutans* ($\sim$ 20x). This selection agrees with organisms comprising a suitable amount of contigs in the meta-assembly as shown in Table A.4. The reads for each organism are then assembled separately with Mira (Chevreux et al., 1999), Soap (Luo et al., 2012) and Velvet (Zerbino and Birney, 2008) with default settings, except in the following cases. For Soap and Velvet assemblies are constructed over a range of kmers from 1 to 75 and for each organism the assembly with the highest N50 is selected for further analysis. Since not all assemblers used here provide alignment information but contig sequences only, the corresponding reads are remapped to the assemblies by using Bowtie2 with default settings to produce sam files as input for SuRankCo. For training and prediction, organisms are assigned to the same groups as for the meta-assemblies.

Mock Single Assembly Merged

For the third evaluation, the single organism assemblies from Mira, Soap and Velvet are merged into combined datasets. Thus the training set and the prediction set consist each of assemblies of three organisms from three assemblers.

In general, SuRankCo is used with default settings for all mock experiments. However, contigs are filtered for a minimum size of 350 bases since commonly no valuable information such as genes are expected to be covered by shorter sequences.

A.1.8. GAGE Study

To further demonstrate the usage of SuRankCo in conjunction with several assemblers, we make use of the bacterial assemblies provided by the GAGE study. We evaluate all available assemblies of *Staphylococcus aureus* and *Rhodobacter sphaeroides* including ABySS, ABySS2, Allpaths-LG, Bambus2, MSR-CA, SGA, SOAPdenovo, Velvet. However, the CABOG assembly of *R. sphaeroides* could not be evaluated since there is no CABOG assembly of *S. aureus* available.

Since none of the GAGE assemblies provide alignment information but contig sequences only, the corresponding reads are remapped to the assemblies by using Bowtie2 with default settings to produce sam files as input for SuRankCo. For each assembly, we used either the original read set or a corrected read set in accordance with the GAGE supplementary material.

The GAGE bacteria assemblies are evaluated in two different settings including the evaluation of single assemblies and a merged evaluation of the assemblies of different assemblers. In both settings *S. aureus* has been used for training and *R. sphaeroides* for prediction. In the first setting, each assembly of *R. sphaeroides* is evaluated by training SuRankCo on the corresponding assemblies of *S. aureus* from the same assembler. For the second evaluation, the assemblies of *S. aureus* are merged into a combined training datasets. Then, each *R. sphaeroides* assembly is evaluated based on this merged training. SuRankCo is used with default settings for all GAGE experiments. However, contigs are filtered for a minimum size of 350 bases since commonly no valuable information such as genes are expected to be covered by shorter sequences.

A.1.9. Additional Result Figures

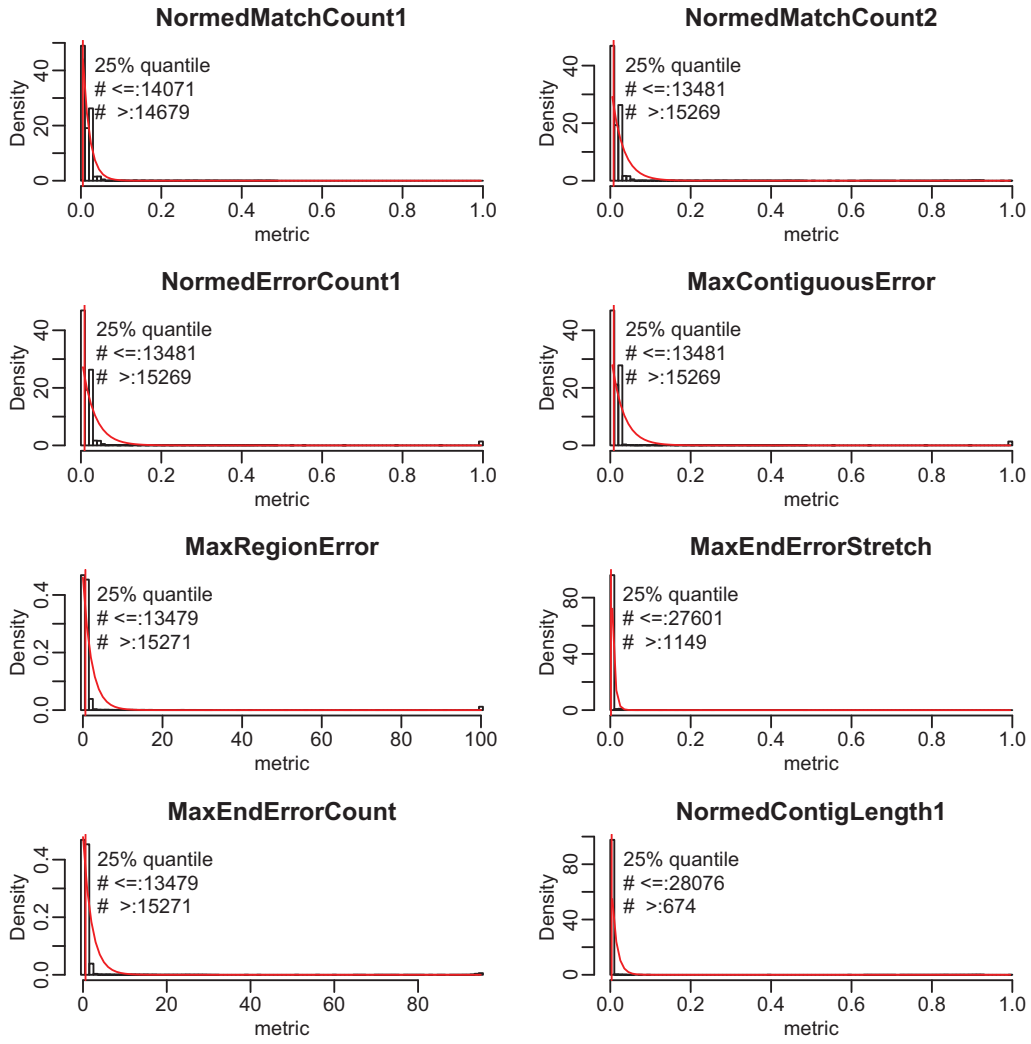

Figure A.2.: Histograms of single contig scores and exponential distribution fittings. For each single score, histograms (in black) are produced from the contigs of the training data. To further support the threshold selection, exponential distributions (in red) are fitted to each score. Finally, for the *E. coli* data 25% quantiles are selected as thresholds (vertical red lines). The numbers inside each plot indicate the amount of contigs below and above the threshold. The score ranges of Normed Match Count 1, Normed Match Count 2 and Normed Contig Length 1 are inverted to enable exponential fittings.

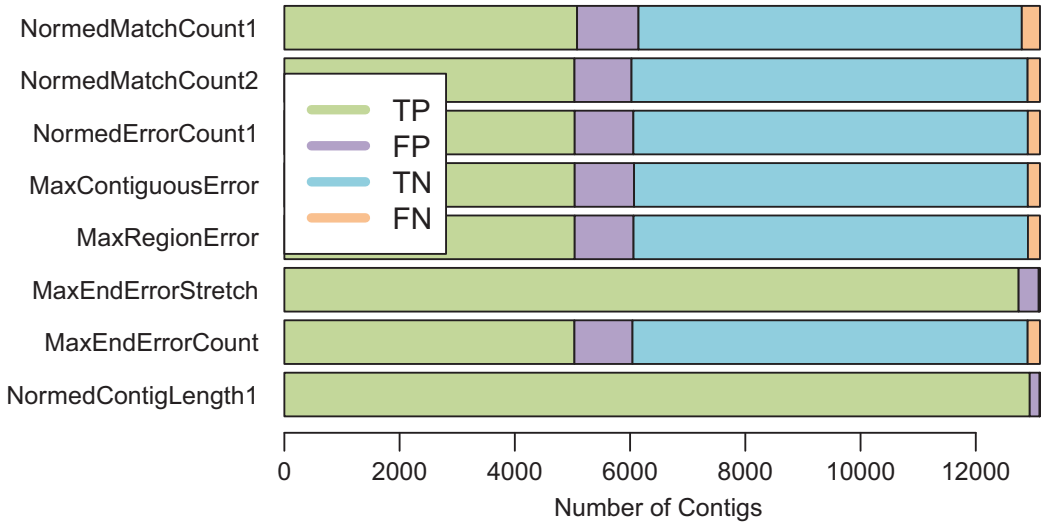

Figure A.3.: Classification metrics. For each score, the proportions of correct and incorrect classifications are indicated by true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Most scores show satisfying numbers of TPs and TNs. However, the scores Max End Error Stretch and Normed Contig Length 1 have mainly true TPs, some FP and almost no TNs and FNs. This is due to the very low variance in the corresponding distributions of the training contigs (see Figure A.2 for comparisons).

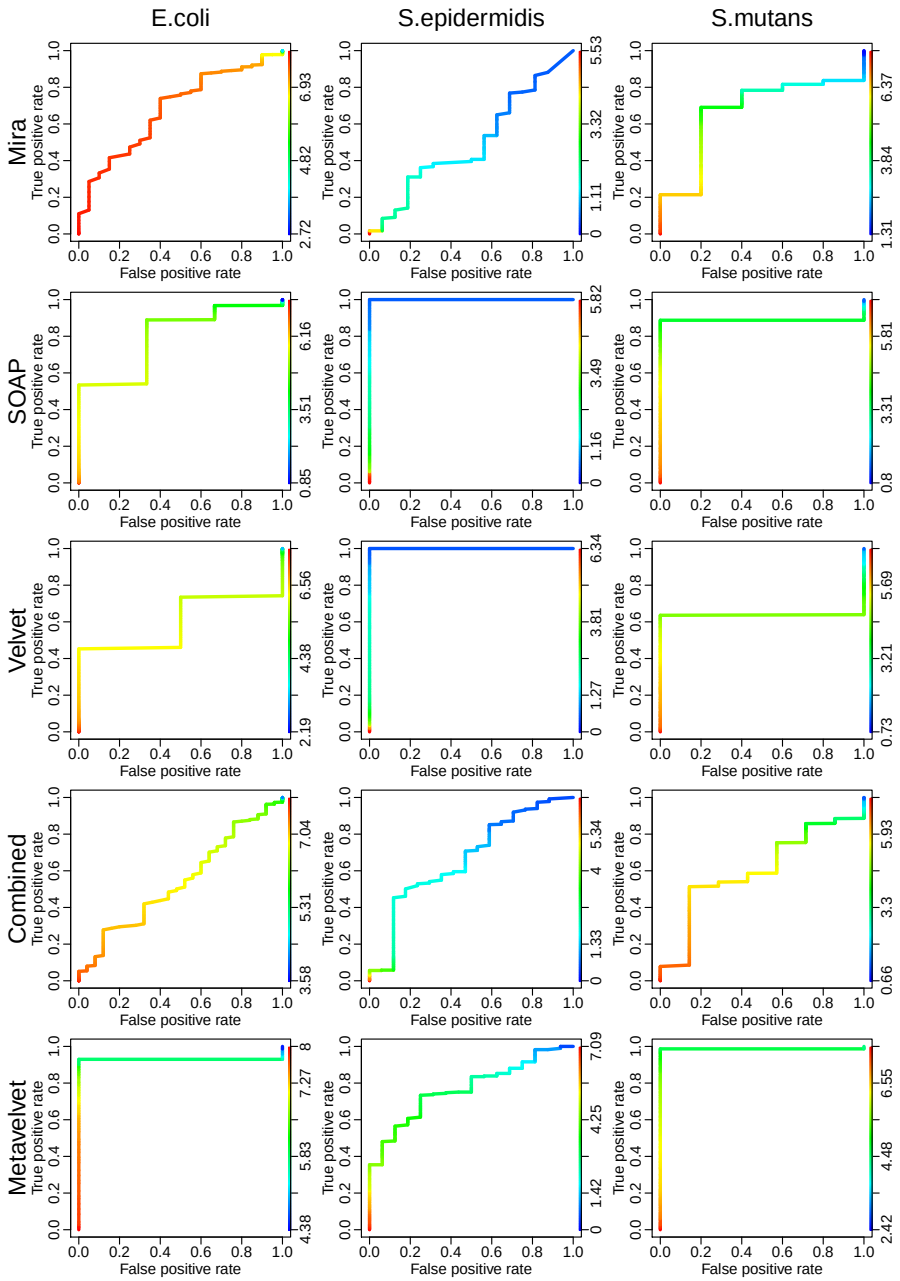

Figure A.4.: Evaluation of the SuRankCo predictions of major organisms in the mock community test data. The results of the mixed prediction sets are separatly illustrated for single organisms. Each plot comprises a ROC curve of the contig evaluation score grouping in contrast to a varying grouping of the SuRankCo scores. Thereby, the changing color of the graph represents the changing threshold for the SuRankCo score grouping.

