Reference and taxonomy-based methods for classification and abundance estimation of organisms in metagenomic samples

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

von

Vitor C. Piro

Berlin, 2018

Betreuer: PD Dr. Bernhard Y. Renard

Erstgutachter: PD Dr. Bernhard Y. Renard

Zweitgutachter: Prof. Dr. Nicola Segata

Tag der Disputation: 23.11.2018

Acknowledgements

I would like to thank my supervisor Bernhard Y. Renard for the trust and opportunity given to me. It was an honor to join such an amazing research group and to have his continuous support and supervision during all those years. I would like to thank all the colleagues I worked, discussed and shared time with in the course of my studies. I specially thank my parents Roseli and Rubens as well as my close family for all the support from the beginning to the end of my PhD, even though they are so far away. Finally, I thank Lisa for being such a lovely and understanding girlfriend.

This work was made possible with the infrastructure and support provided by the Robert Koch Institute and the Freie Universität Berlin and with the funding provided by the Brazilian government via the Coordination for the Improvement of Higher Education Personnel (CAPES) - Programa Ciência sem Fronteiras [BEX 13472/13-5].

Contents

Abstract

Metagenomics provides the means to study the vast and still mostly unknown microbial world which comprises at least half of earth's genetic diversity. Computational metagenomics enables those discoveries via analysis of large amounts of data which are being generated in a fast pace with high-throughput technologies. Reference-based methods are commonly used to study environmental samples based on a set of previously assembled reference sequences which are often linked to a taxonomic classification. Finding the origin of each sequenced fragment and profiling an environmental sample as a whole are the main goals of binning and taxonomic profiling tools, respectively.

In this thesis I present three methods in computational metagenomics. Sets of curated reference sequences jointly with taxonomic classification are employed to characterize community samples. The main goal of those contributions is to improve the state-of-the-art of taxonomic profiling and binning, with fast, sensitive and precise methods.

First I present ganon, a sequence classification tool for metagenomics which works with a very large number of reference sequences. Ganon provides an efficient method to index sequences and to keep those indices updated in very short time. In addition, ganon performs taxonomic binning with strongly improved precision compared to the current available methods. For a general profiling of metagenomic samples and abundance estimation I introduce DUDes. Rather than predicting strains in the sample based only on relative abundances, DUDes first identifies possible candidates by comparing the strength of mapped reads in each node of the taxonomic tree in an iterative top-down manner. This technique works in an opposite direction of the lowest common ancestor approach. Lastly, I present MetaMeta, a pipeline to execute metagenome analysis tools and integrate their results. MetaMeta is a method to combine and enhance results from multiple taxonomic binning and profiling tools and at the same time a pipeline to easily execute tools and analyze environmental data. MetaMeta includes database generation, pre-processing, execution, and integration steps, allowing easy installation, visualization and parallelization of state-of-the-art tools. Using the same input data, MetaMeta provides more sensitive and reliable results with the presence of each identified organism being supported by several methods.

Those three projects introduce new methodologies and improved results over similar methods, constituting valuable contributions to characterize communities in a reference and taxonomy-based manner.

Chapter 1

Introduction

1.1 Computational Metagenomics

The field of metagenomics aims to study and analyze whole genomes directly from environmental samples bypassing the cultivation of clonal cultures and allowing the exploration of the collective genome sequences from an entire community of microbes at once. The modern term metagenomics is often defined as the process to characterize the collection of genomes from the microorganisms present in a defined environment through DNA extraction and shotgun sequencing [Marchesi and Ravel, 2015], a definition that differs from the original use of the term [Handelsman et al., 1998]. Differently from the study of the genome content of a single culturable organism, metagenomics enables the study of genomes of all microorganisms present in a certain environment at the same time. This approach is valuable to many analysis in clinical microbiology like pathogen detection [Schlaberg et al., 2017], outbreak investigation [Loman et al., 2013] and also to bio forensics, bio surveillance, ecology studies, among others.

Whole-metagenome shotgun (WMS) sequencing or just shotgun metagenomics is the name of the technique used to obtain sequence fragments from a microbial community to allow the study of a metagenome. This approach is different in many aspects from other "meta" techniques, for example metataxonomics (also called metabarcoding) which aims to sequence a single or multiple loci from each strain in an environment sample [Marchesi and Ravel, 2015]. Despite similarities in applications, for example community profiling based on the highly conserved gene 16S rRNA for Bacteria, metataxonomics cannot be considered a metagenomic study, since they only cover the contents of specific genomic regions and not the whole genome. Metataxonomics can provide fast and cost effective identification of bacteria and eukaryotes (but not viruses) based on conserved genes, but suffers from amplification bias [Eloe-Fadrosh et al., 2016]. Metagenomics provides broader possibilities of analysis through WMS sequencing, achieving higher resolution and coverage of the studied community, covering their complete genome content while metataxonomics techniques are limited to a chosen sub-set of the genomic content.

In metagenomics, sequence fragments are usually obtained with high-throughput machines which generate millions to billions of small fragments called reads [Goodwin et al., 2016]. Those reads can be generated in a variety of sizes and configurations, a choice that depends on the goal of the research and which questions shall be answered or tackled. Widely used in metagenomics, Illumina machines produce short reads ranging from 150 to

300 base pairs. Newer technologies like the single molecule real-time sequencing (SMRT) from Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies provide longer reads ranging from 10.000 to 16.000 base pairs in average [Ardui et al., 2018] . However, the long reads come with a cost of a higher error rate: 13-15% against ~0.1% from Illumina short reads. Even with many improvements towards reducing such error rates, Illumina is still the main driver for metagenomics. Those small fragments of sequences obtained by high-throughput machines are the main source of input for the computational methods and further analyzed by bioinformaticians. They are used to answer a wide range of questions: which organisms are in the sample? How abundant are they? How are they distributed? What can they do? Are they interacting? Do they change? A vast catalogue of computational tools is currently available for answering those questions. Many of them were catalogued, reviewed and compared against each other on the following publications [Oulas et al., 2015, McIntyre et al., 2017, Lindgreen et al., 2016, Peabody et al., 2015, Sczyrba et al., 2017].

In this thesis I will focus on the primary questions: "which organisms are in the sample?" and "how abundant are they?" and present new methods and approaches to improve the state-of-the-art to answer those questions. Those contributions are commonly classified in the categories: assembly, binning and taxonomic profiling. Those categories will be further detailed in the following sections.

1.1.1 Assembly

Assembly is the process of putting together several small fragments of sequences into one or more bigger segments, recreating the original genome sequence of a certain strain. When successful, a genome assembly will contain all base pairs in the same order and orientation of the DNA molecules of a single organism. Given the complexity and size of genomes compared to the relatively small size of the sequenced reads, genome assembly is by no means a trivial task. Due to repeated and low complexity regions, sequencing errors and other factors [Nagarajan and Pop, 2013] many assemblers cannot reconstruct the entire real sequence from the genome, fragmenting the assembly into smaller pieces, called contigs [Sohn and Nam, 2016].

In metagenomics, the assembly process is even more complex [Vollmers et al., 2017, Teraguchi et al., 2014]. Communities carry genomic information of several organisms in different abundance levels which are later amplified and sequenced. The more strains are present in a sample in diverse quantities, the higher is the complexity of the sample. Further, many microbes have similar genomic segments due to their common origin, conserved genes and also due to possible events of horizontal gene transfer [Ravenhall et al., 2015]. The uneven sequencing depth of reads from a sequenced metagenome, originated from variations in abundance of the members of the community, poses as a big challenge to metagenomic assemblers. Such variation affects the assembly process as a whole, especially the assembly of low abundant strains because fewer reads will cover their genomic content in high complexity and rich communities. Assembly of metagenomic samples is complex and time consuming but when done properly, provides the best resolution to analyze metagenomes with longer stretches of sequences. IDBA-UD [Peng et al., 2012], MEGAHIT [Li et al., 2015], metaSPAdes [Nurk et al., 2017] are some of the state-of-the-art metagenomic assemblers which use multi k-mer approaches to ameliorate the effects of sequencing errors and the different levels of coverage [Greenwald et al., 2017]. Recent studies suggest that, besides advances in the field, the problem of metagenomic assembly will always be complex and more attention should be given to the validation process to better select and identify errors to achieve high quality contigs [Olson et al., 2017].

Assembly is an essential step which created most of the reference sequences currently available. As a direct benefit from advances in metagenomic assembly methods, many metagenome-assembled genome sequences are being published [Mukherjee et al., 2017, Parks et al., 2017]. Those projects are bringing new sequences and diversity to current data repositories very quickly. Many environments which could not be analyzed in a reference-based manner and had almost no representation in the public repositories are now being studied based on a previous metagenomic assembly of the same studied environment [Stewart et al., 2018].

1.1.2 Binning

Sequence binning stands for the assignment of unknown fragments — reads or contigs — to a group representing a biological unit. Binning can be done reference-based (usually together with taxonomic information) by sequence similarity or sequence composition. Sequence composition methods are usually more effective for long reads or contigs since features and unique characteristics are limited in short sequences. Binning can be also performed in a reference-free manner.

Reference-free binning tools usually require long reads or contigs and use clustering methods to infer similarities among sequences and generate groups. The resulting groups are representatives of biological or taxonomic units, aiming at uniqueness per cluster (one genome, one cluster) or clusters representing an operational taxonomic unit (OTU). MaxBin2 [Wu et al., 2016] first assembles short reads into contigs and analyzes their tetranucleotide frequencies to generate clusters. When multiple samples are available, sequence composition and coverage across them can be analyzed to improve completeness and purity of clusters as shown in CONCOCT [Alneberg et al., 2014].

Reference-based binning tools, also called sequence classification tools, make use of assembled reference genome sequences to classify new sequences, long or short, into previously catalogued organisms. Taxonomic binning uses the taxonomic classification as a background structure for sequence classification. When sequence level classification is not possible, the taxonomy hierarchy can be helpful to achieve a low resolution classification. Therefore, the taxonomy provides an extended classification when a specific origin of a read cannot be determined, usually by the lack of specific references available. The taxonomic classification is helpful to explore communities in search for a specific group of organism, characteristic or function. The taxonomy also allows a simple and easy solution for ambiguities in classification, by reporting the lowest common ancestor (LCA) [Huson et al., 2007] when a fragment is classified among two different strains or taxonomic groups. Kraken [Wood and Salzberg, 2014] performs taxonomic classification by searching exact k-mer matches between reads and reference sequences. To achieve high performance, those references are indexed into a look-up table which contains the taxonomic identifier (taxid) assigned to each k-mer, or in case of ambiguities, the taxid of their LCA. That way a read is classified based on the LCA of its k-mer matches. Kaiju [Menzel et al., 2016] uses amino acid reference sequences instead of nucleotides to perform sequence binning. That allows a more sensitive classification due to the higher conservation of the amino acid sequences. For that, reads have to be translated in all six reading frames to be compared to the protein database, accounting for stop codons and frame shifts.

1.1.3 Taxonomic profiling

Differently from the taxonomic binning approach which executes a sequence-by-sequence classification, taxonomic profiling tools aim to characterize a community as a whole, providing a list of organisms or taxonomic groups present and their relative abundances. Even though both categories of tools are similar in their application, taxonomic profiling tools provide the overall composition of the sample using different techniques to infer presence of absence of taxa, for example: coverage profiles [Dadi et al., 2017], distances among identified taxa [Lindner and Renard, 2015], amount of matching reads, among others. Thus, taxonomic profiling can be understood as an extension of the taxonomic binning step, further analyzing the sample and trying to give an interpretable and accurate profile for the studied environment.

Marker gene-based methods like MetaPhlAn2 [Truong et al., 2015] and GOTTCHA [Freitas et al., 2015] profile communities with reduced databases of clade-specific or single copy genes. The reduced database size enables quick profiling, which is crucial for many applications which require fast turnaround time (e.g. outbreaks [Gardy and Loman, 2017]). However this approach can miss potentially important low abundant strains when their genome sequences are not completely covered due to low sequencing depth. In those scenarios, a whole genome-based profiler has more chances to identify such strains. SLIMM [Dadi et al., 2017], a taxonomic profiler based on read mapping classification, works with whole genome references and implements a coverage-based strategy to select better candidates for reads with multiple matches. Multiple matches are a common issue for mapping-based tools and if not properly handled can decrease the specificity in lower taxonomic ranks (e.g. species) since reads, especially short ones, will likely be classified among several similar sequences. Pathoscope [Francis et al., 2013] solves such issue with a Bayesian statistical framework, reassigning reads to their most probable origin, since one read cannot originate from more than one organism.

Taxonomic profiling and relative abundance estimation should always be interpreted and analyzed with caution. First, reference sequences and taxonomy coverage of living organisms are far from complete, meaning that the abundance estimation is only based on the previously known and available references, usually giving overestimates for the abundances of known and more studied species. Second, in a community, organisms have different genome sizes, can be actively dividing during sequencing (yielding more DNA, accumulated on the origin of replication) and can have different susceptibility to DNA fragmentation, affecting the amount of DNA sequenced. Third, quality control should always be the norm, removing sequencing errors, duplicates and low-quality bases to provide a clean and more accurate estimation [Nayfach and Pollard, 2016].

1.1.4 Taxonomy and data repositories

The underlying reference sequences and their taxonomic classification used in reference-based methods are extremely important and crucial to analyze, understand and categorize sequencing data. Both resources are the backbone of any data-based assumption and should be carefully selected and curated to scale down any possible bias.

Taxonomy is defined as the method of arrangement of biological organisms, also referred to as strains, based on shared characteristics. The taxonomic classification describes, organizes and names organisms into taxonomic units called taxa, which is the plural form of taxon. This classification is organized hierarchically, with all elements having the root of the tree as a common node and usually classified along the following main taxonomic ranks: domain, phylum, class, order, family, genus and species. Species is the only real entity while all other higher ranks are abstract [Rosselló-Móra and Amann, 2015]. Still, species definition and boundaries for microbes have been long debated and are not a defined consensus [Baltrus, 2016]. In practical terms, the taxonomy should be linked to reference sequences to bridge taxonomic classification and metagenomics. Out of a few public available and curated taxonomies [Balvočiūtė and Huson, 2017], the NCBI Taxonomy Database [Federhen, 2012] is the one of choice in most metagenomic studies, given its connection with public sequence repositories, size, curatorship, diversity and coverage of current available sequences. NCBI taxonomic definitions and assignments are updated on a daily basis. The current taxonomic classification is based on a polyphasic approach [Chun and Rainey, 2014] which account for phenotypic, chemotaxonomic and genotypic characteristics [Prakash et al., 2007, Ramasamy et al., 2014].

The backbone of the development of most DNA-related studies is the availability of sequencing data. NCBI [Coordinators, 2016], as a representative of INSDC's repositories (International Nucleotide Sequence Database Collaboration [Karsch-Mizrachi et al., 2017]), is the main source of raw data used is metagenomic projects. PATRIC [Antonopoulos et al., 2017], GOLD [Pagani et al., 2012] and environment specific projects like Metagenomics of the Human Intestinal Tract (MetaHit) [Li et al., 2014] and the Human Microbiome Project (HMP) [Turnbaugh et al., 2007] are also valuable resources [Blanco-Míguez et al., 2017]. NCBI's RefSeq [Haft et al., 2018] is a commonly used resource to obtain reference data because it provides a non-redundant set of curated sequences for transcripts, proteins and genomic regions. RefSeq is a sub-set of the GenBank [Benson et al., 2018], the primary nucleotide sequence archive at NCBI.

1.2 Current challenges in Computational Metagenomics

Despite many advances and an extensive number of tools and methods, metagenomics is still a very recent field of study with many open challenges for computational methods [Nayfach and Pollard, 2016]. Considering reference-based analysis, one of the main challenges is the exponential growth of the number of reference sequences in public repositories, especially for whole genome analysis which deals with the biggest data volume. Figure 2.1 shows the growth of NCBI repositories between 2007 and 2017. The growth is continuous and exponential and shows no signs of stopping. In a period of 2 years (2016-2017), RefSeq Microbial which accounts for Bacterial and Archaeal genome sequences, doubled in size. That means that in only 2 years the repository grew as much as the previous 12 years (2004-2015). This amount of data requires tools to be scalable and to process them in reasonable time as well as be able to incrementally update their indices and databases. Recent studies show that currently used methods which were able to deal with such data some years ago can no longer cope with complete sets of reference sequences [Breitwieser et al., 2017, Nasko et al., 2018]. That constrains tools to work with limited data sources, affecting classification sensitivity and resulting in outdated results.

Regardless of the data bonanza, there is still a lack of representation in those repositories in terms of diversity, especially for prokaryotic organisms. A very large number of organisms is completely unknown, being sometimes called microbial dark matter [Robbins et al., 2016], named after the universe's unknown dark matter. Actually, only a small fraction of the estimated number of existing microbes is already catalogued, and a very small fraction of those have sequences available [Locey and Lennon, 2016]. That invariably produces biased analysis for most of the reference-based metagenomic studies. Moreover, the currently available sequences are not evenly distributed among the tree of life. Human pathogens, model organisms and easily cultivable Bacteria are over-represented in public databases since they are the main focus of many researches hitherto. Having a well selected and curated database, keeping up-to-date with the current data releases and covering as most diversity as possible is crucial to ameliorate such problems.

The aforementioned challenges and issues can be verified in the results produced by the first CAMI Challenge (Critical Assessment of Metagenome Interpretation) [Sczyrba et al., 2017]. This challenge covered the three main categories described in this thesis: assembly, taxonomic profiling and binning. Researchers were provided with a set of real metagenomic data to analyze and report their results without any prior knowledge of its composition. As stated in their results and discussion, taxonomic based profiling and binning tools performed well down to the family rank but poorly on more refined levels (e.g. genus, species) in parts due to the lack of references currently available. The low resolution in classification depth is very harmful for specific analysis like pathogen detection where the differences between benign and pathogenic organisms are at the strain level [Loman et al., 2013, Truong et al., 2017]. In addition, taxonomic binning tools performed better in the challenge at strain level than lower taxonomic ranks like species or genus, indicating limitation of the current taxonomic content and their classification.

Taxonomy definition and classification are also open problems, especially for prokaryotic organisms, since taxonomy's early conceptions came from multicellular eukaryotes definitions. There is not a single way or a consensus on how to define the taxonomic hierarchy. Different taxonomies are available and they do not necessarily agree with each other [Balvočiūtė and Huson, 2017] since they were built using diverse methodologies. Species, the basic unit of biological diversity, is also lacking a universally accepted definition for prokaryotes [Rosselló-Móra and Amann, 2015]. The current taxonomic definition for NCBI is polyphasic and not exclusively based on whole genome sequences [Rosselló-Móra and Amann, 2015], making current metagenomic methods relying on such taxonomies prone to errors and misclassifications since they are in their vast majority purely sequence-based. This discrepancy allows same taxonomic groups to have arbitrary levels of sequence similarity. A recent study revealed that the current prokaryotic taxonomy differs in 18% from a newly proposed, purely sequenced-based taxonomy [Varghese et al., 2015]. Moreover, taxonomic-based methods are also biased due to: the lack of classification for uncultured bacteria; unclassified organisms and metagenomes; the lack of versioning for taxonomy releases; taxonomic identifiers and nomenclature changes; duplicated taxonomic identifiers for Fungi; differences between Viral and Bacterial genomes properties and classification among others [Breitwieser et al., 2017]. Many changes are ongoing to incorporate sequence information to current taxonomies and improve current discrepancies [Federhen et al., 2016] but there is still room for improvement [Garrity, 2016].

In general terms, a precise strain-level profiling is the final goal of most methods covered in this thesis: either binning sequences at strain level or accurately estimate the presence of a strain in a community sample. This level of resolution enables in-depth analysis of population genomics and microbial epidemiology [Quince et al., 2017]. Due to the fast increase of sequences in public repositories, many tools are already able to identify strains or extend taxonomic classification to assembly, genome or sequence level. However such identification is still difficult, with taxonomic misclassifications and high similarity among strains as the main causes. In addition, quantitative metagenomics should always account for compositionality [Gloor et al., 2017]. Estimated abundances extracted from one sample are not representative of the real absolute concentration when comparing multiple samples due to normalization bias. Another issue rarely described is the persistence of DNA in the environment after the death of the host cell, resulting in more genome sequences represented in the sample than the active microbial population of interest. All those factors should be accounted for and not overlooked when profiling a community.

Analysis reproducibility is a known issue in science and, not differently, in computational metagenomics. Reference-based methods are data dependent, meaning that analyses will likely change on every update of their underlying reference sequences, databases and indices. Proper metadata like version, date and source should always be provided. However, there is no standardization or guidelines on how to provide reference resources for replication of results and many publications cannot be reproduced. Additionally, as of now, no method can efficiently manage reference data updates in an iterative way, updating previously generated results and showing differences based on reference changes. On a more technical level, format specifications and standards are not a consensus. The CAMI challenge introduced formats to report sequence assembly, binning and taxonomic profiling [Belmann et al., 2015] which are very helpful resources, but not yet fully absorbed by the community.

1.3 Thesis outline

All challenges and problems presented above were motivations to develop the work presented in this thesis. Here I compile three contributions to the field of computational metagenomics described in chapters 2, 3 and 4. Those contributions are new computational methods with focus on reference-based sequence classification, taxonomic profiling and abundance estimation of organisms in environmental samples. Those methods were developed taking into account the content, growth and differences among reference sequences used and how they affect the final results. The approaches here presented make use of NCBI Taxonomy database for reference classification and nomenclature.

In chapter 2 I describe ganon [Piro et al., 2018], a taxonomic binning tool. ganon works with indices based on Interleaved Bloom Filters [Dadi et al., 2018] and classifies reads by their k-mer content. The strength of ganon lies in providing an ultra-fast indexing method for very large sets of reference sequences and a high classification precision. In this work I developed the concept of taxonomy clustering and how to use it in combination with the interleaved bloom filter. I also developed the tool, analyzed the data and wrote the paper, with helpful commentaries of all other authors. T.H. Dadi developed the interleaved bloom filter concept and E. Seiler re-implemented it. K. Reinert and B. Y. Renard conceptualized the project and gave support throughout its development.

ganon: continuously up-to-date with database growth for precise short read classification in metagenomics. Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K., & Renard, B. Y. (2018). bioRxiv, (3), 406017.
https://doi.org/10.1101/406017

Chapter 3 presents DUDes [Piro et al., 2016], a taxonomic profiling tool with a novel top-down approach to analyze metagenomic samples. Rather than predicting an organism presence in the sample based only on relative abundances, DUDes first identifies possible candidates by comparing the strength of the read mapping in each node of the taxonomic tree in an iterative manner. DUDes provides candidate strain classification while other methods would provide a more conservative identification. M. S. Lindner had the idea together with B. Y. Renard who also gave support with the statistical methods used. Both authors also gave valuable advice and provided insightful discussions for the development of DUDes. I wrote the tool and the paper and conducted the data analysis and comparisons.

DUDes: a top-down taxonomic profiler for metagenomics. Piro, V. C., Lindner, M. S., & Renard, B. Y. (2016). Bioinformatics, 32(15), 2272-2280.
https://doi.org/10.1093/bioinformatics/btw150

Chapter 4 addresses the problem of integration and execution of multiple tools and formats in metagenomics, specifically for sequence binning and taxonomic profiling (e.g. ganon and DUDes, respectively). The pipeline MetaMeta [Piro et al., 2017] executes and integrates results from several metagenome analysis tools. MetaMeta provides an easy workflow to run multiple tools with multiple samples and databases producing a single enhanced output profile for each sample. This enhanced output profile summarizes and likely improves the sensitivity and abundance estimation of the overall results based on the merged results of each executed tools. B. Y. Renard and I conceived the project and designed the methods. I developed the pipeline and M. Matschkowski led the sub-sampling analysis. B. Y. Renard and I interpreted the data and I drafted the manuscript with helpful contribution of both co-authors.

MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. Piro, V. C., Matschkowski, M., & Renard, B. Y. (2017). Microbiome, 5(1), 101
https://doi.org/10.1186/s40168-017-0318-y

Chapter 5 will summarize and conclude the impact of the contributions to the field as well as list prospects to future work.

1.3.1 Further contributions

During the development of this thesis, I contributed to side projects which are in some parts related to the presented work and resulted in valuable contributions followed by journal publications.

First, I participated in the experimental design and evaluation of the tool SuRankCo [Kuhring et al., 2015]. The tool uses random forests to predict the quality of contigs after assembly. Such ranking can help researchers to order and select the best segments of the assembly to improve follow-up results in downstream analysis.

SuRankCo: supervised ranking of contigs in de novo assemblies. Kuhring, M., Dabrowski, P. W., Piro, V. C., Nitsche, A., & Renard, B. Y. (2015). BMC Bioinformatics, 16(1), 240

In 2015, the first CAMI challenge was opened to public. This challenge, as explained in the last section, evaluated metagenome analysis tools. By the same time, I was finishing the development of DUDes [Piro et al., 2016] and participated in the category taxonomic profiling. After an extensive and thorough evaluation process, results of participant tools were published [Sczyrba et al., 2017]. DUDes performed relatively well in the challenge, being ranked among the top three best taxonomic profilers of which had the higher average of precision and recall.

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Sczyrba, A., ... Piro, V. C., Renard, B. Y., ... McHardy, A. C. (2017). Nature Methods, 14(11), 1063–1071.

During the development of MetaMeta [Piro et al., 2017] I had the intention to produce an automated and convenient pipeline which, as one of its advantages, could be easily executed in any computational environment. This requirement meets the goal of the BioConda [Grüning et al., 2018], which was recently introduced back in 2016. BioConda provides easy-to-install bioinformatics software through the package manager Conda. Many packages were included in BioConda to make MetaMeta possible and accessible to most platforms, including MetaMeta itself. My contributions to the channel made me one of the members of the Bioconda Team.

Bioconda: sustainable and comprehensive software distribution for the life sciences. Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., ... The BioConda Team, ... Köster, J. (2018). Nature Methods, 15(7), 475–476.

In the last year of my studies, a partnership project with Prof. Dr. Knut Reinert and the PhD student Temesgen Hailemariam Dadi from Freie Universität Berlin was initiated. We came together to create a solution for the recurrent issue in data analyses with huge indices and the growth of reference sequences. Such effort resulted in the Interleaved Bloom Filter data structure and also a distributed read mapper [Dadi et al., 2018]. In this work I designed the taxonomic clustering step and also contributed to the data collection and analysis.

DREAM-Yara: an exact read mapper for very large databases with short update time. Dadi, T. H., Siragusa, E., Piro, V. C., Andrusch, A., Seiler, E., Renard, B. Y., & Reinert, K. (2018). Bioinformatics, 34(17), i766–i772.

Chapter 2

Fast indexing and high precision read binning with ganon

This chapter is based on a preprint article:
ganon: continuously up-to-date with database growth for precise short read classification in metagenomics. Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K., & Renard, B. Y. (2018). bioRxiv, (3), 406017. https://doi.org/10.1101/406017

2.1 Background

Reference- and taxonomy-based short read classification is a fundamental task in metagenomics. Defining the origin of each read from an environmental sample, which can be done during [Tausch et al., 2018] or after sequencing, is usually the first step prior to abundance estimation, profiling and assembly. Many tools have been specifically developed for this task over the last years with different strategies [Oulas et al., 2015, McIntyre et al., 2017, Lindgreen et al., 2016, Peabody et al., 2015, Sczyrba et al., 2017] to achieve good performance classifying huge amount of short reads against a predefined and static set of reference sequences. Many of those use taxonomy based classifications [Balvočiūtė and Huson, 2017] providing a backbone arrangement of already obtained sequences which can be helpful for researchers to understand the composition of samples better.

Due to advances in genome sequencing, improvements in read quality, length and coverage and also better algorithms for genome assembly, the amount of complete or draft genomic sequences in public repositories is growing fast (Figure 2.1). In addition, many partial and some complete genome sequences are coming directly from metagenome-assembled genomes [Parks et al., 2017, Mukherjee et al., 2017], a technique which is boosting the growth of public repositories. This enormous amount of references poses a big challenge for current tools which, in general, were not designed to deal with such amounts of data [Nasko et al., 2018]. The problem that was at first mainly focused on the speed of short read classification is now shifting towards managing the huge reference sizes and their frequent updates [Li et al., 2018].

Figure 2.1 shows the amount of reference sequences available for the last 10 years in the GenBank [Benson et al., 2018] and RefSeq [Haft et al., 2018] repositories from NCBI. The growth is exponential. RefSeq sequences from Archaeal and Bacterial genomes are

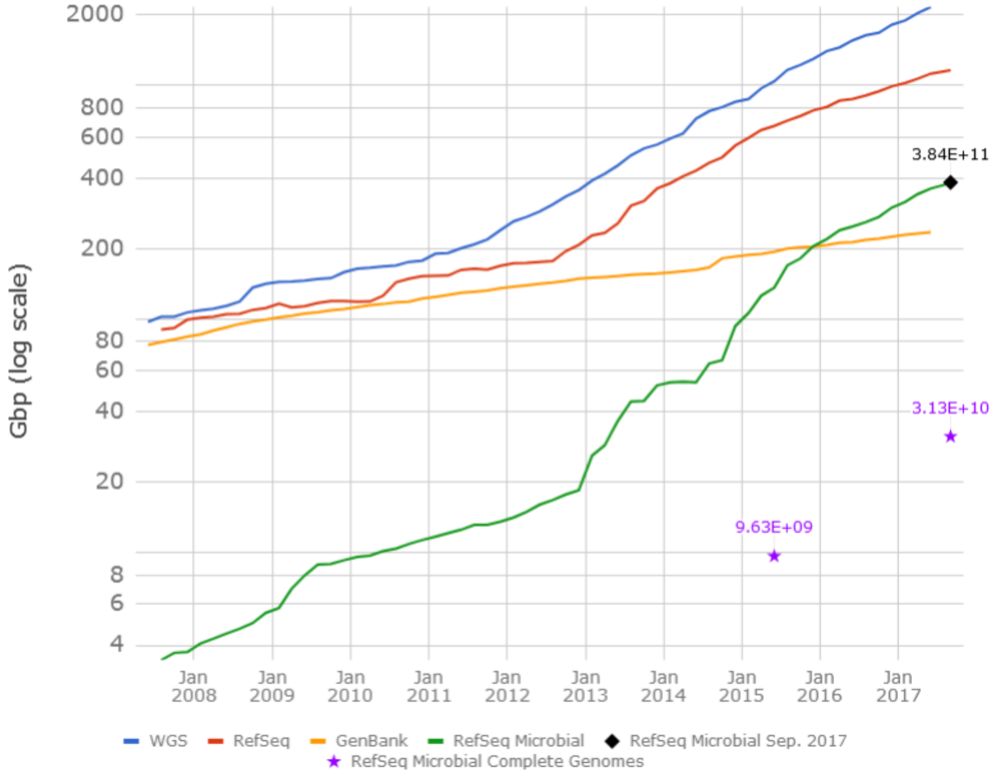


Figure 2.1: Number of available sequences in NCBI repositories from 2007 to 2017. Microbial stands for Archaeal and Bacterial organisms. RefSeq Microbial has an uninterrupted and linear growth on a logarithmic scale. Data collected from: https://ftp.ncbi.nlm.nih.gov/refseq/release/release-statistics/ and https://www.ncbi.nlm.nih.gov/genbank/statistics/

highlighted for being a commonly used reference set for metagenomic classification. In an interval of 2 years (2015-2017) the RefSeq Microbial of Complete Genomes grew more than three times, with almost two times more species contained in the most recent set. Looking at the same data point (end of 2017), the complete RefSeq Microbial had >12 times sequences and >5 times species compared to the Complete Genomes set. These data show that databases are growing fast and the variation among them is very significant. Nowadays, such repositories are too big to be analysed by standard hardware and if the observed growth continues, all this wealth of data will be constrained to just a few groups with available resources to process them.

Therefore, the choice of the database and their subsets to perform reference-based classification is an important step and known issue in metagenomics [Breitwieser et al., 2017]. As a crude rule of thumb, the more sequences the better. But even the complete set of sequences is not evenly distributed throughout the taxonomic tree, with different taxa being represented in different levels of quantity and quality. In addition, most of the taxa are still unknown and do not have any genomic sequence or entry in the taxonomic tree. This requires the tools to always keep up to date with the latest releases of public repositories, a task which is not trivial when dealing with very large amounts of sequences. Most of the tools lack the ability to update their own indices and databases and it is current that many analyses are performed in outdated resources.

The sequence classifiers MetaPhlAn [Truong et al., 2015] and Kaiju [Menzel et al., 2016] created alternatives to cover most of the diversity contained in such repositories by selecting a subset of marker genes and protein sequences, respectively. On the one hand, those methods are very powerful and provide fast and precise community profiles and read classification given their reduced index sizes. On the other hand, when analyzing whole metagenomic sequencing in complex environments, organisms with low abundance are easily missed, since their genomic content may not be completely covered. In addition, current methods which work with complete genome sequences are struggling with the current amount of available data [Nasko et al., 2018].

With those limitations in mind we developed ganon, a new reference and taxonomy-based short read classification tool for metagenomics. Ganon uses interleaved bloom filters [Dadi et al., 2018] to represent very large amounts of sequences into a searchable index. This enables the indexing of large sets of sequences (e.g. complete RefSeq) in short time and with low memory consumption, consequently improving read classification for whole metagenomic sequencing experiments. Ganon also provides updatable indices which can incorporate new released sequences in short time. The classification method, which is based on k-mer counting lemma and a progressive filtering step, improves the precision of the classification without harming sensitivity when compared to state-of-the-art tools. Ganon is open source and available at: https://gitlab.com/rki_bioinformatics/ganon

2.2 Methods

Ganon classifies reads against a set of reference sequences to find their exact or closest taxonomic origin. The method can also work in a further specialized level (e.g. assembly). An indexing step is necessary before classification, where the reference sequences will be clustered into groups based on their taxonomic classification. Ganon indices store all k-mers present in the groups of reference sequences into a specialized type of bloom filter. Once the index is created, ganon classifies the reads based on the k-mer counting lemma together with a post-filtering step providing a unique or multiple classifications for each read. Multiple classifications are solved with the lowest common ancestor algorithm [Huson et al., 2007]. In the following sections we will explain each of those steps in detail.

2.2.1 Indexing

Ganon indices are based on the k-mer content of the reference sequences, meaning it uses all possible substrings of length k of the given sequences. Instead of using standard methods for k-mer storage which can have high memory and space consumption when k is high (>15) we opted for bloom filters [Bloom, 1970], a space-efficient probabilistic data structure. Since the goal of the tool is to classify sequences based on their taxonomic origin, multiple bloom filters would be necessary to represent each distinct group of sequences belonging to a certain taxonomic level (e.g. species). Such approach provides a straightforward solution but with a high cost on the classification step by needing to compare reads multiple times against different filters. This is solved by interleaving the bloom filters (IBF), a technique previously described for the DREAM-Yara tool [Dadi et al., 2018] and also part of the SeqAn library [Reinert et al., 2017]. TaxSBP is used to separate the sequences into taxonomic groups and distribute them better into equal-sized clusters.

TaxSBP

TaxSBP [https://github.com/pirovc/taxsbp] uses the NCBI Taxonomy database [Federhen, 2012] to generate clusters of sequences which are close together in the taxonomic tree. It does that based on an implementation of the approximation algorithm for the hierarchically structured bin packing problem [Codenotti et al., 2004]. As defined by Codenotti et. al this clustering method “(...) can be defined as the problem of distributing hierarchically structured objects into different repositories in a way that the access to subsets of related objects involves as few repositories as possible”, where the objects are sequences assigned to taxonomic nodes of the taxonomic tree. Sequences are clustered together into groups limited by a maximum sequence length size of its components. TaxSBP also supports one level of specialization after the leaf nodes of the tree, making it possible to further cluster sequences by strain or assembly information which is not directly contained in the NCBI Taxonomy database (Figure 2.2). TaxSBP also supports pre-clustering, meaning that members of a certain taxonomic level can be prevented to be split among clusters. TaxSBP can further generate bins with exclusive ranks, which are guaranteed to be unique in their cluster. The tool was developed alongside the distributed indices concept [Dadi et al., 2018] and supports the update of pre-generated clusters. Since TaxSBP is based on the "pre-clustered" taxonomic tree information, the algorithm is very efficient and requires very few computational resources, being potentially useful in many other bioinformatics applications.

IBF

A bloom filter is a probabilistic data structure which comprises a bitvector and a set of hash functions. Each of those functions maps a key value (k-mer in our application) to one of the bit positions in the vector. Collisions in the vector are possible, meaning that distinct k-mers can be set to the same bit positions in the vector. Such overlaps can be avoided with a larger bitvector, reducing the probability of false positives.