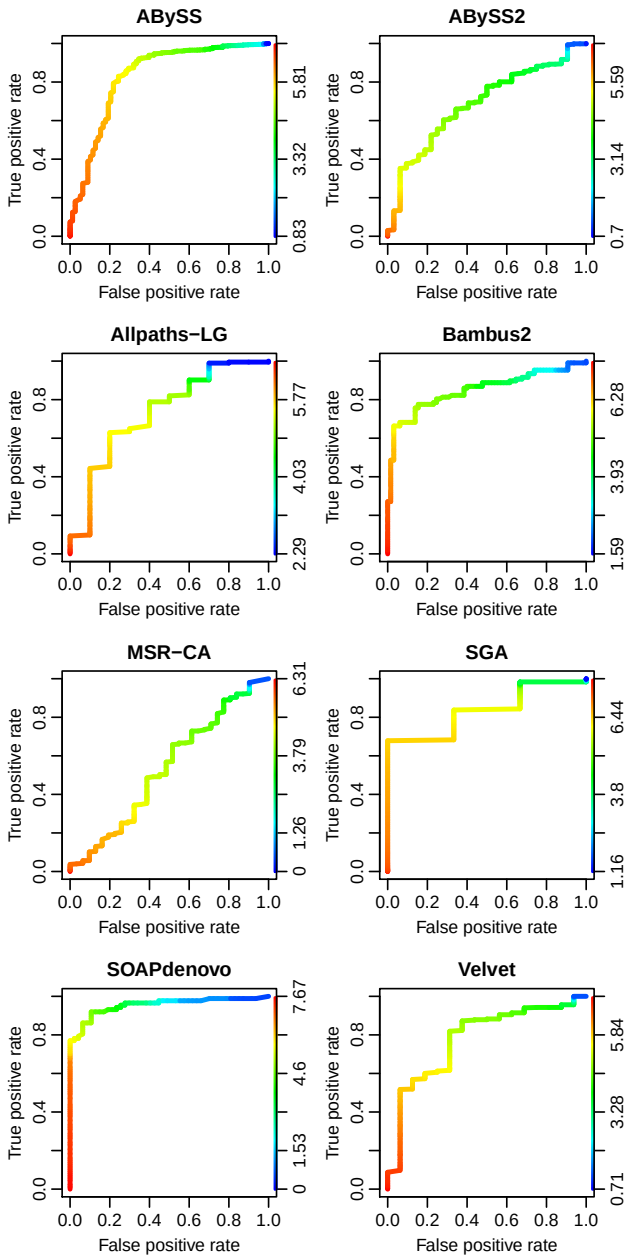

Figure A.5.: Evaluation of the SuRankCo predictions of the GAGE assemblies. Here, one ROC curve represents the evaluation of *R. sphaeroidis* assemblies classified by the single training dataset. Each plot comprises a ROC curve of the contig evaluation score grouping in contrast to a varying grouping of the SuRankCo scores. Thereby, the changing color of the graph represents the changing threshold for the SuRankCo score grouping.

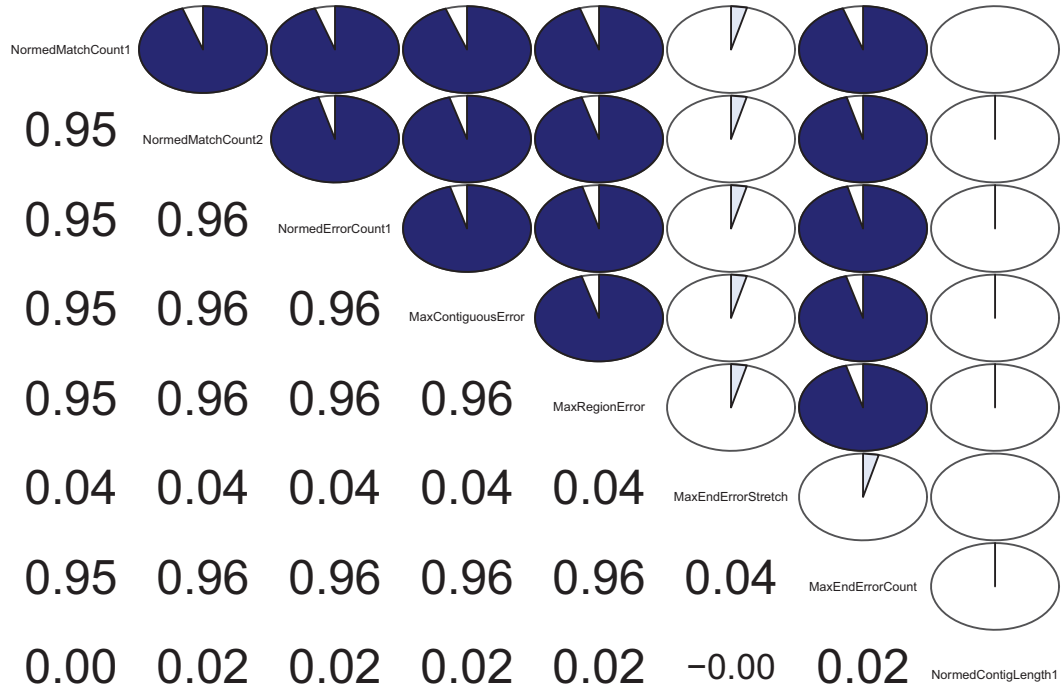

Figure A.6.: Correlation of single score predictions. The predictions of the single scores for the *E. coli* test data are highly correlated except for the Max End Error Stretch and Normed Contig Length 1. In general, contig score predictions may be less correlated since the contigs used in training may have a lower quality with higher variance in their alignments to the reference sequence.

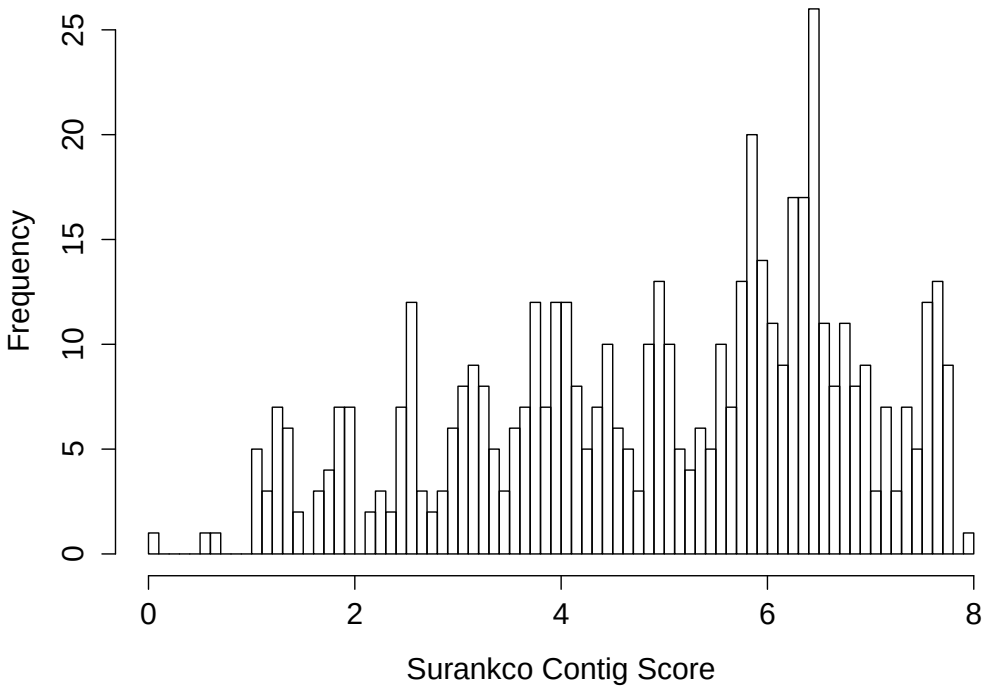

Figure A.7.: Example histogram of the final SuRankCo score. The histogram is constructed for the prediction set of meta-assembled mock community data. It shows a broad distribution of scores in contrast to the clustered scores of the *E. coli* experiment.

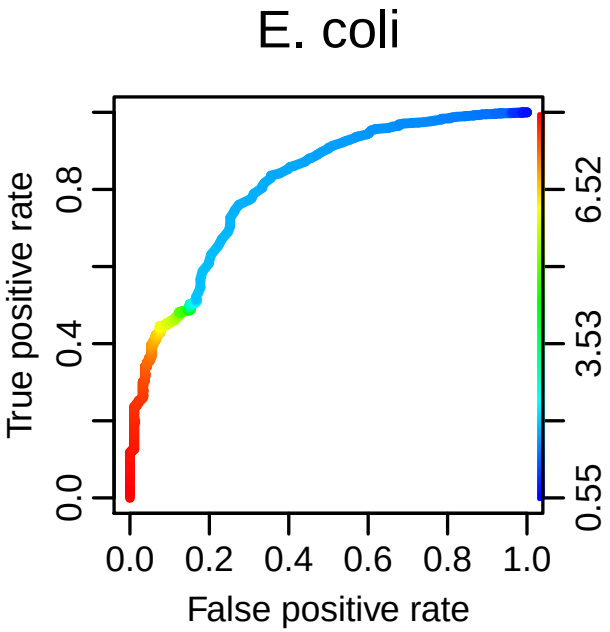

Figure A.8.: Evaluation of the SuRankCo predictions of the *E. coli* assembly. For completeness and comparability, this figure features the ROC curve-based evaluation of the E. coli experiment as applied for the mock and GAGE experiments. The ROC curve is based on the contig evaluation score grouping in contrast to a varying grouping of the SuRankCo scores. Thereby, the changing color of the graph represents the changing threshold for the SuRankCo score grouping.

A.1.10. Additional Result Tables

Table A.4.: Organisms of the mock community data, their grouping for the experiments (T and P), assigned contigs of the meta-assembly and the used references.

Organism	Group	N. Ctgs	Accession N.
Acinetobacter baumannii ATCC 17978	T	0	NC_009085.1, NC_009084.1, NC_009083.1
Actinomyces odontolyticus F0309	T	0	NZ_GG753644.1, NZ_GG753643.1, NZ_GG753642.1, NZ_GG753641.1, NZ_GG753640.1, NZ_GG753639.1
Candida albicans SC5314	T	0	NW_139421.1 - NW_139833.1 (413 sequences)
Enterococcus faecalis V583	T	0	NC_004668.1, NC_004671.1, NC_004670.1, NC_004669.1
Lactobacillus gasseri ATCC 33323	T	0	NC_008530.1
Methanobrevibacter smithii ATCC 35061	T	500	NC_009515.1
Propionibacterium acnes KPA171202	T	0	NC_006085.1
Rhodobacter sphaeroides 2.4.1	T	953	NC_007493.2, NC_007494.2, NC_007488.2, NC_007489.1, NC_007490.2, NC_009007.1, NC_009008.1
Staphylococcus aureus subsp. aureus N315	T	332	NC_002745.2, NC_003140.1
Streptococcus agalactiae 2603V-R	T	0	NC_004116.1
Streptococcus pneumoniae R6	T	0	NC_003098.1
Bacillus cereus ATCC 10987	P	0	NC_003909.8, NC_005707.1
Bacteroides vulgatus ATCC 8482	P	0	NC_009614.1
Clostridium beijerinckii NCIMB 8052	P	3	NC_009617.1
Deinococcus radiodurans R1	P	0	NC_001264.1, NC_001263.1, NC_000959.1, NC_000958.1
Escherichia coli str. K-12 substr. MG1655	P	72	NC_000913.3
Helicobacter pylori 26695	P	0	NC_000915.1
Listeria monocytogenes EGD-e	P	0	NC_003210.1
Neisseria meningitidis MC58	P	1	NC_003112.2
Pseudomonas aeruginosa PAO1	P	7	NC_002516.2

Staphylococcus epidermidis ATCC 12228	P	301	NC_004461.1, NC_005004.1, NC_005006.1, NC_005008.1	NC_005003.1, NC_005005.1, NC_005007.1,
Streptococcus mutans UA159	P	151	NC_004350.2	

Table A.5.: Spearman correlation between SuRankCo Contig Scores and BLAT metrics of the mock community meta-assembly prediction set.

	mismatch	**Qgapcount**	**Tgapcount**	**blockcount**
NormedMatchCount1	-0.8028354	-0.4541743	-0.4047294	-0.4354852
NormedMatchCount2	-0.7737683	-0.4533330	-0.5261784	-0.5462121
NormedErrorCount1	0.7736162	0.4536809	0.5270433	0.5469730
MaxContiguousError	0.7213388	0.4067623	0.4932727	0.5030182
MaxRegionError	0.7496859	0.4630718	0.5411223	0.5592339
MaxEndErrorStretch	0.1095816	0.1295148	0.1097870	0.1188257
MaxEndErrorCount	0.5545051	0.3505190	0.4133438	0.4269306
NormedContigLength1	-0.4434447	-0.7340355	-0.9966364	-0.9461041

Table A.6.: Spearman correlation between SuRankCo Scores and BLAT metrics of the *E. coli* experiment.

	mismatch	**Qgapcount**	**Tgapcount**	**blockcount**
NormedMatchCount1	-0.3865286	-0.0510903	0.1895618	0.1775827
NormedMatchCount2	-0.26312242	-0.05414976	-0.29729808	-0.29618498
NormedErrorCount1	0.26249300	0.05421344	0.29948468	0.29830996
MaxContiguousError	0.21822948	0.04766507	0.29486387	0.29224683
MaxRegionError	0.3624949	0.0614517	0.3318000	0.3309938
MaxEndErrorStretch	-0.33852832	0.03047685	-0.05786300	-0.05406256
MaxEndErrorCount	0.36263360	0.06136507	0.33169810	0.33089688
NormedContigLength1	0.19519882	-0.09248936	-0.99966452	-0.97335101

Table A.7.: Organism relationships based on reads, calculated by a sub-method of the GASiC tool (Lindner and Renard, 2013)

	E.coli	*M.smithii*	*R.sphaeroides*	*S.aureus*	*S.epidermidis*	*S.mutans*
E.coli	1	0	0.0003	0.000105	0.000556	0.000556
M.smithii	0	1	0	0	0	0
R.sphaeroides	0.000736	0	1	0.000065	0.00044	0.000651
S.aureus	0.0004	0	0.000155	1	0.042757	0.003227
S.epidermidis	0.00041	0	0.00015	0.028878	1	0.003173
S.mutans	0.000325	0	0.00014	0.000546	0.002463	1

A.2. Additional Material for Chapter 3

A.2.1. Spectral search parameter

Both *E. coli* and *D. deserti* spectra were analyzed with a tryptic search and fixed modification cysteine carbamidomethylation (+57.0513 Da). The *E. coli* spectra were searched with parent ion mass tolerance of 10 ppm whereas the *D. deserti* spectra were searched with parent ion mass tolerance of 5 ppm, additional variable modification methionine oxidation (+15.9994 Da) and once with and without variable TMPP-Ac modification of N termini and Lys (+572.1811) in accordance with the respective original publications. In addition, the *D. deserti* datasets were subjected to tryptic and chemotryptic searches to identify the original digestion of each dataset.

A.2.2. Preparation of experimental data

To evaluate the *E. coli* O157:H7 strain *Sakai* experiment (PXD000583), we randomly selected 'replicate 1' and merged the two corresponding datasets of membrane and soluble proteins. For both bacteria we introduced a spectra quality filter by searching against databases including the corresponding organisms using MS-GF+ (Kim and Pevzner, 2014). Thus, the simulation validation relies on relevant spectra and excludes contaminants or unidentifiable low quality spectra. The databases used consist of RefSeq proteins of *E. coli* (taxid 562) and *E. coli* O157:H7 strain *Sakai* (taxid 386585) for the *E. coli* O157:H7 strain Sakai spectra and of RefSeq proteins of *D. deserti* (taxid 310783) including *D. deserti* strain VCD115 (taxid 546414) for *D. deserti* strain VCD115 spectra, respectively. Search parameters were chosen as described above. With respect to the simulation, the filtering search only considered spectra identifications with peptide lengths between 8 and 35. After applying an FDR cutoff of 0.01, for *E. coli* the spectra dataset has been reduced to 17105. For *D. deserti*, 110493 spectra have been identified in the tryptic search without and 98857 with TMPP-Ac modification with an overlap of 89348 spectra. For the chemotryptic search, 55488 spectra have been identified in the tryptic search without and 47814 with TMPP-Ac modification with an overlap of 43242 spectra. However, only tryptic spectra were considered for the evaluation of the LiDSiM simulation.

A.2.3. Preparation of genome data integration

We simulated artificial Illumina paired-end reads including quality values from the NCBI reference genome sequences of *E. coli* O157:H7 strain Sakai (NC_002127.1, NC_002128.1, NC_002695.1) and *D. deserti* strain VCD115 (NC_002127.1, NC_002128.1, NC_002695.1) using Mason (v0.1)(Holtgrewe, 2010). We used the default parameters for the Illumina error model with a read length of 150 bp and 30-fold reference

sequence coverage. Reads of reference chromosomes and plasmids were merged into one artificial sequencing dataset per organism. All datasets including both simulated datasets and the SRA dataset were assembled using Mira (v4.0.2)(Chevreux et al., 1999) with job settings genome, denovo and accurate. According to the Illumina reads, additional parameters indicated Illumina (solexa) as technology, autopairing and - for the simulated reads only - no proposed end clipping.

A.2.4. Additional results

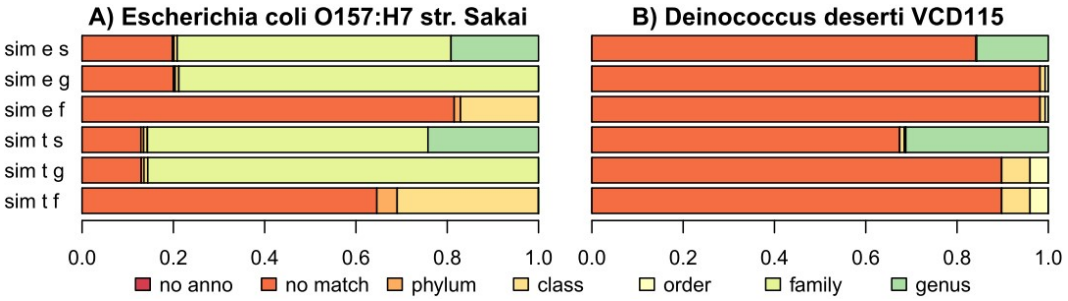

Figure A.9.: Taxonomic level ratios of selected bacteria. For both, *E. coli* O157:H7 strain Sakai (A) and *D. deserti* strain VCD115 (B) the plots show the taxonomic level ratios of the standard simulation once with extracted species (s), genus (g) and family (f), respectively. Results are shown for searches against databases with the corresponding species removed in an exact (e) and error-tolerant (t) search, respectively.

Taxonomic level ratios of selected extractions of higher ranks including genera and families are shown in Figure A.9. For *E. coli*, most peptides with genus matches recur as family matches when the genus is extracted from the database. However, excluding the whole family yields in an increased number of unidentified peptides and only a third is retained by the class. *D. deserti* shows the same results for both, genus and family extractions, as well with an increased number of unidentified peptides. This is due to the fact that the *Deinococcus* genus (taxid 1298) is the only genus with available protein sequences within the *Deinococcaceae* family (taxid 183710).

The analysis of taxonomic level ratio variance due to spectra sampling is illustrated in Figure A.10. The *E. coli* sampling features an increased variance, in particular for unidentified spectra (no match) and genus hits. In contrast, the *D. deserti* sampling shows a rather small variance.

A comparative analysis of proteogenomic enhanced taxonomic level ratio estimations of *E. coli* O157:H7 strain Sakai is shown in Figure A.11. Neither the exact search nor the error-tolerant search features a significant change in taxonomic level

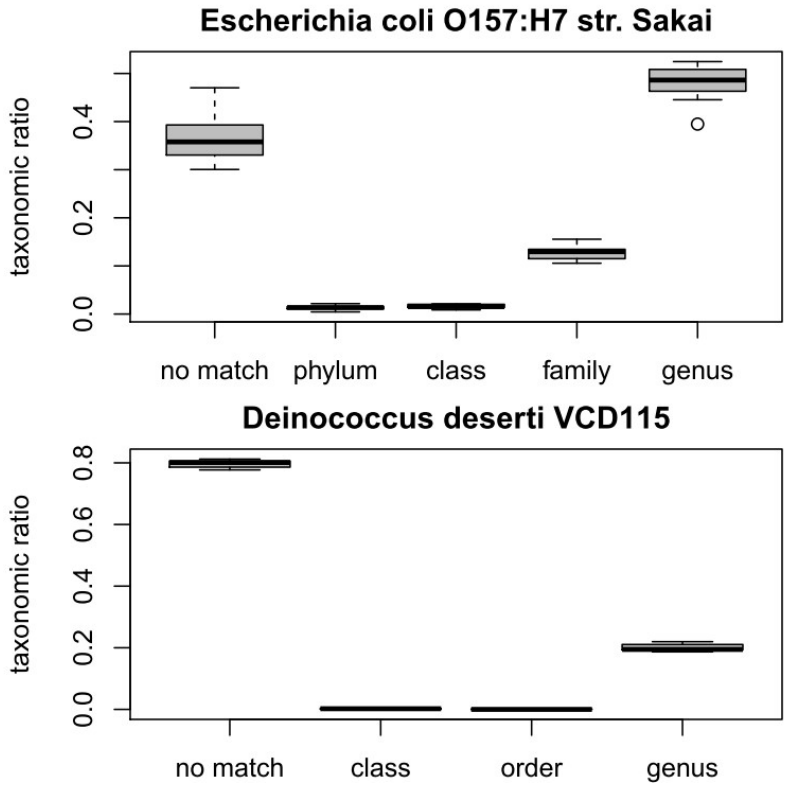

Figure A.10.: Variance of spectra sampling. For *E. coli* and *D. deserti*, taxonomic level ratio differences in the 10 different spectra samples are illustrated by boxplots per taxonomic rank. I.e., each boxplot describes the variance of a rank within the 10 samples.