An interleaved bloom filter (IBF) is a combination of several (b) bloom filters of the same size (n) with the same hash functions into one bitvector. Each i-th bit of every bloom filter is interleaved, resulting in a final IBF of size b * n. Querying in such data structure is possible by retrieving the sub-bitvecors for every hash function and merging them with a logical AND operation, which will result in a final bitvector indicating the membership for the query, as depicted in Figure 2 in the DREAM-Yara manuscript by [Dadi et al., 2018].

Aiming at the classification based on taxonomic levels (e.g. species, genus, ...) or assembly level (Figure 2.2), TaxSBP is set to cluster the input sequences into exclusive groups. That means that every group will contain only sequences belonging to the same taxon or assembly unit, but the same unit can be split into several groups. Groups are limited by a pre-defined limit of the sum of the base pair length of its elements. Sequences defined by NCBI with an accession version are the smallest subunit to be clustered.

Each of those clusters will correspond to a single bloom filter which is interleaved in

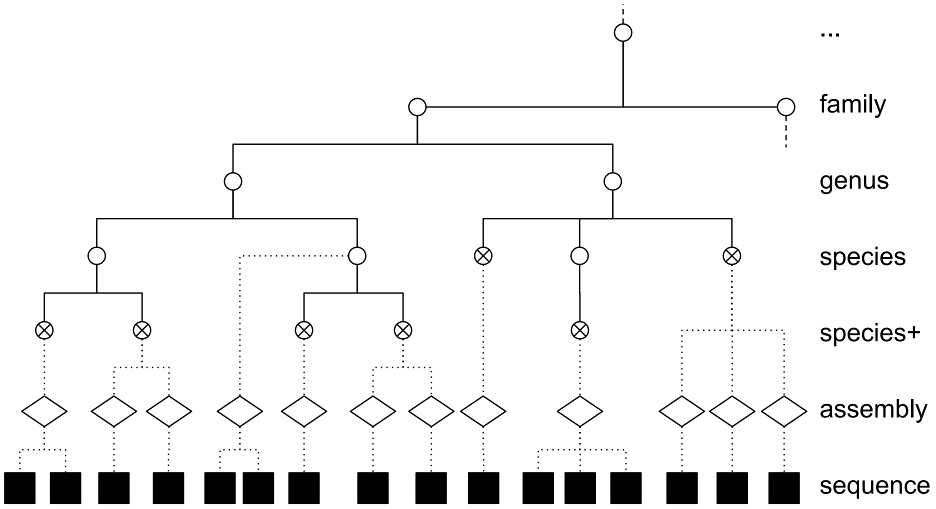


Figure 2.2: classification levels and taxonomic distribution. Empty circles are inner nodes of the tree; marked circles are leaf nodes (also referenced in this manuscript as “taxid” nodes); full lines represent taxonomic relations, dotted lines represent the extension of the taxonomic classification to the assembly and sequence level. Species+ represents all taxonomic groups more specific than species (e.g. subspecies, species group, no rank) with species in the lineage.

a final IBF. Here a trade-off between the number of groups, their maximum size and the k-mer content of each group is important. The false positive rate of a bloom filter depends mainly on its bitvector size and the number of inserted elements. In general, the more sequences a particular cluster has, the higher the number of distinct k-mers. That requires the bloom filter to be bigger to achieve low false positive rates when querying. In ganon indices, the group with most k-mers will define the size and the maximum false positive rate of the final IBF, since they have to be equal sized by definition. Thus the previous clustering step is crucial to achieve a good trade-off between the number of groups, their sizes and k-mer content. The lower the taxonomic level, the more fragmented the clustering will be, requiring bigger filters. The higher the taxonomic level, the lower the number of groups sharing k-mers, thus producing smaller filters.

The IBF has an inherent capability of updating, since it is fragmented into many subparts. Adding new sequences to a previously generated IBF is as easy as setting the bit positions of the k-mers originated from the new sequences, once we know to which cluster it should be added to. To remove sequences from the IBF, the sub-vectors affected should be rebuilt.

At the end of the build process, ganon index will consist of an IBF based on a maximum classification level chosen (taxonomic rank or assembly) and auxiliary files for the classification step.

2.2.2 Classifying

The ganon read classification is based on the well-studied k-mer counting lemma [Jokinen and Ukkonen, 1991, Reinert et al., 2015]. All k-mers from given reads are looked up on the indices previously generated. If a minimum number of mismatches between the read and the reference is achieved, a read is set as classified. Based on incremental error rates, multiple classifications for each read are filtered out and only the best ones are selected. When such filtering cannot define a single origin for a read, an optional LCA step is applied to join multiple matching reads into their lowest common ancestor node in the taxonomic tree.

K-mer counting lemma

The k-mer counting lemma can be defined as the minimum number of k-mer sequences from a read that should match against reference k-mers to be considered present in a set with a certain number of errors allowed. Given a read R with length l, the number of possible k-mers with length k in such read can be defined as:

$$kmers_R = l_R - k + 1 \tag{2.1}$$

An approximate occurrence of R in a set of references has to share at least

$$kcount_R = kmers_R - k \cdot e \tag{2.2}$$

k-mers, where e is the maximum number of errors/mismatches allowed.

Filtering

A read can be assigned to multiple references with different error rates, thus a filtering step is necessary to decrease the number of false assignments. The applied k-mer counting lemma provides k-mer counts for each read against the reference sequences. From this count it is possible to estimate the maximum number of mismatches that a read has. For example: for k=19 and length=100, a read with 50 19-mers matching a certain reference will have at most 2 mismatches. This calculation can be achieved by solving the Equation 2.2 equation for e.

Assuming that reads with fewer mismatches have a higher chance of being correct, the following filtering is applied: first, only matches against references with no mismatches are kept (all k-mers matching). If there are no such assignments, only matches with 1 error are kept. If there are none, only matches with 2 mismatches are kept and so on up to the maximum number of errors allowed (e in Equation 2.2). Similar filtration methods were previously used in read mappers such as Yara. If a read is classified in several references within the same range of mismatches, they are all going to be reported since it is not possible to define which one has higher chance of being correct based on the k-mer count information.

Ganon indices contain groups of reference sequences clustered by taxonomy or assembly group, depending on how the index was created. All k-mers from the reads are extracted and compared against such indices applying the k-mer counting lemma to select candidates, based on a user defined maximum number of errors. All matches within the error rate are filtered as described above and one or more matches are reported. Given our clustering approach, some groups can share the same identification target (e.g. one species was split in two or more groups due to a large amount of sequences). Those cases are treated specially by reporting only the match with more k-mer similarities, since they belong to the same classification group.

Ganon also provides a way to further filter the unique classifications by a different error rate for reads that matched against just one reference group. Such filter will be applied afters the standard filtration and will reclassify a low scored read to its parent taxonomic level. That is useful for a filtering in very low levels (e.g. assembly) since the classification in those levels should be more precise with less mismatches. Ganon supports classification based on multiple indices in a user-defined hierarchy.

At the end, an optional LCA method can be applied to solve reads with multiple matches with a more conservative and less precise taxonomic classification, reporting one match for each read.

2.3 Results

We evaluate ganon against a well-established method for read classification: kraken [Wood and Salzberg, 2014] and also against a new version called krakenhll [Breitwieser and Salzberg, 2018] which uses the basic kraken algorithm and also allows classification on more specific levels after taxonomic assignments (e.g. up to assembly or sequence level). We further compare the results against Centrifuge [Kim et al., 2016] which uses the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index for indexing and aims at reducing the index size by compressing similar sequences together. Here we consider only the direct read classification capabilities of the tools. Further functionalities like the estimation of a presence of a certain organism, abundance estimation or posterior assembly are not covered.

Ganon and the other evaluated tools are reference-based, meaning all classifications are made based on previously generated sequence data. The choice of the underlying database is therefore crucial. We use the same sequences and taxonomic database version for all tools when creating their indices to guarantee a fair comparison. The NCBI RefSeq repository was the chosen source of the reference sequences since it provides a curated and very extensive set of sequences. Two subsets of RefSeq were extracted: a set of complete genomes from Archaeal and Bacterial organisms dating from 26-September-2017 and a complete set of all genomes of Archaeal and Bacterial organisms dating from 25-September-2017 (Table 2.1). Taxonomic information was obtained on the same dates as the sequences.

	Base pairs	# species	# assemblies	# sequences
RefSeq CG	31.344.025.155	2.991	8.053	15.250
RefSeq ALL	383.786.471.018	15.527	95.891	9.861.599

Table 2.1: Reference Sequences used to generate indices

For the classification we use data produced for the first CAMI Challenge [Sczyrba et al., 2017] . Sets of simulated and real datasets mimicking commonly used environments and settings were obtained, representing multiple closely related strains, plasmid and viral sequences. Those samples are divided in 3 categories: low, medium and high complexity with increasing number of organisms and different sequencing profiles providing a well-produced and challenging dataset to analyse. The simulated reads were generated based on public data (NCBI, SILVA46) and an exact ground truth is provided with an origin for each read down to sequence level. The real dataset was obtained from newly sequenced genomes of 700 microbial isolates and 600 circular elements and a ground truth is provided at a taxonomic level (Table 2.2).

	simulated			real		
	# samples	Read len.	Total Size	# samples	Read len.	Total Size
low	1	2 x 100bp	15 Gbp	1	2 x 150bp	15 Gbp
medium	2	2 x 100bp	31.5 Gbp	2	2 x 150bp	40 Gbp
high	5	2 x 100bp	75 Gbp	5	2 x 150bp	75 Gbp

Table 2.2: Evaluated datasets

The classification evaluation was performed in a binary fashion. Every read has an assembly or taxonomic identification defined by the ground truth. If a certain tool gives the correct classification for a read at its original level or at a higher level, this read is marked as a true positive. For example: if a read has an assembly origin but the tool outputs it at its correct species, the read is a true positive for the species level. False positives are reads with a wrong direct classification or a false high level classification. Reads with too-specific classifications are also counted as false positives (e.g. tool classifies a read at assembly level but the ground truth only defines it at species).

2.3.1 Indexing

The set of sequences from RefSeq CG and RefSeq ALL (Table 2.1) were used as input to generate the indices for each evaluated tool. Here evaluation is done by total run-time, memory consumption and final index size (Table 2.3).

We built three different indices for ganon in different depths of classification: assembly, taxonomic leaf nodes (taxid) and species level. The size of the index was defined based on a minimum false positive rate. Ganon indices get smaller in higher taxonomic ranks, due to redundancy of sequences within the taxonomic groups. Centrifuge and krakenhll were run with default parameters. Centrifuge indices allow classification up to a taxonomic or sequence level (assembly must be deduced from sequences) and krakenhll indices allow classifications up to assembly and sequence level. Kraken, which can do only taxonomic level classification, was evaluated in two ways: default build and reduced build size limited to 64GB, a size comparable to the ganon-taxid index.

When indexing the RefSeq CG (Table 2.3, Figure 2.3), the evaluated tools took between 28 minutes and 7 hours, with ganon being the fastest and centrifuge the slowest. Building ganon indices for taxid or assembly level classification requires a higher memory usage compared to the species level due to the necessity of bigger filters. However ganon shows a

	Maximum depth	RefSeq CG			RefSeq ALL		
	of classification	time	Memory (GB)	Index size (GB)	time	Memory (TB)	Index size (GB)
centrifuge	sequence	7:02:17	283	14	Ca. 11d	0.73	211
ganon	species	0:27:47	52	48	2:55:31	0.27	259
ganon	taxid	0:28:07	69	66	2:53:32	0.35	338
ganon	assembly	0:29:56	93	88	2:54:58	0.47	451
kraken	taxid	5:21:52	177	147	Ca. 11d*	2.5*	N.A.
kraken 64	taxid	2:55:25	177	64	N.A.	N.A.	N.A.
krakenhll	sequence/assembly	5:40:47	177	147	N.A.	N.A.	N.A.

Table 2.3: Build times using 24 threads. *data obtained from Nasko, D. J (2018) based on bacterial RefSeq version 80. kraken and krakenhll indices for RefSeq ALL are not available (N.A.) due to very large memory requirements. Computer specifications: 128 logical cpus from 8 x Intel Xeon E5-4667 v4 (2,2 GHz), 512 GiB RAM, 3 TiB SSD running Debian 8.11

significant overall reduction in memory consumption and run-time compared to the other tools: 6 times faster than kraken 64, the second fastest. Centrifuge achieves the lowest index size with the cost of having the highest memory and time consumption.

Indexing the RefSeq ALL was more demanding for all tools. Centrifuge and kraken took around 11 days with high memory consumption, ranging from 730GB to 2.5TB, respectively. Due to computational limitations, we could not run kraken on our infrastructure. Here, kraken values are approximate (obtained from Nasko DJ et al) and could not be further evaluated using this reference set. Ganon built such indices in less than 3 hours, a speed-up of approximately 90 times over the other tools. Ganon also requires 2 to 7 times less memory when compared to centrifuge and kraken, respectively. Centrifuge, however, achieves the smallest index size for the RefSeq ALL data.

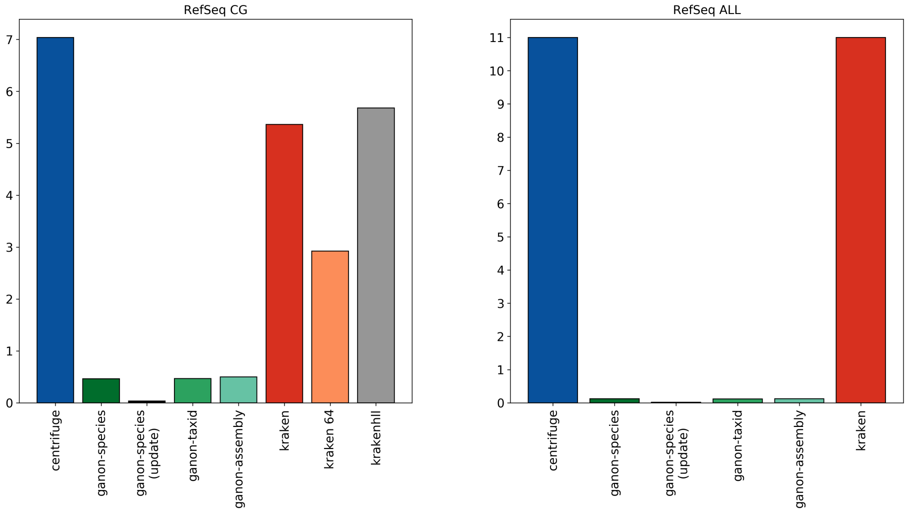


Figure 2.3: Build and update times for indexing RefSeq CG (hours) and RefSeq ALL (days)

Ganon is the only tool among the evaluated ones which allows updates on previously generated indices. We evaluated this functionality with all E. Coli sequences added to RefSeq dating from 26-September-2017 to 19-December-2017 (0.23 Gbp, 155 sequences). Updating the ganon-species index based on RefSeq CG with this dataset took 2 min 11 seconds with the same memory requirements for building the index (Table 2.3). On RefSeq ALL, ganon took around 28 minutes to update the index.

2.3.2 Classifying

The classification results were evaluated in terms of sensitivity and precision at each taxonomic level based on the binary classification. The values are always cumulative up to the evaluated level, for example: sensitivity and precision at the genus level will consider all the classifications up to genus level (species, species+ and assembly). Every taxonomic classification in between and different than the predefined ranks (superkingdom, phylum, class, order, family, genus and species) are counted towards the closest high level parent rank in its lineage.

The results for the CAMI simulated and real datasets can be interpreted in groups (Table 2.4). First, by database: all evaluated tools for the RefSeq CG and only ganon and centrifuge for RefSeq ALL, since kraken and krakenhll requirements for such dataset were not practicable in our computational environments. Second, by tool considering their maximum depth of classification: centrifuge-taxid, ganon-taxid, kraken and kraken 64 classify reads up to the taxonomic leaf nodes (taxid). ganon-assembly, centrifuge-assembly and krakenhll up to assembly level. ganon-species was also included to show how the results changes when classifying at a higher level with a small index. Given the availability of the ground truth, simulated data was evaluated up to assembly level while real data was evaluated up to species+ level. Centrifuge outputs at sequence level, thus an extra step of applying a LCA algorithm for non-unique matches was necessary to generate results at assembly and taxid levels.

		simulated		real	
depth of classification		RefSeq CG	RefSeq ALL	RefSeq CG	RefSeq ALL
species	ganon-species	✓	✓	✓	✓
taxid	centrifuge-taxid	✓	✓	✓	✓
	ganon-taxid	✓	✓	✓	✓
	kraken	✓	✗*	✓	✗*
	kraken 64	✓	✗*	✓	✗*
assembly	centrifuge-assembly	✓	✓	✗**	✗**
	ganon-assembly	✓	✓	✗**	✗**
	krakenhll	✓	✗*	✗**	✗**

Table 2.4: Comparison groups. *) kraken and krakenhll indices for RefSeq ALL are not available. **) real reads do not contain assembly level ground truth data

Figure 2.4 compares the results for one of the simulated high complexity datasets using the indices based on RefSeq CG and RefSeq ALL. With RefSeq CG all tools performed similarly in terms of sensitivity within their depth of classification, with ganon-species achieving a better classification at maximum species level. In terms of precision, ganon shows superior performance in every scenario when compared to its competitors. At assembly level ganon has approximately half of the false positives of krakenhll and centrifuge while having a very similar number of true positives. Using the indices based on the RefSeq ALL both ganon and centrifuge show big improvements compared to the RefSeq CG, given the broader diversity covered by this set. In terms of sensitivity both tools performed once more very similarly, while ganon achieved better precision with very few false positives. Ganon-species had the highest precision of approximately 97% at species level.