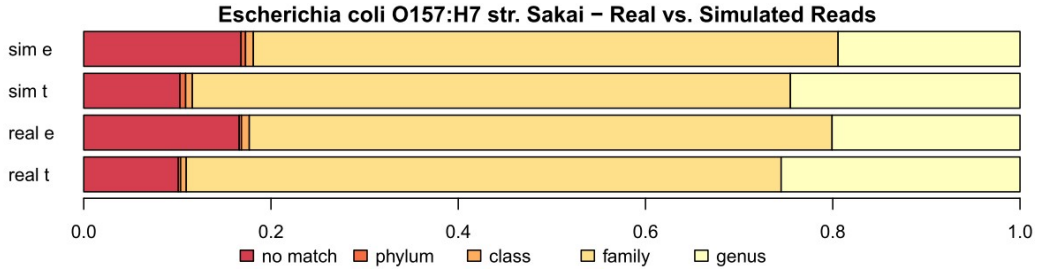

Figure A.11.: Taxonomic level ratios of *E. coli* with simulated and experimental reads. The plots show the taxonomic level ratios of *E. coli* O157:H7 strain Sakai of the evaluation extended with genome sequences based on simulated reads (sim) and experimental reads (real). Results are shown for searches against databases with the corresponding species removed in an exact (e) and error-tolerant (t) search, respectively.

ratios between the integrated genomic sequences based on either simulated or experimental reads.

A.3. Additional Material for Chapter 4

A.3.1. Search Parameters

The cowpox sample of strain *Cowpox virus* (Brighton Red) was acquired in-house. Spectra were analyzed applying a tryptic search with parent ion mass tolerance of 10 ppm, fixed modification cysteine carbamidomethylation (+57 Da) as well as an additional variable modification methionine oxidation (+16 Da).

Bronchitis samples of the strain *Avian infectious bronchitis virus* (strain Beaudette CK) were downloaded from PRIDE (Vizcaíno et al., 2016) (PXD002936) and the sample "BeauR2.raw" was randomly selected for analysis. The raw file was converted to an mgf file using ProteoWizard's MSConvert GUI (3.0.8764) (Chambers et al., 2012). Spectra were analyzed with default settings including a tryptic search with fixed modification cysteine carbamidomethylation (+57 Da) and parent ion mass tolerance of 100 ppm.

The bacillus sample of the strain *Bacillus subtilis* subsp. *subtilis* str. 168 was download from PRIDE (PXD007242, file `614_NG4_BSN238_Urea-Trp_1ug_SR-LFQ_4h_161201.mgf`). Spectra were analyzed with default settings including a tryptic search with fixed modification cysteine carbamidomethylation (+57 Da). However, parent ion mass tolerance was set to 10 ppm in accordance with the original publication.

A.3.2. Additional Figures

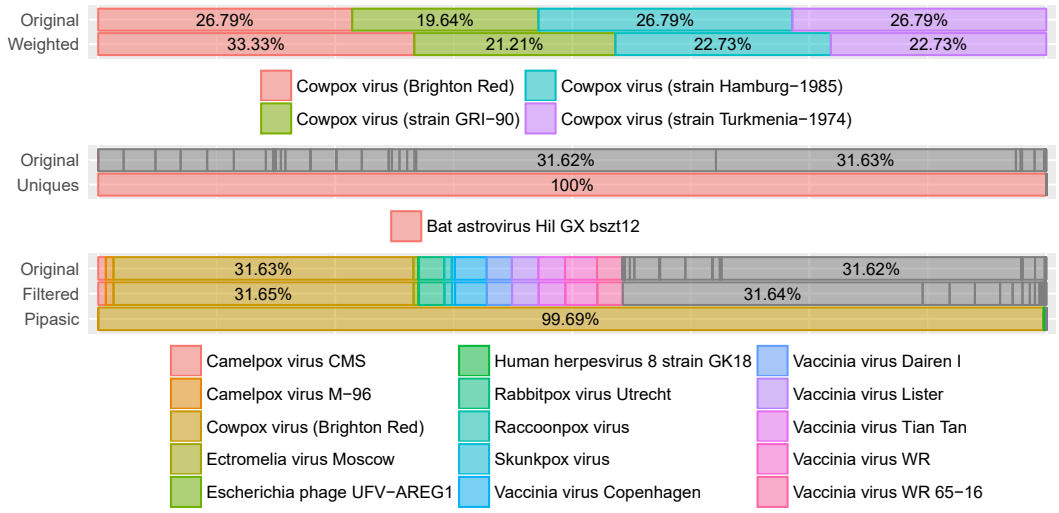

Figure A.12.: Relative counts of cowpox. Relative counts are illustrated for TaxIt (top), uniques- (middle) and Pipasic-based search strategies (bottom). Original, filtered (if applicable) and corrected relative counts are summarized by means of one vertical stacked bar each. Taxa are labeled and color-coded based on a limit of 15 final top candidates (i.e. after correction) with a relative count greater zero. Furthermore, ratios greater 0.05 are highlighted as percentages within bars.

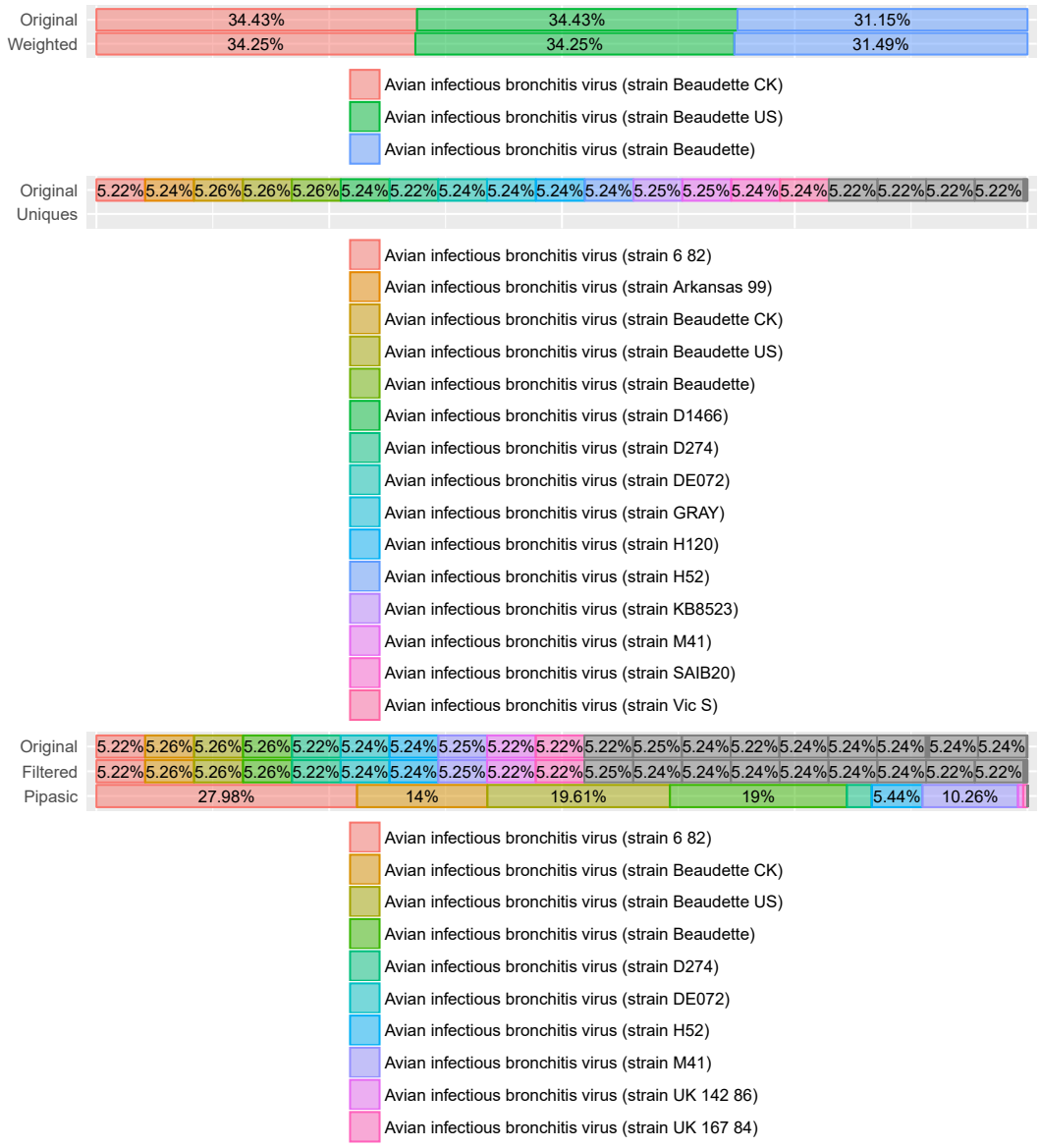

Figure A.13.: Relative counts of bronchitis. Relative counts are illustrated for TaxIt (top), uniques- (middle) and Pipasic-based search strategies (bottom). Original, filtered (if applicable) and corrected relative counts are summarized by means of one vertical stacked bar each. Taxa are labeled and color-coded based on a limit of 15 final top candidates (i.e. after correction, except for uniques) with a relative count greater zero. Furthermore, ratios greater 0.05 are highlighted as percentages within bars.