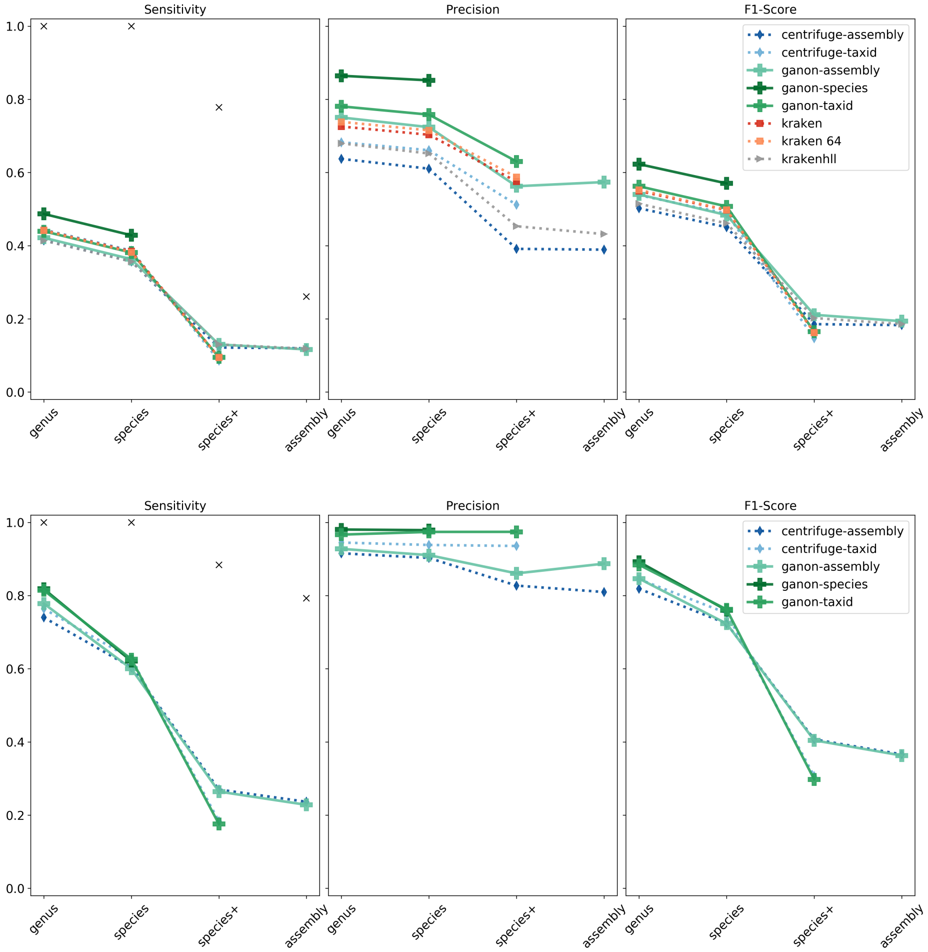


Figure 2.4: Simulated reads - high complexity sample 1. Results on the top with RefSeq CG and on the bottom with RefSeq ALL. Results up to assembly level. Black markers show the maximum sensitivity possible given the ground truth classification.

The same analysis was performed on real data. This set is more challenging, since most of the organisms are novel and not yet present in the public repositories. As stated by the CAMI results [Sczyrba et al., 2017], most tools performed poorly in this dataset in terms of sensitivity, as depicted in Figure 2.5. The results follow the same trend from the previous analysis, with both ganon and centrifuge being similarly sensitive while ganon being more precise with less false positives. Here the use of a larger set of references significantly improved the results in terms of sensitivity and precision, demonstrating that the use of more references is crucial for newly sequenced environments. Ganon and centrifuge enable such analysis. In addition, our evaluated indices were based only on Archaeal and Bacterial genome sequences, while the real dataset also contains viral and other elements, explaining part of the bad performance of the tools.

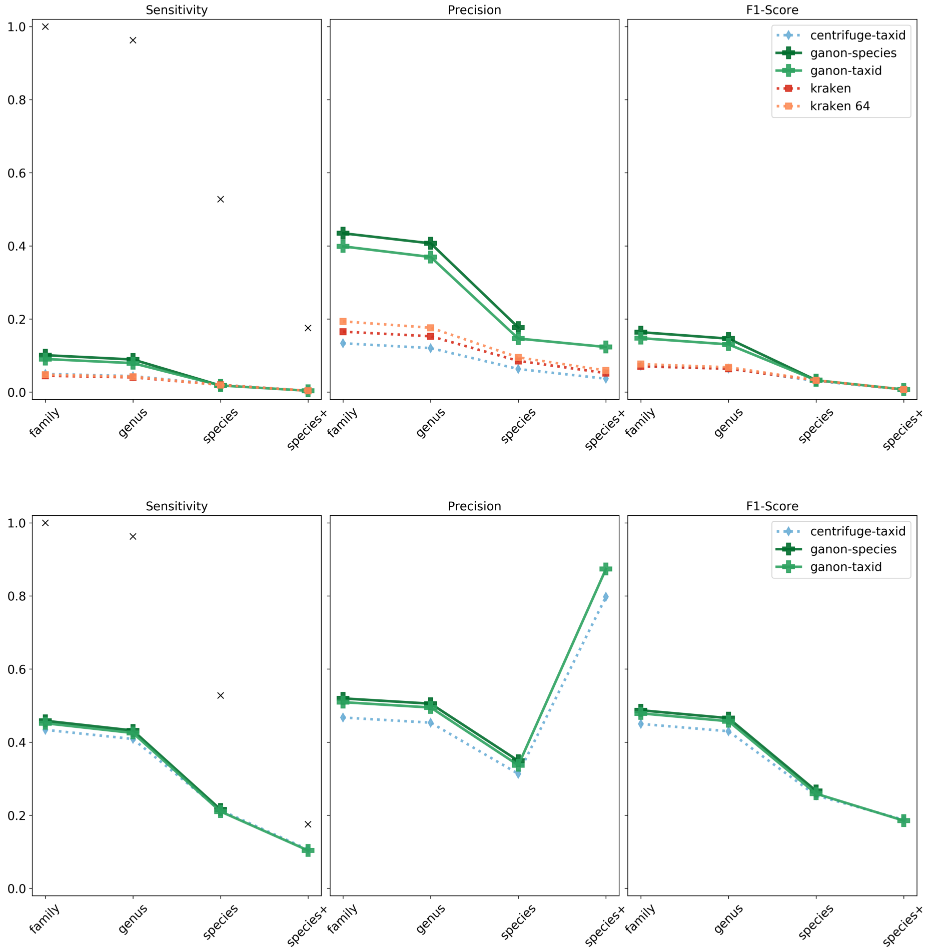


Figure 2.5: Real reads - high complexity sample 1. Results on the top with RefSeq CG and on the bottom with RefSeq ALL. Results up to species+ level. Black markers show the maximum sensitivity possible given the ground truth classification.

Table 2.5 compares the performance of the analysed tools in terms of how many base pairs they can classify per minute. Memory usage is mainly based on the index size of each tool, with little variation besides that. Kraken is the tool with the fastest performance on the classification when using the RefSeq CG indices, followed by ganon-species and centrifuge. When classifying against RefSeq ALL, centrifuge has the fastest classification time followed by ganon-species. The slow down on bigger indices is reasonable, given that

RefSeq ALL indices cover 5x more species, 11x more assemblies, 12x more base pairs and 646x more sequences than the smaller reference set (Table 2.1). Ganon classification time is around 5 times slower on species and taxid level compared to the RefSeq CG. Centrifuge performs around 3x slower in this reference set. A trend in the classification performance is clear, the bigger the index the slower the classification. This trend is evident when using very large indices based on the RefSeq ALL.

	RefSeq CG	RefSeq ALL
centrifuge	428*	150*
ganon-species	481	80
ganon-taxid	379	71
ganon-assembly	324	30
kraken	952	-
kraken 64	1057	-
krakenhll	465	-

Table 2.5: Classification performance of 1 million reads in Mbp/m with 48 threads. * performance estimation based on “Multiseed full-index search” time report without considering post-processing time. Computer specifications: 128 logical cpus from 8 x Intel Xeon E5-4667 v4 (2,2 GHz), 512 GiB RAM, 3 TiB SSD running Debian 8.11

2.4 Discussion

We presented ganon, a novel method to index references and classify short reads for environmental samples based on a taxonomic oriented set of references. Ganon strength lies in providing an ultra-fast indexing method for very large sets of reference sequences and a high classification precision. That is possible due to a novel application of interleaved bloom filters with a k-mer counting and filtering scheme. This is relevant in a time where the number of available sequences keeps increasing daily and current tools struggle to build and update such indices.

Ganon indices are also flexible and can be built for high taxonomic levels, requiring less space and memory, consequently improving classification speed. Evaluations show that indices built to classify at maximum species level had the best results in terms of sensitivity and precision. A trade-off between filter size and false positive rate is also possible, sacrificing precision over performance.

Ganon classification results are on par with state-of-the-art methods with regard to sensitivity, while improving precision rates in every scenario of our evaluations. We attribute such improvement to an application of the k-mer counting lemma together with a progressive filtering step, which can better separate false from true positives.

Even with ganon achieving improved results, in general terms, short read classification tools tested here perform similarly when based on the same underlying set of reference sequences. In addition, the more complete the reference set, the better the classifications. The difference in sensitivity when using RefSeq ALL compared to only complete genomes is very significant and tends to get even bigger with more sequences added to this repository.

Thus the choice of the database is crucial and should not be overlooked when analysing metagenomic data. Even though centrifuge performed well with more reference sequences, the time to index such set is highly prohibitive. In addition the method does not provide a direct one-match-per-read result and an in-house post-processing step was necessary to achieve it. The indexing issue also applies to kraken and krakenhll. Ganon manages to index big set of references sequences in very short time and classifications results are as good or better than the evaluated tools. To the best of our knowledge, Ganon is the only tool with update capabilities, which is performed in a fraction of the complete build time. That poses an advantage to keep up to date with the public repositories and even each single of their frequent updates.

To conclude, we believe that ganon can be a useful tool for metagenomics analysis in a time where reference sequence repositories are growing fast. Ganon provides fast indexing, updatability, competitive results and improved precision for short read classification.

Chapter 3

Taxonomic profiling and abundance estimation with DUDes

This chapter is based on a published article:
DUDes: a top-down taxonomic profiler for metagenomics. Piro, V. C., Lindner, M. S., & Renard, B. Y. (2016). Bioinformatics, 32(15), 2272-2280. https://doi.org/10.1093/bioinformatics/btw150

3.1 Background

The fast increase of complete genome sequences available on public databases has allowed better predictions of the microbial content from sequenced environmental and clinical samples. In addition, the fast evolution and decreasing costs of high-throughput sequencing as well as the development of fast and precise bioinformatics tools to handle huge amounts of data (e.g. read mappers, assemblers) are enabling the integration of automated computational methods in clinical practice [Köser et al., 2012, Pallen, 2014].

Taxonomic or community profiling are common terms to define the process of identification of organisms and their quantification given a targeted or whole shotgun metagenomic sequencing [Mande et al., 2012]. Characterizing the taxonomic diversity is an initial and fundamental step to understand complex biological processes, diversity and functions of a microbial community and it can be applied for pathogen detection studies. Several tools have been recently developed for this characterization, with different approaches and applications [Lindgreen et al., 2016]. Considering only the reference-based methods, that is, methods that use reference sequences to guide their analysis and classify sequences in the sample, community profiling is categorized in two sub-groups: composition and similarity-based. Composition-based methods extract information from the composition of sequences from both sample and database (e.g. GC content, codon usage) and search for similarities between reads and references. Similarity or homology-based techniques are built on mapping or aligning the sequenced reads against reference databases and performing further analysis. Despite significant performance improvements in the last years, similarity analysis is still computationally challenging due to the extremely high throughput of modern sequencing machines and the accelerated growth of available genomic sequences [Benson et al., 2012]. Some tools address this problem by reducing the database space, selecting only marker genes or a specific subset of sequences [Freitas et al., 2015, Segata et al., 2012]. This simplification can speed up analysis but at the same time reduces the complexity and diversification of the references, decreasing precision for more specific identifications. Other methods use custom databases of partial or whole genome sequences [Huson et al., 2007, Francis et al., 2013, Lindner and Renard, 2015]. Additionally, very efficient k-mer based read binning tools [Wood and Salzberg, 2014, Ounit et al., 2015] can also be used in some extension for profiling communities, by selecting the targets with more associated sequences.

Despite the remarkable recent advances in this area and a vast number of tools available, there is still a number of challenges from sequencing methods to bioinformatics tools to integrate computational and automated approaches into molecular and metagenomics diagnostics [Klymiuk et al., 2014, Fricke and Rasko, 2014]. Specifically for community profiling, there is still room for improvement in species and strain level detection, which can have very similar genomic content and at the same time low abundances. The high discordant number of available sequenced genome sequences among several taxonomic groups poses another common problem, where some organisms (e.g. pathogens, model organisms) are more studied and therefore overrepresented in the public sequence databases.

Aiming to solve those limitations, we propose a new method: DUDes, a reference-based taxonomic profiler that introduces a novel top-down approach to analyze metagenomics NGS samples. Rather than predicting an organism presence in the sample based only on relative abundances, DUDes first identifies possible candidates by comparing the strength of the read mapping in each node of the taxonomic tree in an iterative manner. Instead of using the lowest common ancestor (LCA) [Huson et al., 2007], a commonly used bottom-up approach to solve ambiguities in identifications, we propose a new approach: the deepest uncommon descendent (DUD). While the LCA method solves ambiguous identifications by going back in the taxonomic tree to the lowest common ancestor, the DUD approach starts at the root node and tries to go for deeper taxonomic levels, even when ambiguities are found. That way it is possible to have less conservative identifications in higher taxonomic levels. Besides, when the provided data does not allow a specific identification on higher levels, our method identifies a small set of probable candidates among dozens of possibilities (e.g. instead of stopping at species identification, DUDes will provide 5 highly likely strains out of 150). Permutation tests are performed to estimate p-values between taxonomic groups and to identify the presence of them on each level. We show in experiments and comparisons with state-of-the-art tools that DUDes works well for single and multiple organism detection, can handle unequally represented references in the database and identifies low abundant taxonomic groups with high precision. DUDes is open source and it is available at http://sf.net/p/dudes

3.2 Methods

The DUDes workflow requires three main inputs: a set of reads, references sequences and a taxonomic tree structure (Figure 3.1). By mapping the reads against the reference sequences, a SAM file [Li et al., 2009] is created. The linkage between the reference sequences and the taxonomic information is performed by DUDesDB and stored in a database file. Both files serve as input to DUDes profiler which performs an iterative top-down analysis on the taxonomic tree structure.

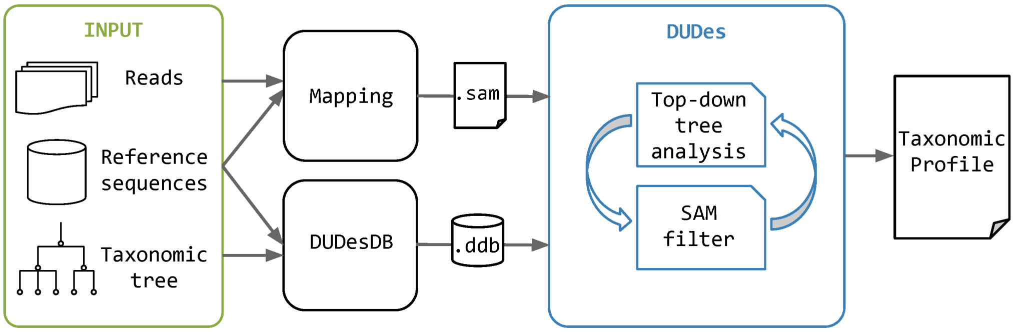


Figure 3.1: DUDes workflow. A set of reads, references sequences and taxonomic tree structure are the pre-requisites for the complete workflow. A SAM file from mapping the reads against references sequences and a database file generated by DUDesDB are the input files required for DUDes profiler, which will perform the top-down tree analysis and generate the taxonomic profile.

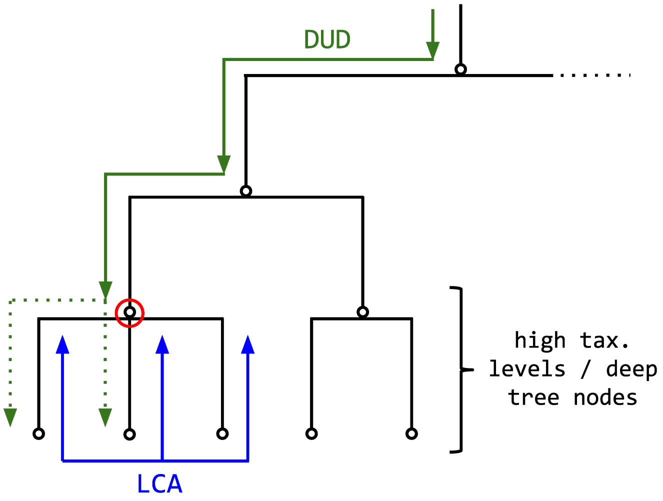
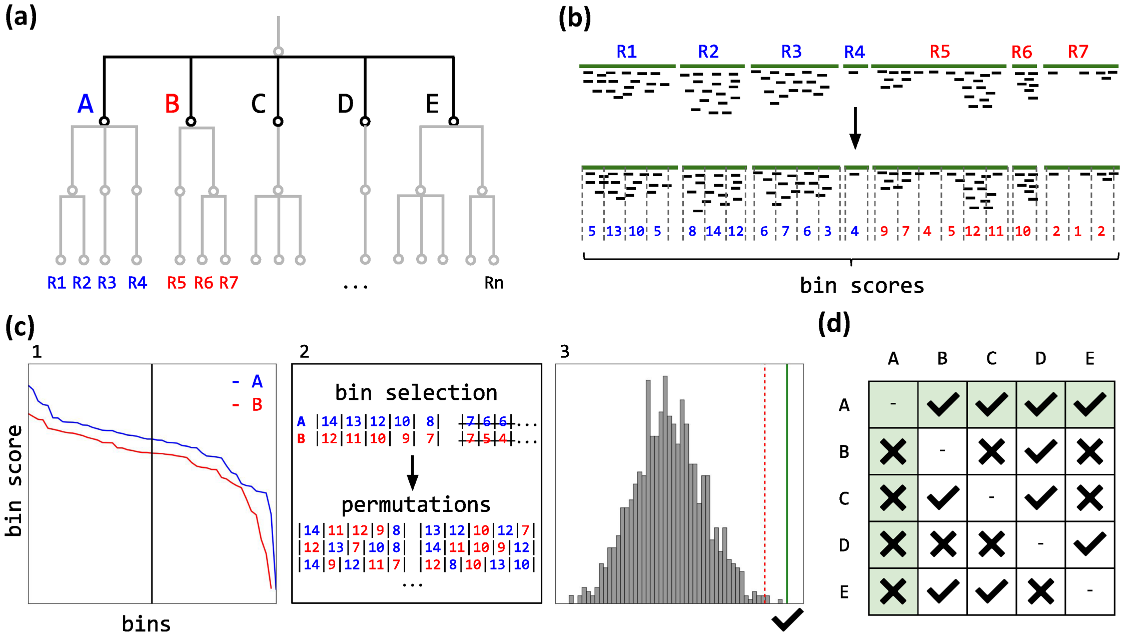


Figure 3.2: Comparison between DUD and LCA. DUD is a top-down approach that starts from low taxonomic levels. LCA is a bottom-up method that solves ambiguities from high taxonomic levels. The node marked in red is considered the lowest common ancestor for its children nodes. Dotted arrows represents how DUD can be more specific in higher taxonomic levels pointing to candidates nodes while LCA is more conservative and goes back to the lowest common ancestor.