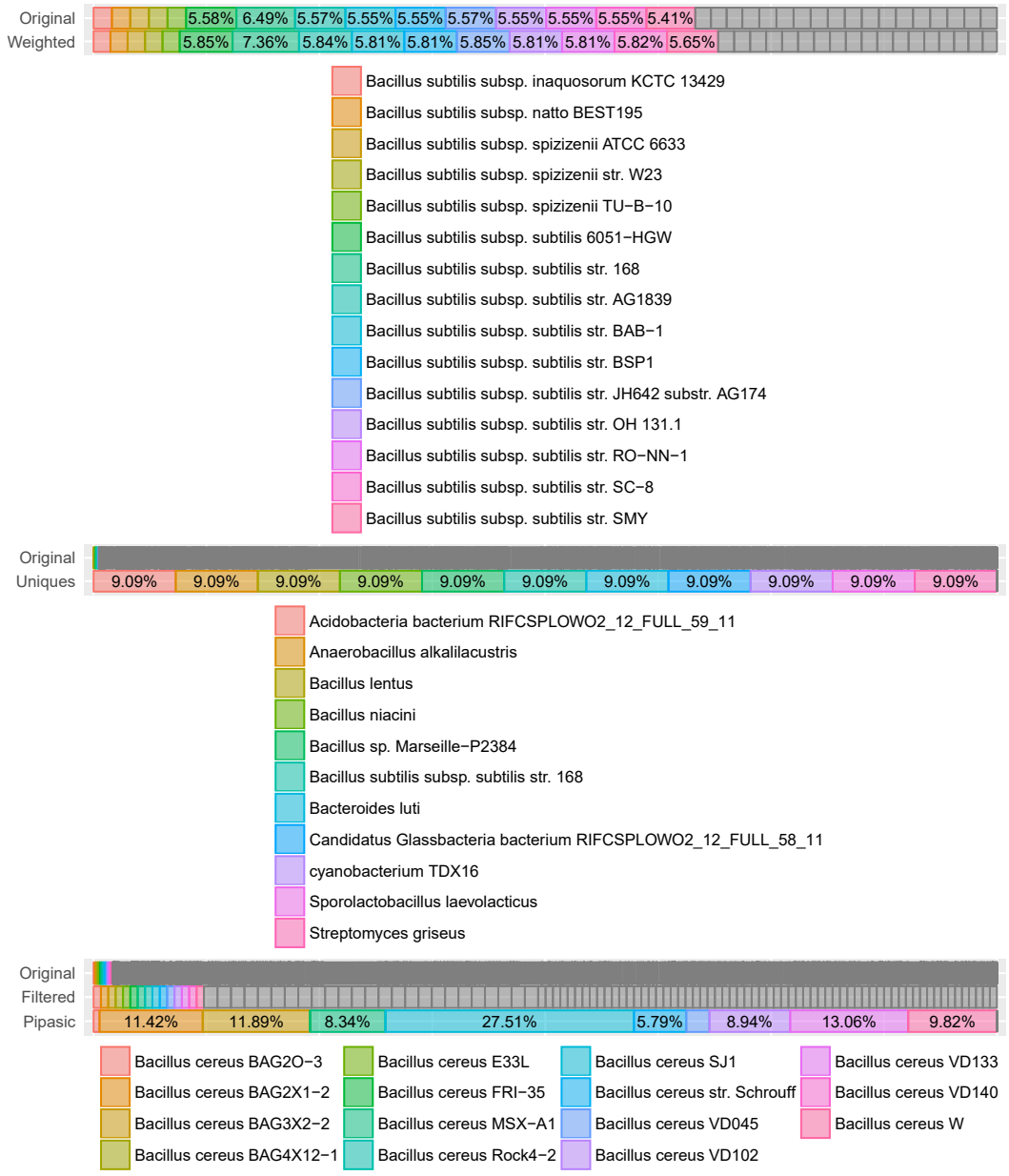

Figure A.14.: Relative counts of bacillus 1k. Relative counts are illustrated for TaxIt (top), uniques- (middle) and Pipasic-based search strategies (bottom). Original, filtered (if applicable) and corrected relative counts are summarized by means of one vertical stacked bar each. Taxa are labeled and color-coded based on a limit of 15 final top candidates (i.e. after correction) with a relative count greater zero. Furthermore, ratios greater 0.05 are highlighted as percentages within bars.

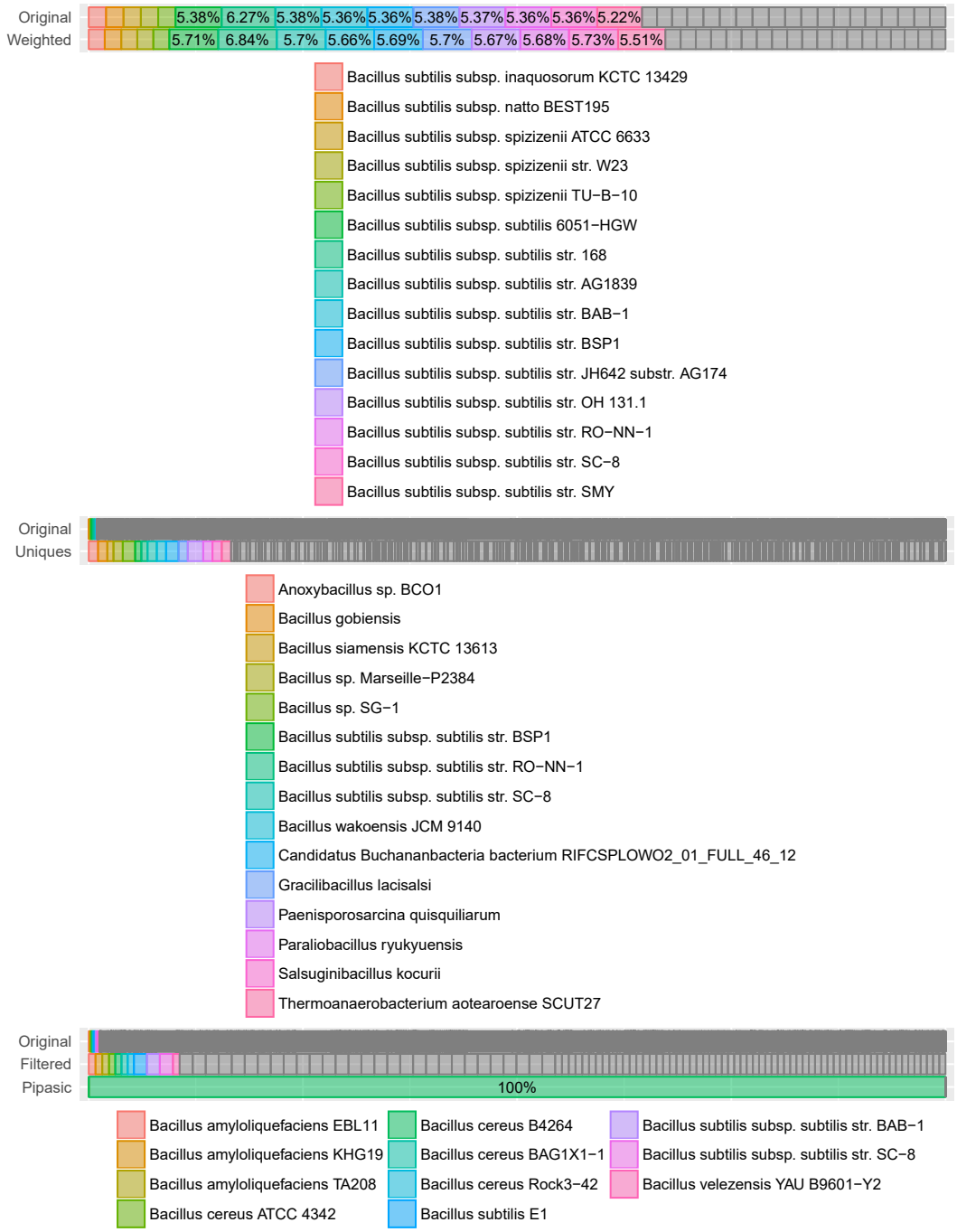

Figure A.15.: Relative counts of bacillus all. Relative counts are illustrated for TaxIt (top), uniques- (middle) and Pipasic-based search strategies (bottom). Original, filtered (if applicable) and corrected relative counts are summarized by means of one vertical stacked bar each. Taxa are labeled and color-coded based on a limit of 15 final top candidates (i.e. after correction) with a relative count greater zero. Furthermore, ratios greater 0.05 are highlighted as percentages within bars.

Bibliography

R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422 (6928):198–207, Mar. 2003. ISSN 0028-0836. doi: 10.1038/nature01511. URL http://dx.doi.org/10.1038/nature01511.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602.

G. Alves, G. Wang, A. Y. Ogurtsov, S. K. Drake, M. Gucek, A. F. Suffredini, D. B. Sacks, and Y.-K. Yu. Identification of Microorganisms by High Resolution Tandem Mass Spectrometry with Accurate Statistical Significance. *Journal of The American Society for Mass Spectrometry*, 27(2):194–210, Feb. 2016. ISSN 1044-0305, 1879-1123. doi: 10.1007/s13361-015-1271-2. URL https://link.springer.com/article/10.1007/s13361-015-1271-2.

C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle. Deep learning for computational biology. *Molecular Systems Biology*, 12(7):878, July 2016. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20156651. URL http://msb.embopress.org/content/12/7/878.

J. Armengaud, E. Marie Hartmann, and C. Bland. Proteogenomics for environmental microbiology. *PROTEOMICS*, 13(18-19):2731–2742, Oct. 2013. ISSN 1615-9861. doi: 10.1002/pmic.201200576. URL http://onlinelibrary.wiley.com/doi/10.1002/pmic.201200576/abstract.

J. Armengaud, J. Trapp, O. Pible, O. Geffard, A. Chaumot, and E. M. Hartmann. Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics*, 105:5–18, June 2014. ISSN 1874-3919. doi: 10.1016/j.jprot.2014.01.007. URL http://www.sciencedirect.com/science/article/pii/S1874391914000190.

M. G. Awan and F. Saeed. MS-REDUCE: an ultrafast technique for reduction of big mass spectrometry data for high-throughput processing. *Bioinformatics (Oxford, England)*, 32(10):1518–1526, May 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw023.

M. Baudet, P. Ortet, J.-C. Gaillard, B. Fernandez, P. Guérin, C. Enjalbal, G. Subra, A. d. Groot, M. Barakat, A. Dedieu, and J. Armengaud. Proteomics-based Refinement of Deinococcus deserti Genome Annotation Reveals an Unwonted Use of Non-canonical Translation Initiation Codons. *Molecular & Cellular Proteomics*, 9(2):415–426, Jan. 2010. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.M900359-MCP200. URL `http://www.mcponline.org/content/9/2/415`.

Z. Bengali, A. C. Townsley, and B. Moss. Vaccinia virus strain differences in cell attachment and entry. *Virology*, 389(1):132–140, June 2009. ISSN 0042-6822. doi: 10.1016/j.virol.2009.04.012. URL `http://www.sciencedirect.com/science/article/pii/S0042682209002608`.

Z. Bengali, P. S. Satheshkumar, and B. Moss. Orthopoxvirus species and strain differences in cell entry. *Virology*, 433(2):506–512, Nov. 2012. ISSN 0042-6822. doi: 10.1016/j.virol.2012.08.044. URL `http://www.sciencedirect.com/science/article/pii/S0042682212004254`.

C. Bielow, S. Aiche, S. Andreotti, and K. Reinert. MSSimulator: Simulation of Mass Spectrometry Data. *Journal of Proteome Research*, 10(7):2922–2929, July 2011. ISSN 1535-3893. doi: 10.1021/pr200155f. URL `http://dx.doi.org/10.1021/pr200155f`.

F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, May 1997. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.277.5331.1453. URL `http://www.sciencemag.org/content/277/5331/1453`.

F. Boulund, R. Karlsson, L. Gonzales-Siles, A. Johnning, N. Karami, O. Al-Bayati, C. Åhrén, E. R. B. Moore, and E. Kristiansson. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics. *Molecular & cellular proteomics : MCP*, 16(6):1052–1063, June 2017. ISSN 1535-9476. doi: 10.1074/mcp.M116.061721. URL `http://europepmc.org/abstract/MED/28420677`.

K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. D. Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, l. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J.

Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, July 2013. ISSN 2047-217X. doi: 10.1186/2047-217X-2-10. URL `http://www.gigasciencejournal.com/content/2/1/10/abstract`.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `http://www.springerlink.com/content/u0p06167n6173512/abstract/`.

Bruker IR. Strain Typing with IR Biotyper Overview | Bruker. URL `https://www.bruker.com/applications/microbiology/strain-typing-with-ir-biotyper/overview.html`. Last visited 2017-09-29.

Bruker MALDI. MALDI Biotyper Systems | Bruker. URL `https://www.bruker.com/products/mass-spectrometry-and-separations/maldi-biotyper-systems.html`. Last visited 2017-09-29.

C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, Dec. 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL `http://www.biomedcentral.com/1471-2105/10/421/abstract`.

M. J. P. Chaisson, R. K. Wilson, and E. E. Eichler. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11):627–640, Nov. 2015. ISSN 1471-0056. doi: 10.1038/nrg3933. URL `http://www.nature.com/nrg/journal/v16/n11/abs/nrg3933.html`.

V. Chaitankar, G. Karakülah, R. Ratnapriya, F. O. Giuste, M. J. Brooks, and A. Swaroop. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. *Progress in Retinal and Eye Research*, 55 (Supplement C):1–31, Nov. 2016. ISSN 1350-9462. doi: 10.1016/j.preteyeres.2016.06.001. URL `http://www.sciencedirect.com/science/article/pii/S1350946216300301`.

M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920, Oct. 2012. ISSN 1087-0156. doi: 10.1038/nbt.2377. URL `http://www.nature.com/nbt/journal/v30/n10/full/nbt.2377.html?foxtrotcallback=true`.

B. Chapman and M. Bellgard. High-throughput parallel proteogenomics: A bacterial case study. *PROTEOMICS*, 14(23-24):2780–2789, Dec. 2014. ISSN 1615-9861. doi: 10.1002/pmic.201400185. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201400185/abstract`.

K. Chen and L. Pachter. Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLOS Computational Biology*, 1(2):e24, July 2005. ISSN 1553-7358. doi: 10.1371/journal.pcbi.0010024. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010024`.

B. Chevreux, T. Wetter, and S. Suhai. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, 99:45–56, 1999. URL `http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html`.

J. Y. Choi, C. D. Sifri, B. C. Goumnerov, L. G. Rahme, F. M. Ausubel, and S. B. Calderwood. Identification of Virulence Genes in a Pathogenic Strain of Pseudomonas aeruginosa by Representational Difference Analysis. *Journal of Bacteriology*, 184(4):952–961, Feb. 2002. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.184.4.952-961.2002. URL `http://jb.asm.org/content/184/4/952`.

S. C. Clark, R. Egan, P. I. Frazier, and Z. Wang. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, 29(4):435–443, Feb. 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts723. URL `http://bioinformatics.oxfordjournals.org/content/29/4/435`.

R. Craig and R. C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrom-*

etry, 17(20):2310–2316, Oct. 2003. ISSN 1097-0231. doi: 10.1002/rcm.1198. URL http://onlinelibrary.wiley.com/doi/10.1002/rcm.1198/abstract.

R. Craig and R. C. Beavis. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–1467, June 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth092. URL https://academic.oup.com/bioinformatics/article/20/9/1466/195237/TANDEM-matching-proteins-with-tandem-mass-spectra.

S. D. Dent, D. Xia, J. M. Wastling, B. W. Neuman, P. Britton, and H. J. Maier. The proteome of the infectious bronchitis virus Beau-R virion. *Journal of General Virology*, 96(12):3499–3506, 2015. doi: 10.1099/jgv.0.000304. URL http://jgv.microbiologyresearch.org/content/journal/jgv/10.1099/jgv.0.000304.

J. Doellinger, L. Schaade, and A. Nitsche. Comparison of the Cowpox Virus and Vaccinia Virus Mature Virion Proteome: Analysis of the Species- and Strain-Specific Proteome. *PLOS ONE*, 10(11):e0141527, Nov. 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0141527. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141527.

B. Domon and R. Aebersold. Mass Spectrometry and Protein Analysis. *Science*, 312(5771):212–217, Apr. 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1124619. URL http://science.sciencemag.org/content/312/5771/212.

D. J. Ecker, J. J. Drader, J. Gutierrez, A. Gutierrez, J. C. Hannis, A. Schink, R. Sampath, L. B. Blyn, M. W. Eshoo, T. A. Hall, M. Tobarmosquera, Y. Jiang, K. A. Sannes-Lowery, L. L. Cummins, B. Libby, D. J. Walcott, C. Massire, R. Ranken, S. Manalili, C. Ivy, R. Melton, H. Levene, V. Harpin, F. Li, N. White, M. Pear, J. A. Ecker, V. Samant, D. Knize, D. Robbins, K. Rudnick, F. Hajjar, and S. A. Hofstadler. The Ibis T5000 Universal Biosensor: An Automated Platform for Pathogen Identification and Strain Typing. *JALA: Journal of the Association for Laboratory Automation*, 11(6):341–351, Dec. 2006. ISSN 1535-5535. doi: 10.1016/j.jala.2006.09.001. URL http://journals.sagepub.com/doi/abs/10.1016/j.jala.2006.09.001.

S. R. Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14 (9):755–763, 1998. ISSN 1367-4803.

R. Ekblom and J. B. W. Wolf. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9):1026–1042, Nov. 2014. ISSN 1752-4571. doi: 10.1111/eva.12178. URL http://onlinelibrary.wiley.com/doi/10.1111/eva.12178/abstract.

M. Fischer, B. Strauch, and B. Y. Renard. Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics (Oxford, England)*, 33(14):i124–i132, July 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx237.

P. Flicek and E. Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11s):S6–S12, Nov. 2009. ISSN 1548-7091. doi: 10.1038/nmeth.1376. URL `https://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1376.html`.

M.-T. Gekenidis, P. Studer, S. Wüthrich, R. Brunisholz, and D. Drissner. Beyond the Matrix-Assisted Laser Desorption Ionization (MALDI) Biotyping Workflow: in Search of Microorganism-Specific Tryptic Peptides Enabling Discrimination of Subspecies. *Applied and Environmental Microbiology*, 80(14):4234–4241, July 2014. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.00740-14. URL `http://aem.asm.org/content/80/14/4234`.

E. Genersch, A. Ashiralieva, and I. Fries. Strain- and Genotype-Specific Differences in Virulence of Paenibacillus larvae subsp. larvae, a Bacterial Pathogen Causing American Foulbrood Disease in Honeybees. *Applied and Environmental Microbiology*, 71(11):7551–7555, Jan. 2005. ISSN 0099-2240, 1098-5336. doi: 10.1128/AEM.71.11.7551-7555.2005. URL `http://aem.asm.org/content/71/11/7551`.

M. Ghodsi, C. M. Hill, I. Astrovskaya, H. Lin, D. D. Sommer, S. Koren, and M. Pop. De novo likelihood-based measures for comparing genome assemblies. *BMC Research Notes*, 6(1):334, Aug. 2013. ISSN 1756-0500. doi: 10.1186/1756-0500-6-334. URL `http://www.biomedcentral.com/1756-0500/6/334/abstract`.

T. C. Glenn. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769, 2011. ISSN 1755-0998. doi: 10.1111/j.1755-0998.2011.03024.x. URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1755-0998.2011.03024.x/abstract`.

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, June 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.49. URL `http://www.nature.com/nrg/journal/v17/n6/full/nrg.2016.49.html`.

A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, Apr. 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt086. URL `http://bioinformatics.oxfordjournals.org/content/29/8/1072`.

J. M. Heather and B. Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, Jan. 2016. ISSN 0888-7543. doi:

10.1016/j.ygeno.2015.11.003. URL http://www.sciencedirect.com/science/article/pii/S0888754315300410.

J. Hedley. jsoup Java HTML Parser, with best of DOM, CSS, and jquery. URL https://jsoup.org/. Last visited 2017-09-29.

J. Henson, G. Tischler, and Z. Ning. Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, 13(8):901–915, June 2012. ISSN 1462-2416. doi: 10.2217/pgs.12.72. URL https://www.futuremedicine.com/doi/abs/10.2217/pgs.12.72.

R. L. Hettich, C. Pan, K. Chourey, and R. J. Giannone. Metaproteomics: Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities. *Analytical Chemistry*, 85(9):4203–4214, May 2013. ISSN 0003-2700. doi: 10.1021/ac303053e. URL http://dx.doi.org/10.1021/ac303053e.

HMMER. HMMER. URL http://hmmer.org/. Last visited 2017-09-28.

M. Holtgrewe. Mason – A Read Simulator for Second Generation Sequencing Data. *Technical Report FU Berlin*, Oct. 2010. URL http://publications.mi.fu-berlin.de/962/.

D. S. Horner, G. Pavesi, T. Castrignanò, D. Meo, P. D'Onorio, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2):181–197, Mar. 2010. ISSN 1467-5463. doi: 10.1093/bib/bbp046. URL https://academic.oup.com/bib/article/11/2/181/216274/Bioinformatics-approaches-for-genomics-and-post.

R. Hrdlickova, M. Toloue, and B. Tian. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):n/a–n/a, Jan. 2017. ISSN 1757-7012. doi: 10.1002/wrna.1364. URL http://onlinelibrary.wiley.com/doi/10.1002/wrna.1364/abstract.

J.-C. Hsu, T.-Y. Chien, C.-C. Hu, M.-J. M. Chen, W.-J. Wu, H.-T. Feng, D. S. Haymer, and C.-Y. Chen. Discovery of Genes Related to Insecticide Resistance in Bactrocera dorsalis by Functional Genomic Analysis of a De Novo Assembled Transcriptome. *PLoS ONE*, 7(8):e40950, Aug. 2012. doi: 10.1371/journal.pone.0040950. URL http://dx.doi.org/10.1371/journal.pone.0040950.

M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto. REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, 14 (5):R47, May 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-5-r47. URL http://genomebiology.com/2013/14/5/R47/abstract.

P. Jagtap, T. McGowan, S. Bandhakavi, Z. J. Tu, S. Seymour, T. J. Griffin, and J. D. Rudney. Deep metaproteomic analysis of human salivary supernatant. *PROTEOMICS*, 12(7):992–1001, Apr. 2012. ISSN 1615-9861. doi: 10.1002/pmic.201100503. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201100503/abstract`.

P. Jagtap, J. Goslinga, J. A. Kooren, T. McGowan, M. S. Wroblewski, S. L. Seymour, and T. J. Griffin. A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *PROTEOMICS*, 13(8):1352–1357, Apr. 2013. ISSN 1615-9861. doi: 10.1002/pmic.201200352. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201200352/abstract`.

P. D. Jagtap, J. E. Johnson, G. Onsongo, F. W. Sadler, K. Murray, Y. Wang, G. M. Shenykman, S. Bandhakavi, L. M. Smith, and T. J. Griffin. Flexible and Accessible Workflows for Improved Proteogenomic Analysis Using the Galaxy Framework. *Journal of Proteome Research*, 13(12):5898–5908, Dec. 2014. ISSN 1535-3893. doi: 10.1021/pr500812t. URL `http://dx.doi.org/10.1021/pr500812t`.

K. Jeong, S. Kim, and N. Bandeira. False discovery rates in spectral identification. *BMC Bioinformatics*, 13(16):S2, Nov. 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S16-S2. URL `https://doi.org/10.1186/1471-2105-13-S16-S2`.