The first step of the DUDes algorithm is to assign a set of reference sequences and read matches for each node of the taxonomic tree (e.g. the root node contains all sequences and matches, while the *Escherichia* node (genus) will contain only the sequences corresponding to the organisms that belong to this classification and their respective matches). The general idea is to start at the root node and go deeper into the tree (Figure 3.2), evaluating all taxonomic levels and identifying nodes with significant matches that are the ones that will be considered present in the sample. The presence of a node in a certain taxonomic level is defined by the following steps (Figure 3.3):



	A	B	C	D	E
A	-	✓	✓	✓	✓
B	✗	-	✗	✓	✗
C	✗	✓	-	✓	✗
D	✗	✗	✗	-	✓
E	✗	✓	✓	✗	-

Figure 3.3: Steps for identifications on a taxonomic level. (a) Taxonomic tree structure with the tested level in black and the reference sequences of each node (R1-Rn). Nodes A and B are the first ones to be compared against each other (b) Bin generation: reference sequences (green) and their respective read matches are subdivided into bins. For each bin, a match-based score is calculated (c) p-value estimation: 1 - Bin scores of the node A (blue, references R1-R4) are compared against bin scores of node B (red, references R5-R7). The black line represents the cutoff value based on the top 50% of bin scores from node A. 2 - Only the best bins are selected based on the cutoff. Permutations are then performed among the selected bins. 3 - The distribution is generated by the randomly permuted difference of means between bin scores. The red dotted line represents the critical value and the green the observed difference of means. In this example node A is therefore significant against node B (p-value $\leq$ critical value). (d) Identification: steps b and c are then repeated for all pair of nodes. In this example, node A is identified because it was the only significant node against all the others (rows) and any other node could be significant against it (columns)

3.2.1 Bin generation

First, we create a set of bins for each node of the tree. Bins are sub-sequences from the reference sequences assigned to the node (Figure 3.3b). They are non-overlapping and equal-sized. The bin size is fixed for all nodes and it is defined by default as the 25th percentile of the sequence lengths of the whole reference database. The bin size should be a balance point between the number of small sequences in the database and speed requirements: the smaller the bin size, the more calculations are needed. The larger the bin size, the bigger the chance that higher nodes lack of full bins. In our analysis the 25th percentile provides a good trade-off between those, ensuring that the majority of reference sequences will have at least one full bin. Each one of these sub-regions will have a bin score based on read matches, defined as the sum of all match scores. The match score ms is defined for each match m as:

$$ms_m = l - e \tag{3.1}$$

where l is the length and e the edit distance of the match (minimum number of edit operations — insertions, deletions, and substitutions of letters — transforming one sequence into another [Levenshtein, 1966]). Bin scores do not only take the number of matches to the bins into account, but also the number of mismatches and indels in those sub-regions.

3.2.2 p-value estimation

DUDes performs a pairwise comparison between all nodes on a taxonomic level, estimating a p-value for each pair with permutation tests. The permutations occur between the nodes' bin scores. Only bins with scores higher than zero are considered. Additionally, only the best bins are permuted: a cutoff is chosen, based on the number of bins representing the top 50% scores of the main node (Figure 3.3c-1). This cutoff is useful in order not to prioritize nodes with larger or more references sequences, meaning that the comparison of a certain node X against node Y can have a different cutoff value from the comparison of the same node Y against X. For example: consider a node X with 100 bins and a node Y with only 70 bins. When comparing X against Y, only the first 50 bins from node X are going to be compared against the first 50 bins of node Y. When comparing Y against X, only 35 bins from both nodes are going to be considered (Appendix A - Figure 1). When one of the nodes does not have enough bins to be compared, the cutoff will be reset to the total number of bins of the node with fewer bins (Appendix A - Figure 2). The cutoff is also useful to remove poorly mapped bins with low scores and to normalize the number of bins between the two compared nodes, allowing a fair comparison between them. A limit of 5 bins is required for each node to allow a significant permutation. Permutations between the selected bin scores are performed 1000 times by default. The values are randomly shuffled and separated in two groups based on the cutoff value. The random difference of means between the groups is calculated, generating a distribution like the one shown in Figure 3.3c-3. A one-sided p-value is then estimated based on the observed difference of means between their actual bin scores (Appendix A - p-value estimation).

The estimated p-value is considered significant if it is lower than a certain critical value. Since many hypotheses are being tested, multiple testing correction is necessary to control the type I error, that is, the risk of falsely rejecting a hypothesis that is true [Goeman and Solari, 2014]. Here we applied two methods: Bonferroni correction locally (taxonomic level) and the Meinshausen procedure globally (tree level) [Meinshausen, 2008]. Two methods were applied together because multiple tests occur in two ways: several taxonomic levels on the tree and several nodes for each taxonomic level are tested. The Meinshausen procedure was chosen for being a hierarchical approach that can achieve larger power for coarser resolution levels. It was applied to control the multiple testing error generated by several comparisons on each taxonomic level of the tree. Bonferroni correction was applied locally to correct for the multiple tests performed among the nodes in a taxonomic level. While the Bonferroni method is highly conservative in general, this is not critical in our application since the number of nodes on a taxonomic level is usually fairly small. The critical value cv for each node n is calculated as:

$$cv_n = \frac{\alpha}{N-1}\frac{L_n}{L} \tag{3.2}$$

where α is the significance level threshold (default 0.05) and N is the total number of nodes in the tested taxonomic level (minus itself), comprising the Bonferroni correction. Additionally, L is the total number of leaves of the tree and L_n the total number of leaves below node n. They stand for the Meinshausen procedure correction. All nodes are tested against each other at the same level and, if the p-value of the comparison is below the critical value, the comparison is set as significant as pictured in the table in Figure 3.3d.

3.2.3 Identification

After all nodes within a taxonomic level have been compared against each other, they are evaluated using two criteria: how many times they are not significant and how many times other nodes are significant against them (columns and rows in Figure 3.3d, respectively). The number of occurrences of those two metric are summed for each node and DUDes selects the one with the minimum value as identified, therefore present on the sample. In a metagenomic dataset it is possible that more than one node is identified at once, when their bins have similar abundances. In this case, two or more nodes with sufficient number of bins and matches that could not be significant against each other are going to be identified concurrently. That means that more than one node is going to be considered present in the current taxonomic level, leading to more than one path on the following tree analysis.

The three-step-algorithm (consisting of bin generation, p-value estimation and identification) continues to the next taxonomic level only for the children of the identified nodes. The process is iterated until it reaches the leaf nodes of the tree (here considered the deepest possible uncommon descendent — at species level by default) or until no more identifications are possible, due to lack of minimum matches or bins support. Here we can point out an advantage of the DUD method over the LCA. When the LCA is applied, the results tend to be very conservative because it solves ambiguous identifications by going back one taxonomic level to the lowest common ancestor. Instead, the DUD approach will always try to go for a deeper taxonomic level, even when ambiguities are found (Figure 3.2). That way it is possible to have identifications in higher taxonomic levels. Besides, when the provided data does not allow a specific identification on higher levels, it is still possible to propose a set of likely candidates based on the concurrent identification, being more specific than going back in the taxonomic tree.

At the end of the tree iteration, one or more paths on the tree and their leaf nodes are identified as candidate taxonomic groups to be present on the sample. Since metagenomic samples can contain hundreds to thousands of organisms [Handelsman et al., 1998], a filtering step is performed to remove identified matches and allow more iterations. DUDes perform this step by filtering out the direct matches on the identified candidates' references sequences. Furthermore, all matches from the reads that had at least one direct match are analyzed. If those read matches have a match score lower than the direct match, they are considered indirect matches, and are filtered out as well. With this new set of matches, a new iteration is started from the root node. Several iterations are performed until the number of matches is below a certain threshold or until all matches were filtered.

At the end, a relative abundance value is calculated for the final candidates. These are based on the direct matches of the identified leaf nodes and normalized by the length of their respective reference sequences. Each identified leaf node n has an abundance ab calculated as:

$$ab_n = \sum_{i=1}^{r} \frac{\sum_{j=1}^{t} ms}{l} \tag{3.3}$$

where r is the number of references sequences belonging to the node n, t are the matches belonging to the reference i, ms is the match score and l is the length of the reference i. The abundance of the parent nodes are based on the cumulative sum of their children nodes' abundance.

DUDes outputs a file with a set of final candidates in the BioBoxes Profiling Output Format v0.9.3 (https://github.com/bioboxes/rfc/blob/master/data-format/profiling.mkd). When strain identification is selected, DUDes outputs an additional file with all identified strains and their relative abundances.

3.2.4 Strain identification

Optionally, DUDes will try to extend the species identification and provide a set of probable strains present in the sample. The process of strain identification works identically as the three-step-algorithm but starting the analysis from each one of the identified species nodes. Sequences among strains usually have high similarity in their composition. This makes the identification process more challenging. For that reason we implemented a post-filtering process to better select a candidate strain. Given a set of identified strains by the three-step method, we choose one representative candidate, which has the maximum summed value of match scores in this set. Alternatively we provide a second output, reporting all other strains identified and their relative abundance.

3.2.5 DUDesDB

DUDesDB pre-processes the taxonomic tree structure and the reference sequences, generating a database file to be used by DUDes profiler. The current version of DUDesDB supports the NCBI taxonomic tree [Coordinators, 2015] and uses the GI to make the link between reference sequences (fasta files) and tree nodes (nodes.dmp and gi_taxid_nucl.dmp files). Since the *strain* level is not directly defined in the NCBI taxonomic tree structure, we considered any unclassified node with the tag *no rank* after the species level as a strain node.

3.2.6 Mapping

DUDes can handle multiple matches and account for the mapping quality with match scores, improving its identification capabilities. By default, the number of allowed matches should be as high as possible, allowing all matches when feasible. Since that can be computational impracticable, we used a default value of 60 matches for each read. Other mapping parameters can be found in the Appendix A - Tools and parameters.

3.2.7 Experiments

DUDes evaluation was performed in four distinct datasets: two synthetic communities and two real metagenomic samples.

First we analyzed synthetic metagenomic data with available ground truth to evaluate how precise is our identification method. We chose a common set for metagenomics evaluations — the Human Microbiome Project (HMP) mock community [Turnbaugh et al., 2007], an *in vitro* synthetic mixture of 22 organisms (20 Bacteria, 1 Archaea and 1 Eukaryote) that mimics errors and organism abundances from real metagenomics samples. Only Bacteria and Archaea were considered in this evaluation. Further, this set was also divided in sub-sets of different percentages of reads and compared against other metagenomics analysis tools: kraken [Wood and Salzberg, 2014], GOTTCHA [Freitas et al., 2015] and MetaPhlAn2 [Truong et al., 2015]. They were chosen for having rather different approaches to solve the taxonomic profiling problem and for having good results in recent metagenomics studies [Lindgreen et al., 2016]. Kraken is a read binning tool that uses a k-mer approach to classify each read in a given sample with focus in high performance. GOTTCHA is a taxonomic profiler that uses non-redundant signature databases and aims for lower false discovery rates. MetaPhlAn2 relies on a curated database of approximately 1 Million unique clade-specific marker genes for profiling metagenomic samples. A second synthetic community consisted of 64 laboratory-mixed microbial genomic DNAs was also evaluated [Shakya et al., 2013]. This community made of organisms of known sequences has a very broad diversity among bacteria and archaea and a wide range of genetic variation at different taxonomic levels. At the same time, this dataset provides a large number of sequenced reads (approximately 110M), allowing a more realistic performance evaluation.

We also applied DUDes to real metagenomic samples of gut microbiomes from the outbreak of Shiga-toxigenic *Escherichia coli* (STEC) in Germany in 2011 [Loman et al., 2013]. With this dataset we evaluated how well DUDes performs in a real scenario to profile a pathogenic sample, and compared the results with the previously known experiments. Furthermore, we evaluated this set based on previous known information (e.g. lab experiments, other tools based on the LCA approach) performing a more specific profiling. Lastly, we profile a marine dataset from Tara Oceans [Sunagawa et al., 2015] with Bacteria, Archaea, Virus and Eukaryotes present on the sample, showing the versatility of the tool. Datasets' details are shown in Appendix A - Table 1.

The HMP and STEC samples were pre-processed with the digital normalization algorithm [Brown et al., 2012] for decreasing sampling variation and for error correction. In both analysis, the reference database for DUDes and kraken was generated with the set of complete genomes sequences (Bacteria and Archaea) together with the taxonomic tree structure, both from NCBI [Coordinators, 2015] from 26-Mar-2015. The 64-organim set was used without any filter. For this set we used the above database with the addition of 4 non-complete genome sequences [taxid: 901, 52598, 314267, 304736] to have all species in the sample available. For the Tara dataset we made a custom database, containing only expected marine organisms. Bacterial, Archaeal and Viral taxons were obtained from the references sequences used in the Tara Oceans Project [Sunagawa et al., 2015] and the

Eukaryotic set of taxons were obtained from the MMETSP [Keeling et al., 2014]. All NCBI refseq sequences relative to those taxons were collected to generate the database (from 31-Jan-2016). For MetaPhlAn2 and GOTTCHA (all sets) we used their provided database, v20 and v20150825, respectively. Bowtie2 [Langmead and Salzberg, 2012] was used for read mapping. Parameters and usage details of each tool can be found in the Appendix A - Tools and parameters.

We evaluated the output from each tool based on a binary classification of the sorted taxonomic profile. The binary classification is valid for taxonomic groups (TG) of a certain taxonomic level. True positives are all TG present in the sample and correctly identified, false positives are the identified TG known to not be present in the sample, false negatives are the TG that are present but could not be identified and true negatives are all TG on the database not identified and known not to be present in the sample.

3.3 Results

HMP mock community

We first assessed DUDes' taxonomic profiling capabilities with the set of Illumina reads from the HMP staggered mock community. Digital normalization was applied to correct errors, reduce data size and decrease sample variation. The normalization reduced around 18% from the complete set of 7.932.819 reads. Then, we mapped this normalized set against the reference database using Bowtie2 and the resulting mapping file was used in DUDes to profile the mock community sample.

The performance in terms of sensitivity and specificity for the identifications at each taxonomic level is shown in Table 3.1. DUDes successfully identified 20 of the 21 known species present in the sample (Bacteria and Archaea). *Actinomyces odontolyticus* does not have reference sequences in our database, therefore it could not be identified directly as species. However, at lower levels — Actinomycetales (order), Actinobacteria (class), Actinobacteria (phylum), Bacteria (superkingdom) — all present taxonomic groups could be identified, reaching a true positive rate of 1. Even in deeper levels — family, genus and species — DUDes achieved a sensitivity level of approximately 0.95. Furthermore, high specificity values show that DUDes can precisely select organisms present in a given metagenomic sample even with a large number of references sequences in the database.

	s.kingdom	phylum	class	order	family	genus	species
Sensitivity	1	1	1	1	0.944	0.944	0.952
Specificity	1	1	1	0.973	0.987	0.991	0.994

Table 3.1: DUDes' sensitivity and specificity values for of the HMP mock community evaluation

We further analyze the same dataset, comparing the results against the established metagenomics tools kraken, GOTTCHA and MetaPhlAn2 — at species level. In addition, we evaluated random sub-sets of the normalized set of reads, in a range from 1 to 100%. The sub-set analysis allowed us to show how the tools perform over a wide range of coverages: from the set where all data is available and each organism is well represented to a scenario where very few reads are present for some low abundant organisms. All results were based on the ordered output provided by each tool. Kraken is a read binning tool and outputs an unsorted report with all identified organisms. From this output we selected only species level identifications and sorted the candidates by the percentage of reads assigned. We limited the output from all tools by the first 30 entries when necessary.

Figure 3.4 shows ROC curves comparing the output from each tool on three sub-sets: 1%, 50% and 100% (all sub-sets in the Appendix A - Figure 3). DUDes and kraken had stable and similar performances in all sets, with DUDes achieving the best AUC values. When only 1% of the reads was used, MetaPhlAn2 and GOTTCHA did not achieve good results. That can be explained by the fact that both tools use specific gene/signature databases, decreasing the chance of finding matches when the number of reads is very low. Both DUDes and kraken use the whole genome sequence as database and therefore achieved good predictions, even at extremely low coverages. DUDes could successfully identify *Bacteroides vulgatus* at species level although there were only 14 matches on the references. Since kraken is not a taxonomic profiler, it outputs each organism that has at least one read, resulting in a long tail of false positives. DUDes produced fewer false positives. With more reads, all tools improved their results (Figure 3.5) , with DUDes being slightly superior, followed by kraken and GOTTCHA. Both DUDes and kraken identified 20 of 21 organisms present in the sample using only 10% of the reads, keeping the results stable for the next sub-sets. GOTTCHA needed at least 50% of the data to reach the same level of identification. MetaPhlAn2 performed poorly in this scenario, even with the whole set of normalized reads, but it had the best overall running time ranging from 1 min 49 sec with the 1% dataset to 3 min 03 sec for the complete dataset, disregarding database generation (Table 3.2). DUDes had a very similar time for the smaller dataset but an increase in running time for bigger datasets (15 min 38 sec for 50% and 32 min 35 sec for 100%), since it had to deal with larger SAM files (Appendix A - Table 2).

64-organism Archaea and Bacteria synthetic community

In addition to the HMP dataset, we also evaluated another synthetic community. [Shakya et al., 2013] created this set not to simulate any specific environment but to represent phylogenetic and genomic heterogeneity within Bacteria and Archaea usually encountered in communities. It comprises 64 organisms from 62 species. Similarly to the HMP mock set, we evaluated the output and performance of DUDes and three other tools with this dataset. Figure 3.6 shows ROC curves comparing the results of all tools. Except GOTTCHA, all tools had very similar results, identifying all 62 species in the sample (Sensitivity 0.9 and 1, respectively). Kraken achieved the best AUC value, but with the drawback of a long list (962) of false positives (showing only top 150 entries). DUDes performed as well as MetaPhlAn2 in terms of AUC but with only 2 false positives against 14, respectively. MetaPhlAn2 was the fastest tool among all (Table 3.2). GOTTCHA performs poorly in this set, with 15 false positives among its 56 identified organisms.