M. Junqueira, V. Spirin, T. S. Balbuena, H. Thomas, I. Adzhubei, S. Sunyaev, and A. Shevchenko. Protein identification pipeline for the homology-driven proteomics. *Journal of Proteomics*, 71(3):346–356, Aug. 2008. ISSN 1874-3919. doi: 10.1016/j.jprot.2008.07.003. URL `http://www.sciencedirect.com/science/article/pii/S1874391908001115`.

D. R. Kelley, M. C. Schatz, and S. L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11(11):R116, Nov. 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-11-r116. URL `http://genomebiology.com/2010/11/11/R116/abstract`.

W. J. Kent. BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, Jan. 2002. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.229202. URL `http://genome.cshlp.org/content/12/4/656`.

A. Kertesz-Farkas, U. Keich, and W. S. Noble. Tandem Mass Spectrum Identification via Cascaded Search. *Journal of Proteome Research*, 14(8):3027–3038, Aug. 2015. ISSN 1535-3893. doi: 10.1021/pr501173s. URL `http://dx.doi.org/10.1021/pr501173s`.

S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, 5:5277, Oct. 2014. doi: 10.1038/ncomms6277. URL `http://www.nature.com/ncomms/2014/141031/ncomms6277/full/ncomms6277.html`.

L. Käll and O. Vitek. Computational mass spectrometry-based proteomics. *PLoS computational biology*, 7(12):e1002277, Dec. 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002277.

C. Kocharunchitt, T. King, K. Gobius, J. P. Bowman, and T. Ross. Global Genome Response of Escherichia coli O157:H7 Sakai during Dynamic Changes in Growth Kinetics Induced by an Abrupt Downshift in Water Activity. *PLoS ONE*, 9(3): e90422, Mar. 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0090422. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3940904/`.

A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5):513–520, May 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4256.

J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, Jan. 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts480. URL `http://bioinformatics.oxfordjournals.org/content/28/19/2520`.

M. Kuhring and B. Y. Renard. iPiG: Integrating Peptide Spectrum Matches into Genome Browser Visualizations. *PLoS ONE*, 7(12):e50246, Dec. 2012. doi: 10.1371/journal.pone.0050246. URL `http://dx.doi.org/10.1371/journal.pone.0050246`.

M. Kuhring and B. Y. Renard. Estimating the computational limits of detection of microbial non-model organisms. *PROTEOMICS*, 15(20):3580–3584, Oct. 2015. ISSN 1615-9861. doi: 10.1002/pmic.201400598. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201400598/abstract`.

T. Kwon, H. Choi, C. Vogel, A. I. Nesvizhskii, and E. M. Marcotte. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. *Journal of proteome research*, 10(7):2949–2958, July 2011. ISSN 1535-3893. doi: 10.1021/pr2002116. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3128686/`.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Apr. 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923. URL `http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html`.

M. Laue. Chapter 1 - Electron Microscopy of Viruses. In T. Müller-Reichert, editor, *Methods in Cell Biology*, volume 96 of *Electron Microscopy of Model Systems*, pages 1–20. Academic Press, Jan. 2010. URL `http://www.sciencedirect.com/science/article/pii/S0091679X10960019`. DOI: 10.1016/S0091-679X(10)96001-9.

R. Leinonen, H. Sugawara, and M. Shumway. The Sequence Read Archive. *Nucleic Acids Research*, 39(Database issue):D19–D21, Jan. 2011. ISSN 0305-1048. doi: 10.1093/nar/gkq1019. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013647/`.

M. S. Lindner and B. Y. Renard. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research*, 41(1):e10, Jan. 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks803. URL `http://nar.oxfordjournals.org/content/41/1/e10`.

T. Liu, M. E. Belov, N. Jaitly, W.-J. Qian, and R. D. Smith. Accurate Mass Measurements in Proteomics. *Chemical Reviews*, 107(8):3621–3653, Aug. 2007. ISSN 0009-2665. doi: 10.1021/cr068288j. URL `http://dx.doi.org/10.1021/cr068288j`.

M. Locard-Paulet, O. Pible, A. Gonzalez de Peredo, B. Alpha-Bazin, C. Almunia, O. Burlet-Schiltz, and J. Armengaud. Clinical implications of recent advances in proteogenomics. *Expert Review of Proteomics*, 13(2):185–199, 2016. ISSN 1744-8387. doi: 10.1586/14789450.2016.1132169.

L. Luo, E. Boerwinkle, and M. Xiong. Association studies for next-generation sequencing. *Genome Research*, 21(7):1099–1108, Jan. 2011. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.115998.110. URL `http://genome.cshlp.org/content/21/7/1099`.

R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, Dec. 2012. ISSN 2047-217X. doi: 10.1186/2047-217X-1-18. URL `http://www.gigasciencejournal.com/content/1/1/18/abstract`.

K. Ma, O. Vitek, and A. I. Nesvizhskii. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*, 13 (16):S1, Nov. 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-S16-S1. URL `https://doi.org/10.1186/1471-2105-13-S16-S1`.

M. Mann and N. L. Kelleher. Precision proteomics: The case for high resolution and high mass accuracy. *Proceedings of the National Academy of Sciences*, 105 (47):18132–18138, Nov. 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0800788105. URL http://www.pnas.org/content/105/47/18132.

M. Mascher, G. J. Muehlbauer, D. S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, M. Muñoz-Amatriaín, T. J. Close, R. P. Wise, A. H. Schulman, A. Himmelbach, K. F. Mayer, U. Scholz, J. A. Poland, N. Stein, and R. Waugh. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *The Plant Journal*, 76(4):718–727, Nov. 2013. ISSN 1365-313X. doi: 10.1111/tpj.12319. URL http://onlinelibrary.wiley.com/doi/10.1111/tpj.12319/abstract.

A. Masselot. InSilicoSpectro-Databanks-0.0.43 - search.cpan.org. URL http://search.cpan.org/~alexmass/InSilicoSpectro-Databanks-0.0.43/. Last visited 2017-09-29.

L. McHugh and J. W. Arthur. Computational Methods for Protein Identification from Mass Spectrometry Data. *PLoS Comput Biol*, 4(2):e12, Feb. 2008. doi: 10.1371/journal.pcbi.0040012. URL http://dx.plos.org/10.1371/journal.pcbi.0040012.

G. Menschaert and D. Fenyö. Proteogenomics from a bioinformatics angle: A growing field. *Mass Spectrometry Reviews*, pages n/a–n/a, Dec. 2015. ISSN 1098-2787. doi: 10.1002/mas.21483. URL http://onlinelibrary.wiley.com/doi/10.1002/mas.21483/abstract.

J. Muntel, S. A. Boswell, S. Tang, S. Ahmed, I. Wapinski, G. Foley, H. Steen, and M. Springer. Abundance-based classifier for the prediction of mass spectrometric peptide detectability upon enrichment. *Molecular & Cellular Proteomics*, page mcp.M114.044321, Dec. 2014. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.M114.044321. URL http://www.mcponline.org/content/early/2014/12/03/mcp.M114.044321.

T. Muth, M. Vaudel, H. Barsnes, L. Martens, and A. Sickmann. XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results. *PROTEOMICS*, 10(7):1522–1524, Apr. 2010. ISSN 1615-9861. doi: 10.1002/pmic.200900759. URL http://onlinelibrary.wiley.com/doi/10.1002/pmic.200900759/abstract.

T. Muth, D. Benndorf, U. Reichl, E. Rapp, and L. Martens. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular BioSystems*, 9(4):578–585, Mar. 2013. ISSN 1742-2051. doi: 10.1039/C2MB25415H. URL http://pubs.rsc.org/en/content/articlelanding/2013/mb/c2mb25415h.

T. Muth, B. Y. Renard, and L. Martens. Metaproteomic data analysis at a glance: advances in computational microbial community proteomics. *Expert Review of Proteomics*, 13(8):757–769, Aug. 2016. ISSN 1744-8387. doi: 10.1080/14789450.2016.1209418.

N. Nagarajan and M. Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, Mar. 2013. ISSN 1471-0056. doi: 10.1038/nrg3367. URL `http://www.nature.com/nrg/journal/v14/n3/full/nrg3367.html`.

T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40(20):e155, Nov. 2012. ISSN 1362-4962. doi: 10.1093/nar/gks678.

G. Narzisi and B. Mishra. Comparing De Novo Genome Assembly: The Long and Short of It. *PLoS ONE*, 6(4):e19175, Apr. 2011. doi: 10.1371/journal.pone.0019175. URL `http://dx.doi.org/10.1371/journal.pone.0019175`.

A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73(11):2092–2123, Oct. 2010. ISSN 1874-3919. doi: 10.1016/j.jprot.2010.08.009. URL `http://www.sciencedirect.com/science/article/pii/S1874391910002496`.

A. I. Nesvizhskii. Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11):1114–1125, Nov. 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3144. URL `http://www.nature.com/nmeth/journal/v11/n11/full/nmeth.3144.html`.

A. I. Nesvizhskii, F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, S. Baginsky, and R. Aebersold. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides. *Molecular & Cellular Proteomics*, 5(4):652–670, Jan. 2006. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.M500319-MCP200. URL `http://www.mcponline.org/content/5/4/652`.

A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nature Methods*, 4(10):787–797, Oct. 2007. ISSN 1548-7091. doi: 10.1038/nmeth1088. URL `http://www.nature.com/nmeth/journal/v4/n10/full/nmeth1088.html`.

C. Nielsen, S. Jackman, I. Birol, and S. Jones. ABySS-Explorer: Visualizing Genome Sequence Assemblies. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):881–888, Nov. 2009. ISSN 1077-2626. doi: 10.1109/TVCG.2009.116.

K. Ning, D. Fermin, and A. I. Nesvizhskii. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. *PROTEOMICS*, 10(14): 2712–2718, July 2010. ISSN 1615-9861. doi: 10.1002/pmic.200900473. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.200900473/abstract`.

N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, Jan. 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1189. URL `https://academic.oup.com/nar/article/44/D1/D733/2502674/Reference-sequence-RefSeq-database-at-NCBI-current`.

F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12(2):87–98, Feb. 2011. ISSN 1471-0056. doi: 10.1038/nrg2934. URL `http://www.nature.com/nrg/journal/v12/n2/abs/nrg2934.html`.

A. Penzlin, M. S. Lindner, J. Doellinger, P. W. Dabrowski, A. Nitsche, and B. Y. Renard. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics*, 30(12):i149–i156, June 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu267. URL `http://bioinformatics.oxfordjournals.org/content/30/12/i149`.

S. Pfrunder, J. Grossmann, P. Hunziker, R. Brunisholz, M.-T. Gekenidis, and D. Drissner. Bacillus cereus Group-Type Strain-Specific Diagnostic Peptides. *Journal of Proteome Research*, 15(9):3098–3107, Sept. 2016. ISSN 1535-3893. doi: 10.1021/acs.jproteome.6b00216. URL `http://dx.doi.org/10.1021/acs.jproteome.6b00216`.

A. M. Phillippy, M. C. Schatz, and M. Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology*, 9(3):R55, Mar. 2008. ISSN 1465-6906. doi: 10.1186/gb-2008-9-3-r55. URL `http://genomebiology.com/2008/9/3/R55`.

O. Pible, E. M. Hartmann, G. Imbert, and J. Armengaud. The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteomics*, 3:246–249, June 2014. ISSN 2212-9685. doi: 10.1016/j.euprot.2014.04.001. URL `http://www.sciencedirect.com/science/article/pii/S2212968514000269`.

M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13:341, July 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-341. URL `https://doi.org/10.1186/1471-2164-13-341`.

A. Rahman and L. Pachter. CGAL: computing genome assembly likelihoods. *Genome Biology*, 14(1):R8, Jan. 2013. ISSN 1465-6906. doi: 10.1186/gb-2013-14-1-r8. URL `http://genomebiology.com/2013/14/1/R8/abstract`.

K. Reinert, B. Langmead, D. Weese, and D. J. Evers. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, 16(1):133–151, 2015. doi: 10.1146/annurev-genom-090413-025358. URL `https://doi.org/10.1146/annurev-genom-090413-025358`.

B. Y. Renard, B. Xu, M. Kirchner, F. Zickmann, D. Winter, S. Korten, N. W. Brattig, A. Tzur, F. A. Hamprecht, and H. Steen. Overcoming Species Boundaries in Peptide Identification with Bayesian Information Criterion-driven Error-tolerant Peptide Search (BICEPS). *Molecular & Cellular Proteomics*, 11(7):M111.014167, Jan. 2012. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.M111.014167. URL `http://www.mcponline.org/content/11/7/M111.014167`.

S. Renuse, R. Chaerkady, and A. Pandey. Proteogenomics. *PROTEOMICS*, 11 (4):620–630, Feb. 2011. ISSN 1615-9861. doi: 10.1002/pmic.201000615. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201000615/abstract`.

P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277, June 2000. ISSN 0168-9525. doi: 10.1016/S0168-9525(00)02024-2. URL `http://www.sciencedirect.com/science/article/pii/S0168952500020242`.

K. Rooijers, C. Kolmeder, C. Juste, J. Doré, M. d. Been, S. Boeren, P. Galan, C. Beauvallet, W. M. d. Vos, and P. J. Schaap. An iterative workflow for mining the human intestinal metaproteome. *BMC Genomics*, 12(1):6, Jan. 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-6. URL `http://www.biomedcentral.com/1471-2164/12/6/abstract`.

S. L. Salzberg and J. A. Yorke. Beware of mis-assembled genomes. *Bioinformatics*, 21(24):4320–4321, Dec. 2005. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bti769. URL `http://bioinformatics.oxfordjournals.org/content/21/24/4320`.

S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marçais, M. Pop, and J. A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 22(3):557–567, Jan. 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.131383.111. URL `http://genome.cshlp.org/content/22/3/557`.

W. Sanders, N. Wang, S. Bridges, B. Malone, Y. Dandass, F. McCarthy, B. Nanduri, M. Lawrence, and S. Burgess. The Proteogenomic Mapping Tool. *BMC Bioinformatics*, 12(1):115, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-115. URL `http://www.biomedcentral.com/1471-2105/12/115`.

C. N. Schlaffner, G. Pirklbauer, A. Bender, and J. S. Choudhary. PoGo: Jumping from Peptides to Genomic Loci. *bioRxiv*, page 079772, Nov. 2016. doi: 10.1101/079772. URL `https://www.biorxiv.org/content/early/2016/11/26/079772`.

J. Seifert, F.-A. Herbst, P. Halkjær Nielsen, F. J. Planes, N. Jehmlich, M. Ferrer, and M. von Bergen. Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. *PROTEOMICS*, 13(18-19):2786–2804, Oct. 2013. ISSN 1615-9861. doi: 10.1002/pmic.201200566. URL `http://onlinelibrary.wiley.com/doi/10.1002/pmic.201200566/abstract`.

W. Shen, S. Le, Y. Li, and F. Hu. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, 11(10):e0163962, Oct. 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0163962. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163962`.

N. Singhal, M. Kumar, P. K. Kanaujia, and J. S. Virdi. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in Microbiology*, 6, 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00791. URL `http://journal.frontiersin.org/article/10.3389/fmicb.2015.00791/full`.

J.-i. Sohn and J.-W. Nam. The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, 2016. doi: 10.1093/bib/bbw096. URL `https://academic.oup.com/bib/article/doi/10.1093/bib/bbw096/2339783/The-present-and-future-of-de-novo-whole-genome`.

M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher. OpenMS – An open-source software framework for mass spectrometry. *BMC Bioinformatics*, 9 (1):163, Mar. 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-163. URL `http://www.biomedcentral.com/1471-2105/9/163/abstract`.

J. Tarhio and E. Ukkonen. Approximate Boyer–Moore String Matching. *SIAM Journal on Computing*, 22(2):243–260, Apr. 1993. ISSN 0097-5397. doi: 10.1137/0222018. URL `http://epubs.siam.org/doi/abs/10.1137/0222018`.

M. The and L. Käll. MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics. *Journal of Proteome Research*, 15(3):713–720, Mar. 2016. ISSN 1535-3907. doi: 10.1021/acs.jproteome.5b00749.

The Global Proteome Machine. cRAP protein sequences. URL `http://www.thegpm.org/crap/`. Last visited 2017-09-29.

The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(D1):D191–D198, Jan. 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1140. URL `http://nar.oxfordjournals.org/content/42/D1/D191`.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, Jan. 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1099. URL `https://academic.oup.com/nar/article/45/D1/D158/2605721/UniProt-the-universal-protein-knowledgebase`.

K. Trappe, B. Wulf, J. Doellinger, S. Halbedel, T. Muth, and B. Y. Renard. Hortense: Horizontal gene transfer detection directly from proteomic MS/MS data. Sept. 2017. URL `https://peerj.com/preprints/3248`. DOI: 10.7287/peerj.preprints.3248v1.

T. J. Treangen, D. D. Sommer, F. E. Angly, S. Koren, and M. Pop. Next generation sequence assembly with AMOS. *Current Protocols in Bioinformatics*, Chapter 11: Unit 11.8, Mar. 2011. ISSN 1934-340X. doi: 10.1002/0471250953.bi1108s33.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804–810, Oct. 2007. ISSN 0028-0836. doi: 10.1038/nature06244. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3709439/`.

M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, and H. Barsnes. PeptideShaker enables reanalysis of MS-derived

proteomics data sets. *Nature Biotechnology*, 33(1):22–24, Jan. 2015. ISSN 1546-1696. doi: 10.1038/nbt.3109.

W. N. Venables and B. D. Ripley. *Modern applied statistics with S*. Springer, New York, NY [u.a., 2007. ISBN 978-0-387-95457-8 0-387-95457-0.

F. Vezzi, G. Narzisi, and B. Mishra. Feature-by-Feature – Evaluating De Novo Sequence Assembly. *PLoS ONE*, 7(2):e31002, Feb. 2012a. doi: 10.1371/journal.pone.0031002. URL UR-http://dx.doi.org/10.1371/journal.pone.0031002, http://dx.doi.org/10.1371/journal.pone.0031002.

F. Vezzi, G. Narzisi, and B. Mishra. Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, 7(12):e52210, Dec. 2012b. doi: 10.1371/journal.pone.0052210. URL http://dx.doi.org/10.1371/journal.pone.0052210.

J. A. Vizcaíno, A. Csordas, N. del Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q.-W. Xu, R. Wang, and H. Hermjakob. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, 44(D1):D447–D456, Jan. 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1145. URL https://academic.oup.com/nar/article/44/D1/D447/2502640/2016-update-of-the-PRIDE-database-and-its-related.

J. F. Vázquez-Castellanos, R. García-López, V. Pérez-Brocal, M. Pignatelli, and A. Moya. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*, 15(1):37, Jan. 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-37. URL http://www.biomedcentral.com/1471-2164/15/37/abstract.

J. Wang and H. Jia. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology*, 14(8):508–522, Aug. 2016. ISSN 1740-1526. doi: 10.1038/nrmicro.2016.83. URL http://www.nature.com/nrmicro/journal/v14/n8/abs/nrmicro.2016.83.html.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan. 2009. ISSN 1471-0056. doi: 10.1038/nrg2484. URL http://www.nature.com/nrg/journal/v10/n1/abs/nrg2484.html.

D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov,

T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 36(Database issue):D13–21, Jan. 2008. ISSN 1362-4962. doi: 10.1093/nar/gkm1000.

S. Wu and U. Manber. A Fast Algorithm for Multi-Pattern Searching. *Technical Report TR94-17, Department of Computer Science, The University of Arizona*, May 1994.

yafeng. An in silico trypsin digestion tool, June 2017. URL `https://github.com/yafeng/trypsin`. Last visited 2017-09-29.

D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, May 2008. ISSN 1088-9051. doi: 10.1101/gr.074492.107. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2336801/`.

Zusammenfassung

Sequenzdaten bilden das Rückrad für viele biologische Forschungsbereiche, einschließlich (aber nicht beschränkt auf) Genomik, Proteomik sowie Proteogenomik. Sequenzierung wird durch eine breite Auswahl an modernen Technologien ermöglicht, wie beispielsweise Next-Generation-Sequenzierung und Massenspektrometrie. Diese Hochdurchsatzverfahren erzeugen erhebliche Datenmengen mit immer geringerem zeitlichen und finanziellen Aufwand. Die anfallenden Datenvolumina lassen manuelle Aufbereitung nicht mehr zu und benötigen deshalb modernste rechnerische Methoden für eine adäquate Analyse und Interpretation. In der Proteogenomik wird das Potential die verschiedene Omik-Technologien zu kombinieren häufig betont, insbesondere für Non-Model-Organismen. In dieser Dissertation möchten wir einige Herausforderungen im „Lebenszyklus“ der Sequenzdaten hervorheben und uns eingehender mit ihnen befassen, von Genomsequenzierung über integrative Evaluierung zu extensiver Anwendung umfangreicher Sequenzdatenbanken.

Wir beschreiben einige Methoden mit ihrer Anwendung in unterschiedlichen Omik-Gebieten und betrachten zusätzlich die Möglichkeiten einer potentiell integrativen Analyse. Zunächst stellen wir eine Methode für das Ranking von *de novo* assemblierten Contigs basierend auf maschinellem Lernen vor. Dabei heben wir das besondere Potential für die Anwendung auf metagenomische Sequenzdaten hervor, welche für gewöhnlich ein große Vielfalt an zuvor sequenzierten als auch unsequenzierten Non-Model-Organismen aufweisen. Des Weiteren untersuchen wir den Einfluss von Sequenz-Verfügbarkeit in angewendeten Datenbanken in Bezug auf taxonomischem Klassifizierungspotential von Tandem-MS-Spektren. Dabei analysieren wir die Effekte verschiedener Sequenzquellen und Such-Strategien auf die taxonomische Tiefe. Abschließend stellen wir einen neuen Ansatz für eine extensive taxonomische Klassifizierung durch iterativer Aufarbeitung möglichst aktueller und umfangreicher Protein-Sequenz-Datenbanken. Wir diskutieren Potential und Grenzen unserer Methoden mit Hinblick auf aktuelle Sequenzdaten-Verfügbarkeit. Dabei zeigen wir potentiellen Nutzen für Non-Model Organismen auf.

Eigenständigkeitserklärung

Ich versichere, dass ich die hier vorgelegte Dissertation selbstständig angefertigt habe und die benutzten Quellen und Hilfsmittel vollständig angegeben sind.

Ein Promotionsverfahren wurde zu keinem früheren Zeitpunkt an einer anderen in- oder ausländischen Hochschule oder bei einem anderen Fachbereich beantragt. Die Bestimmungen der Promotionsordnung sind mir bekannt.

Mathias Kuhring, Berlin, 29.09.2017