Shiga-toxigenic *Escherichia coli* (STEC)

In real metagenomics applications, direct performance as with the HMP mock community is typically not possible due to lack of ground truth information. To simulate those scenarios, we analyzed samples where some information about their compositions is already known based on previous conventional microbiology and metagenomics analysis, but the true

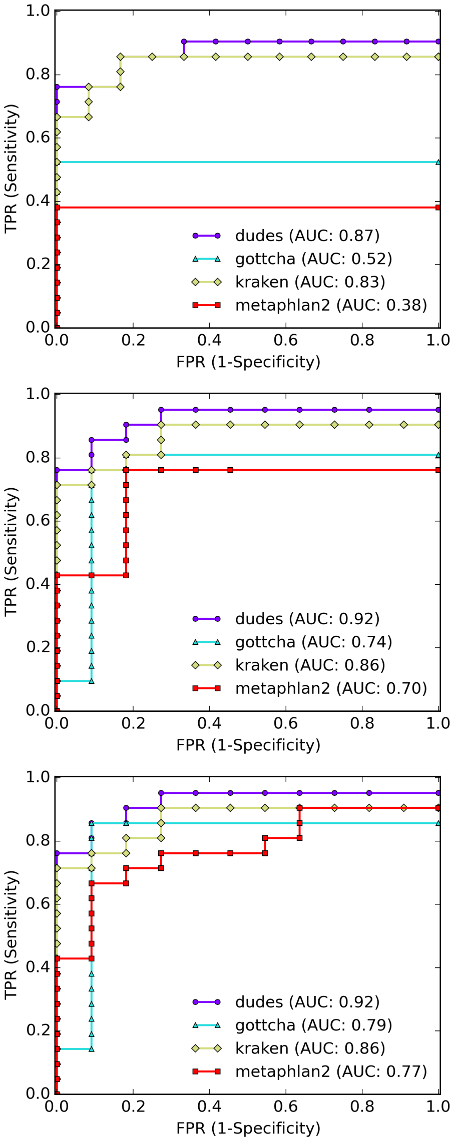


Figure 3.4: ROC curves comparing results based on three sub-sets (1, 50 and 100%) of the normalized HMP Illumina set of reads

composition is not completely known. The data were obtained from stool samples of diarrhea patients during the Shiga-toxigenic *E. coli* (STEC) outbreak in Germany [Loman et al., 2013]. We analyzed four samples, described in the Table 3.3. In this evaluation we try to verify if DUDes is capable of identify previously known pathogens. We opted to frame the discussion here in terms of relative abundance given the lack of complete ground truth, just estimations based on the evaluations provided by [Loman et al., 2013].

We first applied digital normalization to all datasets and mapped the normalized reads against the reference database. We then performed DUDes' top-down analysis, this time allowing strain identifications. In this mode, DUDes will always try to find a set of possible candidate strains in the sample and choose one among them to be the representative

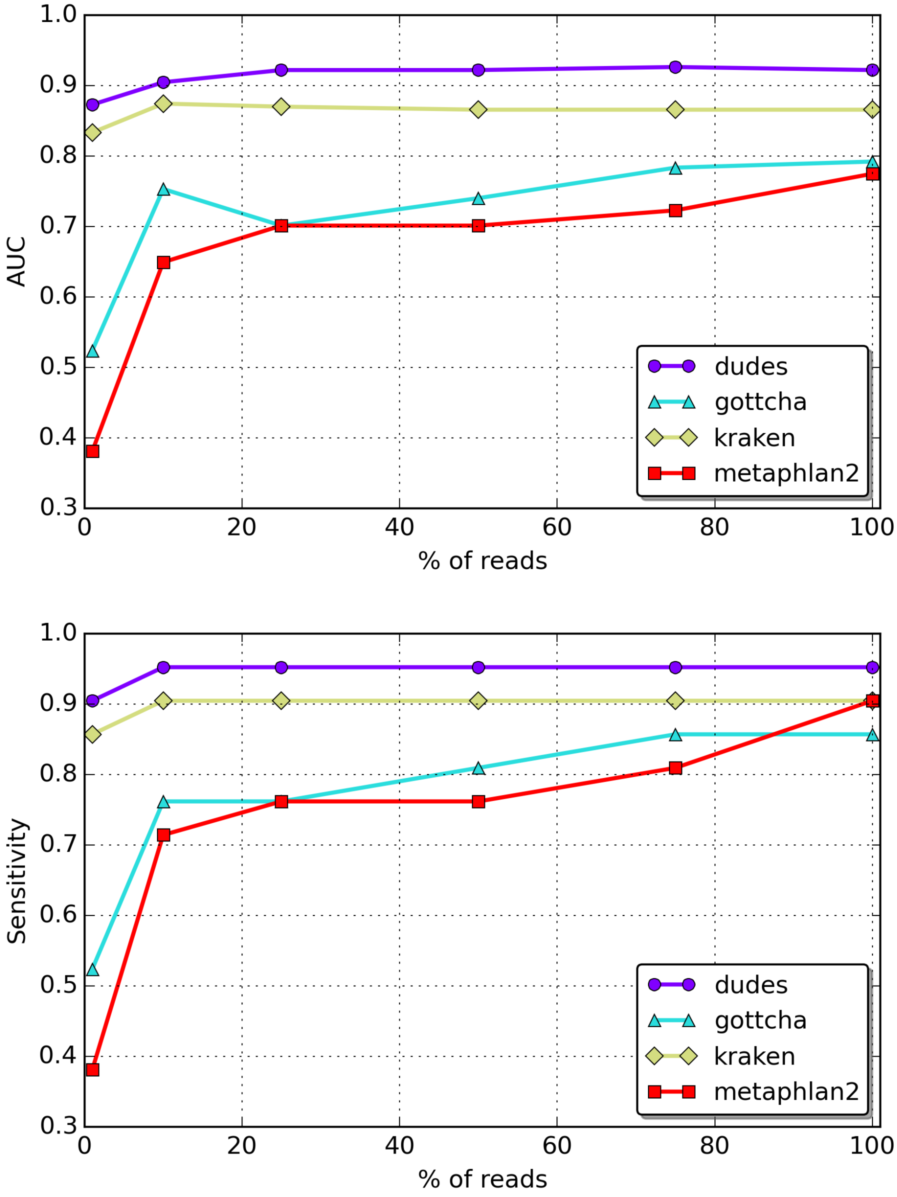


Figure 3.5: AUC and Sensitivity values for the HMP sets. 6 data points were evaluated for each tool (1%, 10%, 25%, 50%, 75% and 100%) representing the percentage of reads in each sub-set

strain on the final output. In the sample 1122, DUDes identified 53 likely strains with relative abundance higher than 0.01%, with *C. difficile* (strain M68) present in a low abundance (approximately 0.35%) (Table 3.3). In sample 1253, DUDes identified 47 strains with relative abundance higher than 0.01%, with *C. difficile* (strain 630) and *Campylobacter concisus* (strain 13826) among them , with abundances of approximately 0.15% and 0.32%, respectively. For samples 2535 and 2638 DUDes identified STEC (*E. coli* O104:H4

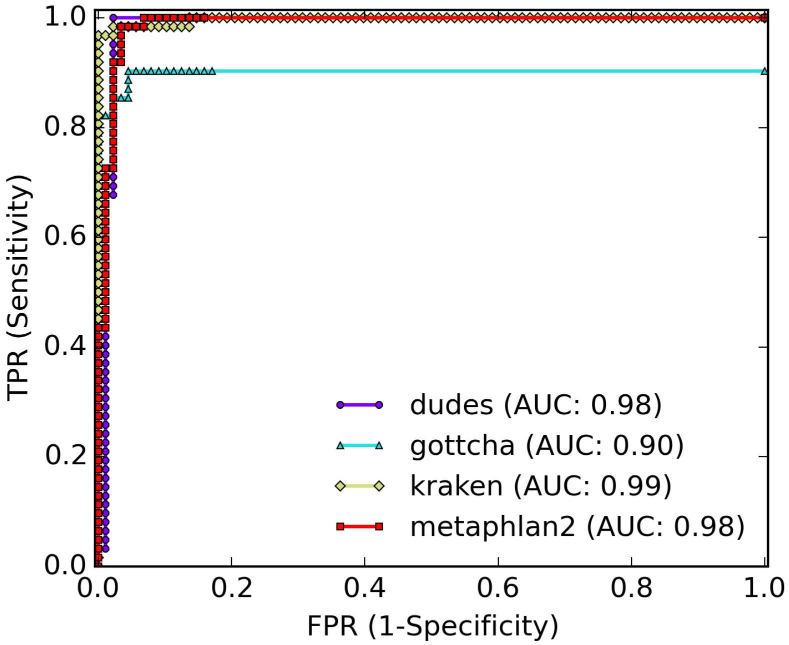


Figure 3.6: ROC curves comparing results of the 64-organism Archaea and Bacteria synthetic community

Sub-set	DUDes	GOTTCHA	kraken + kraken_report	MetaPhlAn2
HMP 1%	01:29	02:30	06:42 + 00:26	01:49
HMP 10%	04:13	03:13	09:08 + 00:28	02:00
HMP 25%	08:06	04:28	09:53 + 00:29	02:04
HMP 50%	15:38	06:24	10:50 + 00:35	02:16
HMP 75%	23:42	08:46	11:56 + 00:39	02:47
HMP 100%	32:35	15:37	27:13 + 00:47	03:03
64-organism	02:46:14	05:09:50	01:05:33 + 05:05	39:52

Table 3.2: Running time of the evaluated tools ([hh:]mm:ss). Computer specifications: AMD Opteron(tm) Processor 6174 2.2-3.2 Ghz, 48 cores, 256GB Ram.

str. 2011C-3493) as candidate strain with a medium and high relative abundance values (approximately 10% and 26%, respectively), matching the previous analysis results and selecting pathogenic strains as candidates.

DUDes can also provide a guided identification, starting the analysis from a defined node of the taxonomic tree. It can be applied when a certain taxonomic group is known to be present in the sample from previous analyses or other approaches (e.g. LCA-based tools) but a specific identification in deeper levels is necessary. For example, conventional tests can identify that a certain family is highly abundant on the sample. Starting the analysis from a specific family node, DUDes will provide a more accurate profile for the following taxonomic levels, giving abundance values relative to the specified starting node.

Sample	Pathogens	Expected abundance	DUDes estimated abundance
1122	*C. difficile*	low	0.35%
1253	*C. difficile*	low	0.15%
	C. concisus	low	0.32%
2535	STEC	medium	10%
2638	STEC	high	26%

Table 3.3: STEC Samples evaluated. Known pathogens were discovered from conventional microbiology and computational metagenomics analysis [Loman et al., 2013]. All samples were generated with a single Illumina MiSeq run (2x151 paired-end sequencing)

We applied the guided identification to samples 2535 and 2638, starting the analysis from previous MetaPhlAn2 results, aiming at strain identification. MetaPhlAn2 identified the *E. coli* species as the most abundant organism in both samples but it did not identify any specific strain. Starting the analysis from the *E. coli* species level (taxid:562) DUDes precisely identified the STEC (*E. coli* O104:H4 strain 2011C-3493) as the first candidate and most likely strain as well as other probable strains in smaller abundance. Further, we performed the same guided identification at each taxonomic rank above STEC (Appendix A - Table 3). The higher the starting taxonomic level, the more precise are the identifications with less strain candidates. DUDes' specific identification method can be applied as a complement for other methods of identification, improving their results and providing a deeper classification.

Tara oceans

We choose one marine sample from the Tara Oceans Project [Sunagawa et al., 2015] to evaluate DUDes' performance on a mixed and complex environment. Differently from human host-associated data, marine environments communities are very challenging to be analyzed, mainly for the lack of reference sequences available. We built a custom database, focused on organisms commonly found in marine environments, composed of Bacterial, Archaeal, Viral and Eukaryotic organisms.

From approximately 289 million reads on our selected sample (Appendix A - Table 1), only 2.2% (less than 5 million reads) could be mapped against the reference database. That shows how difficult it is to profile metagenomic samples in a whole genome fashion lacking known references. Still, we evaluated DUDes output from this set and compared against the 16S rRNA evaluation provides by the Tara Oceans Project. At the Kingdom level, DUDes estimated Bacteria as the most common organism (>90%) with the Proteobacteria phylum being the most representative (55%), as indicated in the 16S-based study. DUDes profiled Archaeal organisms representing approximately 2% of the set. Viruses and Eukaryotes sum up to more than 4%, low as expected from such environment.

Since only 2% of the reads could be analyzed, this analysis is still biased towards the references. But it is shown here that DUDes can cope with mixed non-human host-associated environments and that it could profile such samples with a wider range of reference sequences.

3.4 Discussion

We described here a new method for profiling metagenomics samples with a completely new approach to explore the taxonomic tree structure. In our experiments with mock communities, DUDes achieved high accuracy even in lower taxonomic levels. In an extreme scenario, using only 1% of the data with very few reads for some organisms, DUDes was the best tool overcoming GOTTCHA and MetaPhlAn2. Surprisingly, kraken, a read binning tool, had an excellent performance in this set, but with the disadvantage of a high number of false positives, as expected in this sort of application. DUDes achieved two times fewer false positives than kraken with the best AUC and sensitivity value among all tested tools in the HMP experiment. With 10% of the dataset, DUDes performed as well as with the full set of reads, showing that it can be very precise and stable with small sample sizes. In a broader synthetic community set with more than 100 million reads, DUDes performed equally well even with high organism diversity. From those results we see two main advantages of our tool: first, the method can be applied to profile datasets with low abundant organisms, identifying taxonomic groups with very few matches. Secondly, selecting only part of the dataset poses as a good approach to reduce sample size, allowing faster analysis. Smaller samples and the application of digital normalization generated good results in the HMP datasets without information loss (Appendix A - Figure 4). This strategy can be useful with the increasing amount of data generated by NGS technologies, decreasing mapping and execution time as well as memory usage. In the meantime, a thorough analysis is necessary to better estimate the effects of such techniques. Digital normalization can skew abundances downwards and should be carefully used when there is no ground truth available.

DUDes also performed well when applied to outbreak samples for pathogen detection. Our method's results corroborated previous findings and could be used as a fast alternative or confirmation tool to the pathogen detection problem. DUDes can also be a fine-tuning tool for posterior analysis of LCA-based methods identifications when they cannot achieve high taxonomic levels.

The candidate selection approach used in DUDes is not unique to DUD method and it could also be applied in LCA tools. However DUDes provides this functionality out-of-the-box. In addition the choice is not based only in a presence of a certain taxon (given by counting read matches) but also based on a comparison against the other taxons of the same taxonomic level, giving more significance to the candidate selection.

By transforming the information from read mapping to bins of the same size, and subsequently selecting the same number of bins for comparisons we could ameliorate the fact that taxonomic groups are not evenly represented in the database. This technique provides a fair comparison among taxonomic groups regardless of their number of reference sequences.

The deepest uncommon descendent method implemented in DUDes provides a new way to analyze the taxonomic tree structure. We see several advantages in this method: first it provides a reliable way to identify the presence of taxonomic groups, making a comparison on each taxonomic level. It also can solve ambiguities in a less conservative manner, first by allowing concurrent identifications and second by selecting a set of candidates when a specific identification is not possible. In comparison with LCA-based tools we can point out some methodological differences: kraken, like many LCA-based methods, applies LCA on a sequence level and set the presence of taxonomic groups after all sequences were classified in the LCA system. DUDes uses the taxonomic information to guide the analysis and it does not use the DUD algorithm directly on a sequence level but on a taxonomic group level. It is important to notice that those two methods have opposite approaches (Figure 3.2) and at the same time have been applied in a different way. DUDes' implementation of the deepest uncommon descendent method relies on permutations tests among bin scores of nodes of the tree with correction for multiple testing. That introduces statistical significance to our comparisons and decisions with a rigid control of type I errors values to avoid false identifications.

DUDes is a flexible tool that does not rely on a specific or custom-built databases or read mappers and it can run using any set of reference sequences (e.g. draft genomes, marker genes, proteins) not only whole genomes sequences as here presented. The creation of any other custom database is straightforward with DUDesDB. Our tool also provides a possibility to analyze sub-trees from the taxonomic tree structure by setting a start node and a final taxonomic level desired, giving a guided and fast identification.

In conclusion, DUDes propose a novel approach to the taxonomic profiling problem, with a top-down technique to analyze taxonomic tree structures. The deepest uncommon descendent can be less conservative than current methods for solving ambiguities in the identifications, and showed superior performance in our experiments compared to recent tools, being very precise at low coverages. Additionally the tool provides a strain identification method that can propose one or more strains presents in a sample. We believe that DUDes can be useful for several applications, from complete metagenomics profiling to pathogen detection studies.

Chapter 4

Integrating metagenome analysis tools with MetaMeta

This chapter is based on a published article:
MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. Piro, V. C., Matschkowski, M., & Renard, B. Y. (2017). Microbiome, 5(1), 101 https://doi.org/10.1186/s40168-017-0318-y

4.1 Background

A large and increasing number of metagenome analysis tools are presently available aiming to characterize environmental samples [Bazinet and Cummings, 2012, Oulas et al., 2015, Peabody et al., 2015, Lindgreen et al., 2016]. Motivated by the large amounts of data produced from whole metagenome shotgun (WMS) sequencing technologies, profiling of metagenomes has become more accessible, faster and applicable in real scenarios and tends to become the standard method for metagenomics analysis [Köser et al., 2012, Pallen, 2014, Fricke and Rasko, 2014]. Tools which perform sequence classification based on WMS sequencing data come in different flavors. One basic approach is the *de novo* sequence assembly [Namiki et al., 2012, Peng et al., 2012, Li et al., 2015], which aims to reconstruct complete or near complete genomes from fragmented short sequences without any reference or prior knowledge. It is the method which provides the best resolution to assess the community composition. However it is very difficult to produce meaningful assemblies from metagenomics data due to short read length, insufficient coverage, similar DNA sequences, and low abundant strains [Howe and Chain, 2015].

More commonly, methods use the WMS reads directly without assembly and are in general reference-based, meaning that they rely on previously obtained genome sequences to perform their analysis. In this category of applications, two standard definitions are employed: taxonomic profiling and binning tools. Profilers aim to analyze WMS sequences as a whole, predicting organisms and their relative abundances based on a given set of reference sequences. Binning tools aim to classify each sequence in a given sample individually, linking each one of them to the most probable organism of the reference set. Regardless of their conceptual differences, both groups of tools could be used to characterize microbial communities. Yet binning tools produce individual classification for each sequence and should be converted and normalized to be used as a taxonomic profiler.

Methods available among these two categories make use of several techniques, e.g. read mapping, k-mer alignment, and composition analysis. Variations on the construction of the reference databases, e.g. complete genome sequences, marker genes, protein sequences, are also common. Many of those techniques were developed to overcome the computational cost of dealing with the high throughput of modern sequencing technologies as well as the large number of reference genome sequences available.

The availability of several options for tools, parameters, databases and techniques create a complicated scenario to researchers to decide which methods to use. Different tools provide good results in different scenarios, being more or less precise or sensitive in multiple configurations. It is hard to rely on their output for every study or sample variation. In addition when more than one method is used, inconsistent results between tools using different reference sets are difficult to be integrated. Furthermore, installation, parameterization, database creation as well as the lack of standard outputs are challenges not easily overcome.

We propose MetaMeta, a new pipeline for the joint execution and integration of metagenomic sequence classification tools. MetaMeta has several strengths: easy installation and set-up, support for multiple tools, samples and databases, improved final profile combining multiple results, out-of-the-box parallelization and high performance computing (HPC) integration, automated database download and set-up, custom database creation, integrated pre-processing step (read trimming, error correction, and sub-sampling) as well as standardized rules for integration of new tools. MetaMeta achieves more sensitive profiling results than single tools alone by merging their correct identifications and properly filtering out false identifications. MetaMeta was built with SnakeMake [Koster and Rahmann, 2012] and is open-source. The pipeline has six pre-configured tools that are automatically installed using Conda through the BioConda channel (https://bioconda.github.io). We encourage the integration of new tools, making it available to the community through a participative Git repository (via pull request). MetaMeta source-code is available at: https://github.com/pirovc/metameta

4.2 Methods

MetaMeta executes and integrates metagenomic sequence classification tools. The integration is based on several tools' output profiles and aims to improve organism identification and quantification. An optional pre-processing and sub-sampling step is included. The pipeline is generalized for binning and profiling tools, categories that were previously described in the CAMI (Critical Assessment of Metagenome Interpretation) challenge (http:// www.cami-challenge.org). MetaMeta provides a pre-defined set of standardized rules to facilitate the integration of tools, easy parallelization and execution in high performance computing infrastructure. The pre-configured tools are available at the BioConda channel to facilitate download and installation, avoiding set-up problems and broken dependencies.

The pipeline accepts one or multiple WMS samples as well as one or more databases and the output is an integrated taxonomic profile for each sample per database (as well as a separated output from each executed tool). The MetaMeta pipeline can be described in 4 modules: database generation, pre-processing, tool execution, and integration (Figure 4.1).

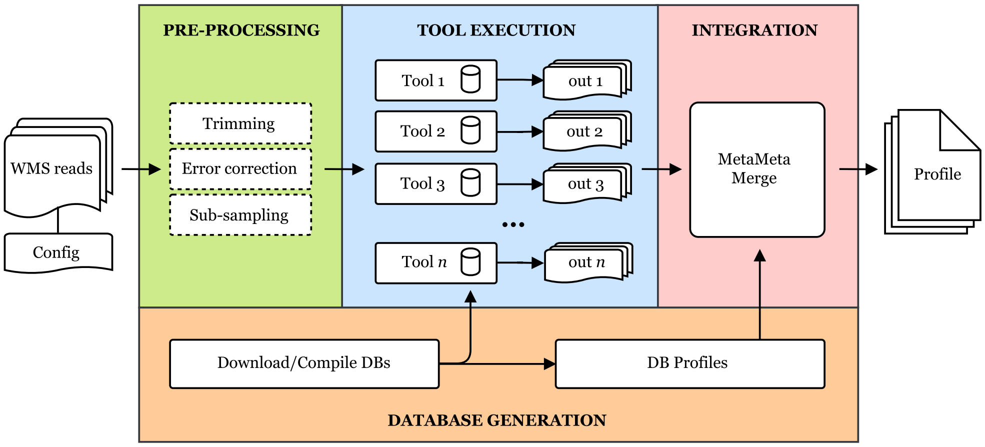


Figure 4.1: MetaMeta Pipeline: The MetaMeta Pipeline. One or more WMS read samples and a configuration file are the input. The pipeline consists of 4 main modules: Database Generation (only on the first run), Pre-processing (optional), Tool Execution and Integration. The output is a unified taxonomic profile integrating the results from all configured tools for each sample, generated by the MetaMetaMerge module.

4.2.1 Database generation

On the first run, the pipeline downloads and builds the databases for each of the configured tools. Pre-configured databases (Appendix B - Table 1) are provided as well as a custom database creation option based on reference sequences. Since each tool has its own database with a specific version of reference sequences, database profiles are generated, collecting which taxonomic groups each tool can identify. Given a list of accession version identifiers for each sequence on the reference set, MetaMeta automatically generates a taxonomic profile for each tool's database.

4.2.2 Pre-processing

An optional pre-processing step is provided to remove errors and improve sequence classification: Trimommatic [Bolger et al., 2014] for read trimming and BayesHammer [Nikolenko et al., 2013] for error correction. A sub-sampling step is also included, allowing the sub division of large read sets among several tools by equally dividing them or by taking smaller random samples with or without replacement, to reduce overall run-time.

4.2.3 Tool execution

In this step, the pre-processed reads are analyzed by the configured tools. Tools can be added to the pipeline if they follow a minimum set of requirements. They should output their results based on the NCBI Taxonomy database [Federhen, 2012] (by name or taxonomic id). Profiling tools should output a rank separated taxonomic profile with relative abundances while binning tools should provide an output with sequence id, length used in the assignment and taxon. The BioBoxes [Belmann et al., 2015] data format for binning and profiling (https://github.com/bioboxes/rfc/tree/master/data-format) is directly accepted. Tools which provide non-standard output should be configured with an additional step converting their output to be correctly integrated into the pipeline (More details are given in the Appendix B - File Formats).

4.2.4 Integration

The integration step will merge identified taxonomic groups and abundances and provide a unified profile for each sample. MetaMeta aims to improve the final results based on the assumption that the more identifications of the same taxon by different tools are reported, the higher its chance to be correct. This task is performed by the MetaMetaMerge module. This module accepts binning and profiling results and relies on previously generated database profiles. Taxonomic classification can change over time and each tool can use a different version/definition of it. For that reason a recent taxonomy database version is used to solve name and rank conflicts (e.g. changing name specification, species turning into sub-species, etc.).

Abundance estimation - binning tools

Binning tools provide a single classification for each sequence in the dataset instead of relative abundances for taxons. An abundance estimation step is necessary for a correct interpretation of such data and posterior integration. The lengths of the binned sequences are summed up for each identified taxonomic group and normalized by the length of their respective reference sequences, estimating the *abundance* for each identified taxon n as:

$$abundance_n = \sum_{i=1}^{r} \frac{\sum_{j=1}^{t_i} b_j}{l_i} \tag{4.1}$$

where r is the number of reference sequences belonging to the taxonomic group n, t_i is the total of reads classified to the reference i, b_j is the number of aligned bases of a read j and l_i is the length of the reference i. The abundance of the parent nodes is based on the cumulative sum of their children nodes' abundance.

Merging approach

The first step on the merging approach is to normalize estimated abundances to 100% for each taxonomic level. That is necessary because some tools do account for the unclassified reads and others do not. MetaMetaMerge only considers classified reads. Once normalized, all profiles are then integrated to a single profile. In this step, MetaMetaMerge saves the number of occurrences of each taxon among all profiles. This occurrence count is used to better select taxons that are more often identified, assuming that they have higher chances of being a correct identification. MetaMetaMerge also calculates an integrated value for the relative abundance estimation, defined as the harmonic mean of all normalized abundances for each taxon, avoiding outliers and obtaining a general trend among the estimated abundances. All steps taken in the merging process are performed for each taxonomic level independently, from super kingdom to species by default.

Since tools use different databases of reference sequences it is necessary to account for this bias. Previously generated database profiles provide which taxons are available for each tool. By merging all database profiles, it is possible to anticipate how many times each taxon could be identified among all tools used. The number of occurrences of each taxon from the tools' output and the database presence number are integrated to generate a score S for each taxon, defined as:

$$S_{ij} = \frac{(i+1)^2}{j+1} \tag{4.2}$$

where i is the number of times the taxon was identified and j the number of times it is contained in the databases. This score calculation accounts for the presence/absence of taxonomic groups on different databases. It gives higher scores to the most identified taxons present in more databases. At the same time, lower scores are assigned to taxons present in many databases but not identified too many times. The score calculation is purposely biased for higher scores when $i = j$ (Appendix B - Figure 1), given the benefit of the doubt for taxons with low identification that are available only in few databases.

Commonly, metagenome analysis methods have to deal with a moderate to high number of false positive identifications at lower taxonomic levels. That occurs mainly because metagenomes can contain very low abundant organisms with similar genome sequences. This problem is even extended in our merged profile by collecting all false positives from different methods, generating a long tail of false positives with lower scores mixed together with true identifications. A filtering step is therefore necessary to avoid wrong assignments. This step is usually performed by an abundance cutoff value. Setting up this value is subject to uncertainty since the real abundances are usually not known and the separation between low abundant organisms and false identifications is not clear [Zepeda Mendoza et al., 2015]. A simple cutoff would not provide a good separation between true and false results in this scenario.

To overcome this problem, MetaMetaMerge classifies each taxon in a set of bins (four by default) based on the calculated score (Equation 4.2). Bins are defined by equally dividing the range of scores in the numbers of bins selected. Now each taxon has a score and a bin assigned to it. Taxons with higher scores are more likely to be true identifications and are going to be grouped together in the same bin. With this strategy it is possible to obtain a general separation among taxons which are prone to be true or false identifications.

Within each taxon grouped in a bin (sorted by relative abundance) a cutoff is applied to remove possible false identifications with low abundance. Here, the cutoff value is a percentile relative to the number of taxons on each bin and it is selected based on predefined functions, which can achieve more sensitive or precise results (Appendix B - Mode functions). Each bin will have a different cutoff value depending on the chosen function.

If precision is chosen, a gradually more stringent cutoff will be used, selecting only the most abundant taxa for each bin. If sensitivity is selected, cutoffs will be set higher, allowing more identifications to be kept. Sensitive results have an increased chance of containing more true positives but at the same time they will likely have more false identifications due to less strict cutoffs.

Based on this percentile cutoff, MetaMetaMerge keeps only the top abundant taxa on each bin and removes taxons below it. After this step, the remaining taxons on each bin are re-grouped and sorted by relative abundance to generate the final profile.

At the end, MetaMeta will provide a final taxonomic profile, integrating all tools results, a detailed profile with co-occurrence and individual abundances, an interactive Krona pie chart [Ondov et al., 2011] to easily compare taxonomic abundances among the tools as well as single profiles for each executed tool.

4.3 Results

4.3.1 Tool selection

MetaMeta was evaluated with a set of six tools: CLARK [Ounit et al., 2015], DUDes [Piro et al., 2016], GOTTCHA [Freitas et al., 2015], Kraken [Wood and Salzberg, 2014], Kaiju [Menzel et al., 2016], and mOTUs [Sunagawa et al., 2013]. The choice was partially motivated by recent publications comparing the performance of such tools [Peabody et al., 2015, Lindgreen et al., 2016, Sczyrba et al., 2017]. CLARK, GOTTCHA, Kraken and mOTUs achieved very low false positive numbers according to [Lindgreen et al., 2016]. DUDes was an in-house developed tool which achieves good trade-off between precision and sensitivity according to [Sczyrba et al., 2017]. Kaiju uses a translated database, bringing diversity to the current whole genome based methods. We also considered the amount of data/run time performance for each tool, selecting only the ones that can handle large amounts of data as commonly used today in metagenomics analysis in an acceptable time (less than 1 day for our largest CAMI dataset - 7.4 Gbp). MetaPhlAn [Truong et al., 2015] a widely used metagenomics tool could not be included due to taxonomic incompatibility. Any other sequence classification tool could be configured and used in MetaMeta, as long as it fits with our pipeline requirements described in the Methods - Tool execution section. We selected an equal number of tools for each category: DUDes, GOTTCHA and mOTUs are taxonomic profiling tools, while CLARK, Kraken and Kaiju are binning tools. Databases were created following the default guidelines for each tool, considering only bacteria and archaea as targets (Appendix B - Table 1).

4.3.2 Datasets and evaluation

The pipeline was evaluated with a set of simulated and real samples (Table 4.1). The simulated data were provided as part of the CAMI Challenge (toy samples) and the real samples were obtained from the Human Microbiome Project (HMP) [Methé et al., 2012, Huttenhower et al., 2012]. MetaMeta was compared to each single result from each tool configured in the pipeline. Although the pipeline can work on the strain level, we evaluate the results until species levels since most of the tools still do not provide strain level identifications. We compare the results to the ground truth in a binary (true and false positives, sensitivity and precision) and quantitative way with the L_1 norm, which is the sum of absolute differences between predicted and real abundances, when abundance profiles are available. Computer specifications and parameters can be found on the Appendix B.

Sets	# Samples	Total bases	# species	cpu time/sample	estimated wall time/sample
CAMI Toy Low	1	14.8 Gbp	30	31:04:52	02:35:24
CAMI Toy Medium	4	31.3 Gbp	199	15:18:16	01:16:31
CAMI Toy High	5	74.5 Gbp	375	33:20:30	02:46:42
HMP Stool	147	1.44 Tbp	299*	19:39:39	01:38:18

Table 4.1: Samples used in this study and run-time (based on the Computer specifications on Appendix B). cpu time/sample stands for the mean cpu time for each sample without paralellization. estimated wall time/sample considers a double speed-up by using 12 threads and concurrently running all 6 tools (when computational resources are available the pipeline can run all tools/samples/databases at the same time). * expected number of species from isolated genomes from the gastrointestinal tract

CAMI data

The CAMI challenge provided three toy datasets of different complexity (Table 4.1) with known composition and abundances. From low to high complexity, they provide an increasing number of organisms and samples. The samples within a complexity group contain the same organisms with variable abundances among samples. The sets contain real and simulated strains from complete and draft bacterial and archaeal genome sequences. The simulated CAMI datasets, especially those of medium and high complexity, provide a very challenging and realistic data in terms of complexity and size.

In Figure 4.2 it is possible to observe the tools performance in terms of true and false positives for the CAMI high complexity set. All configured tools perform similarly in the true positive identifications but vary among the false positives. Binning tools have a higher number of false positive identifications due to the fact that even single classified reads are considered. The MetaMetaMerge profile surpassed all other methods in true positive identifications while keeping the false positive number low. The same trend occurs in the other complexity sets (Appendix B - Figures 3-8). Figure 4.3 shows the trade-off between precision and sensitivity for all high complexity samples. MetaMetaMerge achieved the best sensitivity while GOTTCHA the best precision among the compared tools with default parameters. Those results show how the merging module of the MetaMeta pipeline is capable of better selecting and identifying true positives based on the co-occurrence information. MetaMetaMerge also has the flexibility to provide more precise or sensitive results (Figure 4.3) just by changing the *mode* parameter (details are given in the Appendix B - Mode functions). In the very precise mode, the merged profile outperformed all tools in terms of precision, but with the cost of losing sensitivity. In the very sensitive mode, the merged profile could improve the sensitivity compared to the run with default parameters, with some loss of precision. It is important to notice that the trade-off between precision and sensitivity could also be explored by the *cutoff* parameter (default 0.0001), depending on what is expected to be the lowest abundant organism in the sample. The MetaMetaMerge *mode* parameter will give more precise or sensitive results based on this cutoff value.

In terms of relative abundance, MetaMetaMerge provides the most reliable predictions with smaller difference from the real abundances, as shown in Figure 4.4 with regard to the L_1 norm measure. By taking the harmonic mean, we succeed in reducing the effect of outliers that occur among the tools and capture the trend of the estimated relative

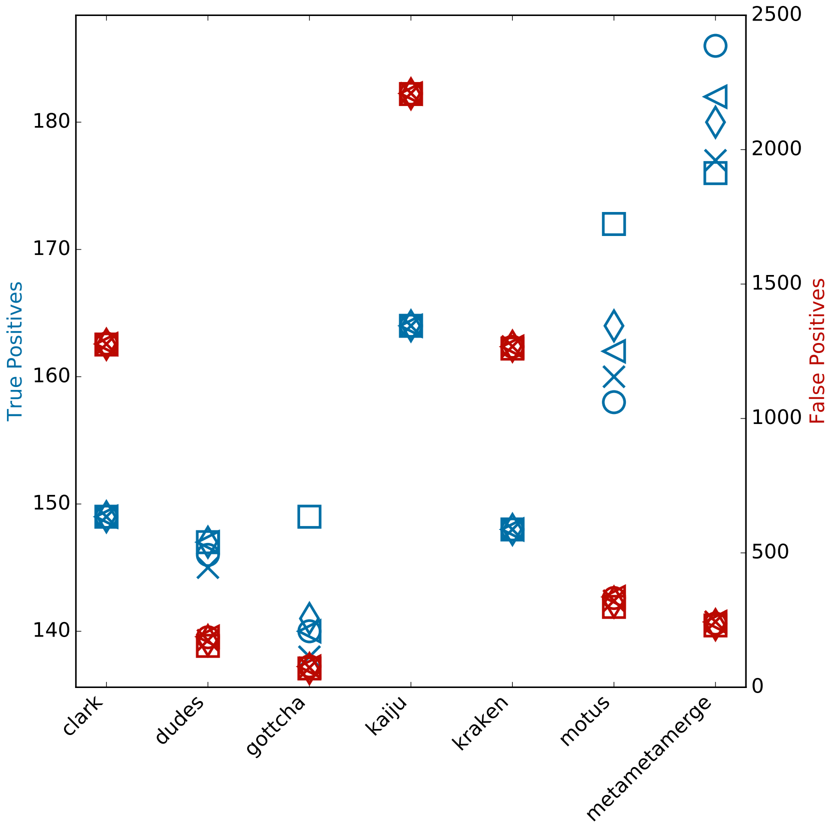


Figure 4.2: True and False Positives - CAMI high complexity set. In blue (left y axis): True Positives. In red (right y axis): False Positives. Results at species level. Each marker represents one out of five samples from the CAMI high complexity set.

abundances, providing a new, more robust estimate.

Pre-processing and sub-sampling effects

We explore here the effects of pre-processing and sub-sampling on the CAMI toy sets. Results shown in this section were trimmed and sub-sampled in several sizes, with and without replacement and executed five times for each sub-sample. Trimming effects were small on this set, slightly increasing precision (data not shown). Figure 4.5 shows the effects of sub-sampling in terms of sensitivity and run-time (wall time for the full pipeline) for one of the high complexity CAMI sets. Sub-sampling provides a high decrease on run-time for every tool and consequently for the whole pipeline. However, only below 5% it is possible to see a significant but still small decrease on sensitivity. All tools behave similarly on the sub-sampled sets, with GOTTCHA and mOTUs having a high decrease of sensitivity when using only 1% of the data. With the same sub-sample configuration

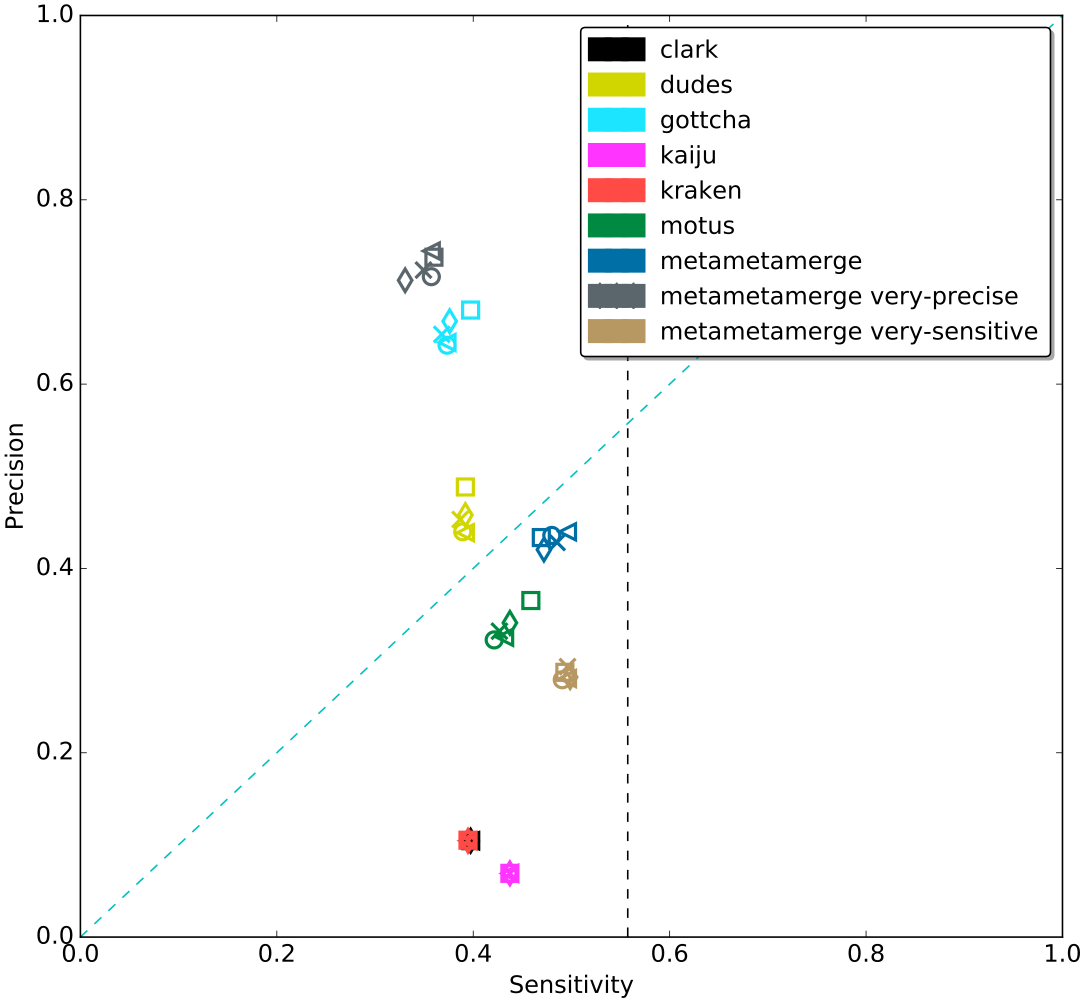


Figure 4.3: Precision and Sensitivity - CAMI high complexity set. Dotted black line marks the maximum possible sensitivity value (0.57) that could be achieved with the given tools and databases. Results at species level. Each marker represents one out of five samples from the CAMI high complexity set.

(1%), MetaMetaMerge achieved a sensitivity higher than any other tool alone using 100% of the set. It also runs the whole pipeline approximately 17 times faster than with the full set (from 05h41m36s to 20m19s on average), being faster than the fastest tool with 100% of the data (kraken 29m26s on average) and the second best sensitive tool (kaiju 1h47m44 on average). As expected, precision is slightly increased in small sub-samples due to less data (Appendix B - Figure 9).

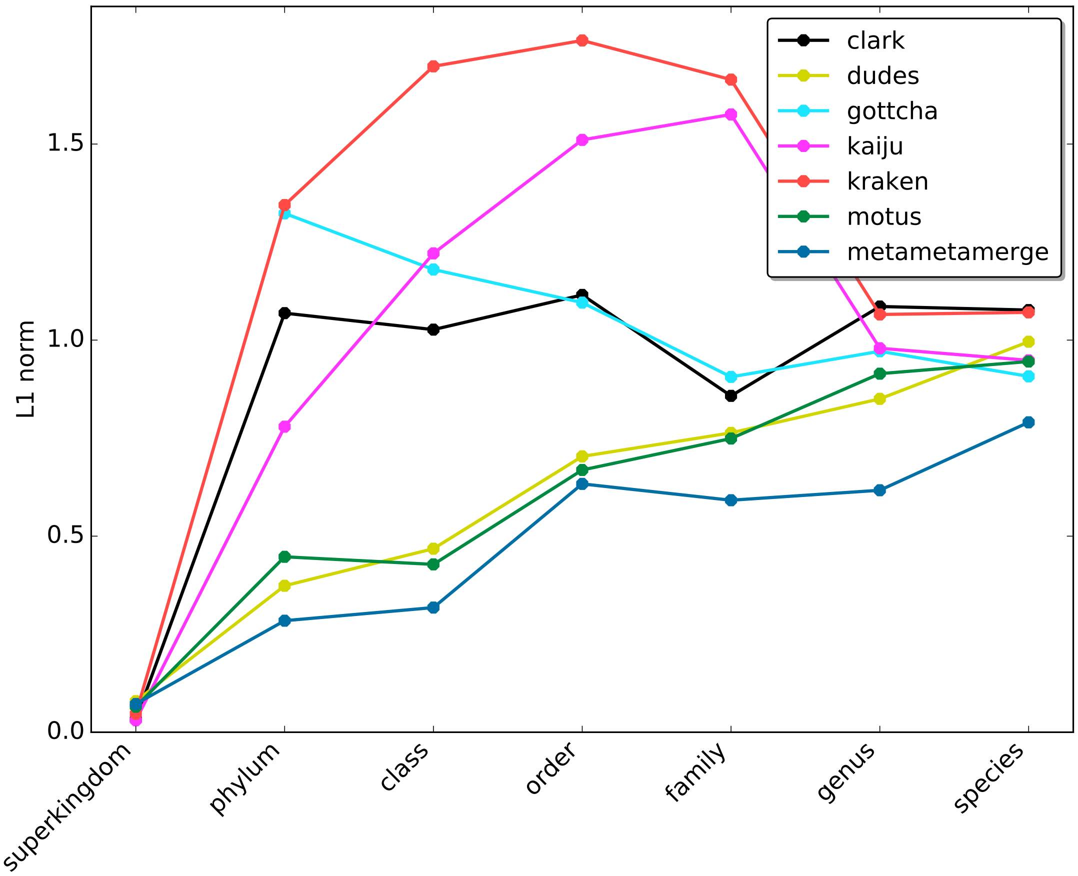


Figure 4.4: L_1 norm error. Mean of the L_1 norm measure at each taxonomic level for five samples from the high complexity CAMI set.

Human Microbiome Project data

The HMP provided several resources to characterize the microbial communities at different sites of the human body. MetaMeta was tested on stool samples to evaluate the performance of the pipeline on real data. For evaluation we used a list of reference genome sequences that were isolated from specific body sites and sequenced as part of the HMP. They do not represent the complete content of microbial diversity in each community but serve as a guide to check how well the tools are performing. Stool samples were compared against the isolated genomes obtained from the gastrointestinal tract.

Figure 4.6 shows the results for 147 samples. In sensitive mode, MetaMetaMerge achieved the highest number of true positive identifications with a moderate number of false positives, below all binning tools but above all taxonomic profilers. mOTUs produced good results in the selected samples mainly because its database is based on the isolated genomes

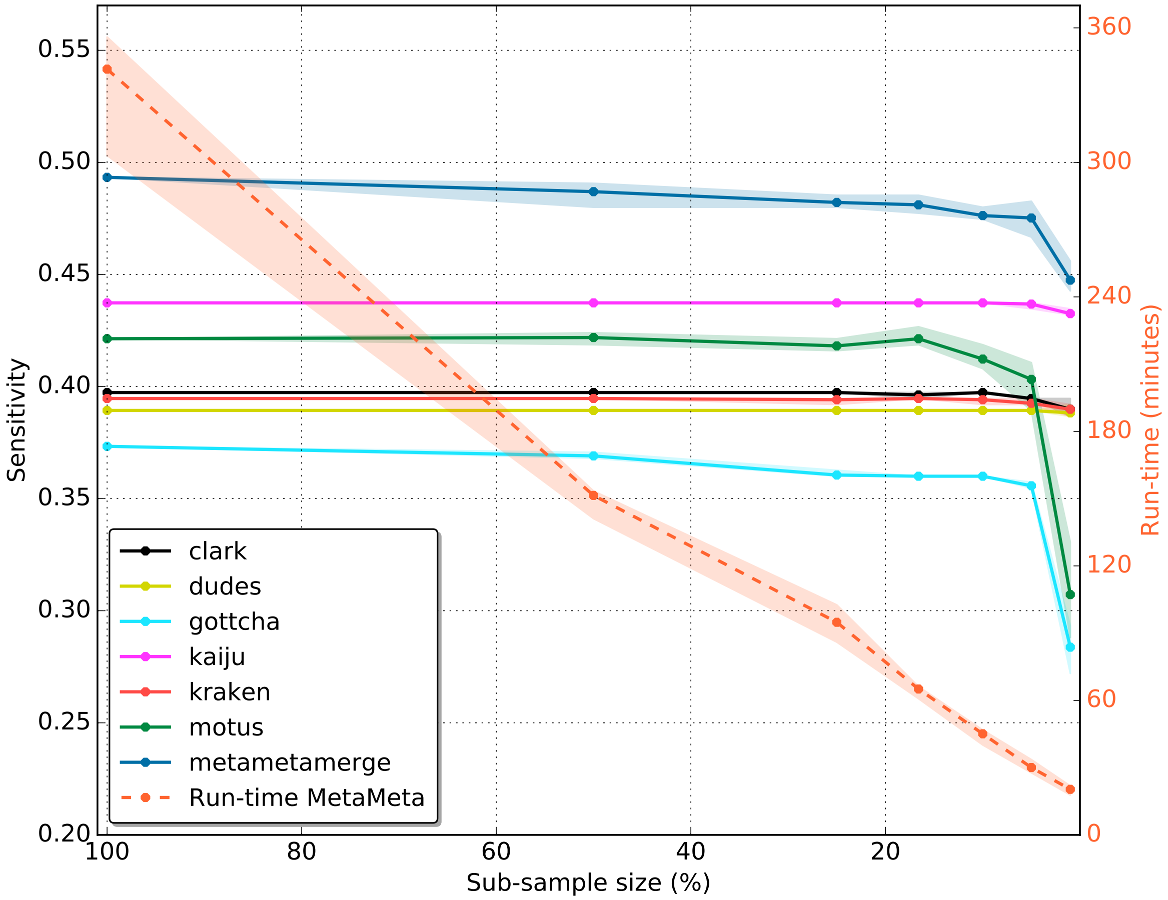


Figure 4.5: Sub-sampling. Sensitivity (left y axis) and run-time (right y axis) at species level for one randomly selected CAMI high complexity sample. Each sub-sample was executed five times. Lines represent the mean and the area around it the maximum and minimum achieved values. Run-time stands for the time to execute the MetaMeta pipeline. The evaluated sample sizes are: 100%, 50%, 25%, 16.6%, 10%, 5%, 1%. 16.6% is the exact division among 6 tools, using the the whole sample. Sub-samples above that value were taken with replacement and below without replacement. The plot is limited to a value of 0.57 (left y axis) that is the maximum possible sensitivity value that could be achieved with the given tools and databases.

from the HMP (the same as the ground truth used here). Since mOTUs is the only tool with a distinct set of reference sequences that could classify this set, the scores (from Equation 4.2) attributed to mOTUs' unique identifications were low. Still, MetaMetaMerge could improve the true identifications keeping a lower rate of false positives by incorporating the true identifications from other methods.

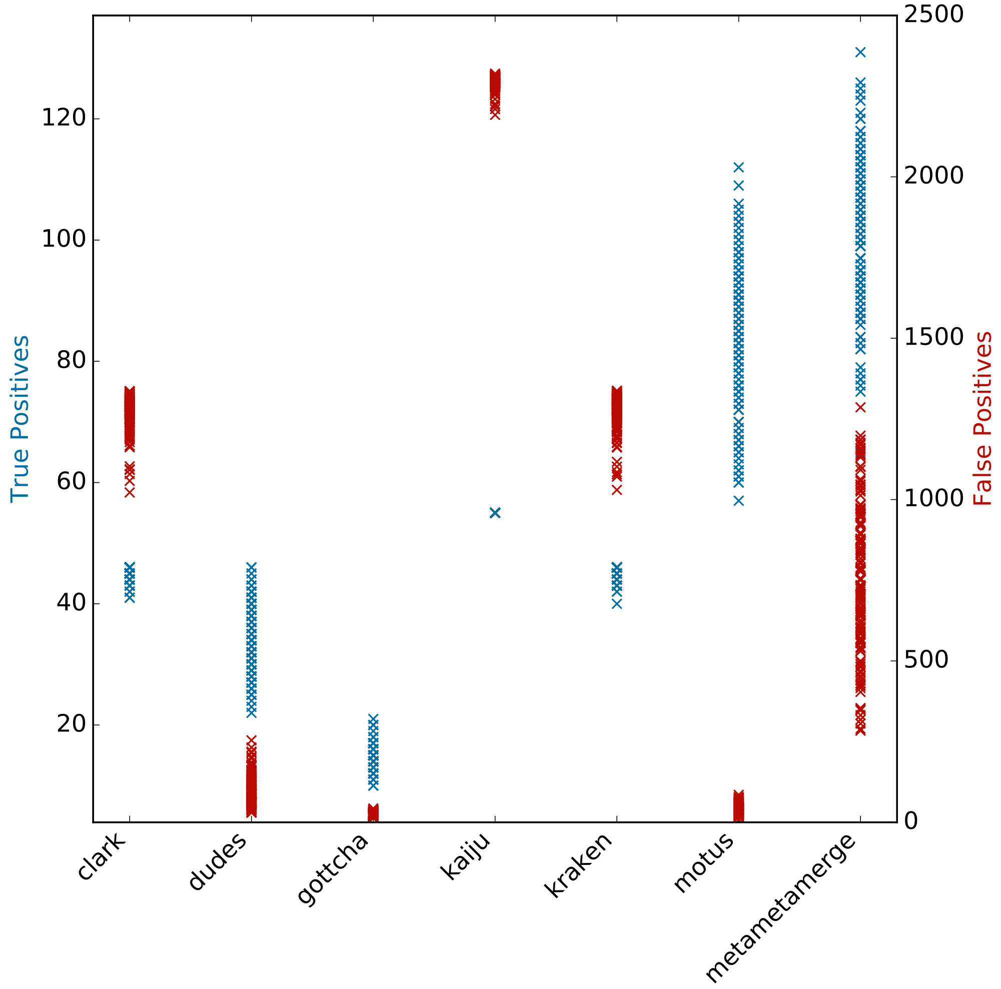


Figure 4.6: True and False Positives - HMP stool samples. In blue (left y axis): True Positives. In red (right y axis): False Positives. Results at species level. Each marker represents one out of 147 stool samples from the HMP.

4.4 Discussion

MetaMeta is a complete pipeline for classification of metagenomic datasets. It provides improved profiles over the tested tools by merging their results. In addition, the pipeline provides easy configuration, execution and parallelization. With simulated and real data, MetaMeta is capable to achieve higher sensitivity. That is possible due to the MetaMetaMerge module, which extracts information of co-occurrence of taxons on databases and profiles, collecting complementary results from different methods. Further, the guided cutoff approach avoids false positives and keeps most of the true identifications, enhancing final sensitivity and exploring the complementarity of currently available methods.

By running several tools, MetaMeta has an apparently prohibitive execution time. In reality, the parallelization provided by Snakemake makes the pipeline run in a reasonable time using most of the computational resources (Table 4.1). That is possible by the way the rules are chained and executed among several cores, lasting not more than the slowest tool plus pre- and post-processing time, which are very small in comparison to the analysis time. In addition, sub-sampling allows the reduction of input data and a high decrease of execution time with small if any impact on the final result. That is viable due to redundant data contained in many metagenomic samples as well as redundant execution by several tools provided in the MetaMeta environment. However sub-sampling should be used with caution, taking in consideration the coverage of low abundant organisms.

All tools presented here are available at the BioConda channel and are automatically installed in the first MetaMeta run, working out-of-the-box for several computer environments and avoiding conflicts and broken dependencies. MetaMeta can also handle multiple large samples and databases at the same time, with options to delete intermediate files and keep only necessary ones, being well suited to large scale projects. It also reduces idle computational time by smartly parallelizing samples among one or more tools (Appendix B - Figures 10-13). The parallelization noticeably decreases the run time when computational power is available and manages to serialize and control the run when access to computational power is limited. Integration into HPC systems is also possible and we provide a pre-configured file for queuing systems (e.g. slurm). As stated by Lee et al. [Lee et al., 2016], solid-state drives accelerate the run time of many bioinformatics tools. Such drives were used in some evaluations shown in this paper and are beneficial for the MetaMeta pipeline.

MetaMeta makes it easier for the user to obtain more precise or sensitive results by providing a single default parameter as well as advanced options for more refined results. This parameter when set towards sensitivity tends to output an extensive list of taxons, being at the same time less stringent with the minimum abundance cutoff. When set towards precision it will apply a more strict abundance cutoff and provide a smaller but more accurate list of predicted taxons. Since all tools were used in default mode, it is possible to obtain problem-centric optimized results only by changing the way MetaMeta works. That facilitates and simplifies the task for researchers that are in search for a specific goal.

MetaMeta supports strain level identification. Nevertheless all evaluations were made at species level due to lack of support to strain identification in some tools. Also the lack of standard was a limiting factor. Taxonomic IDs are no longer assigned to strain levels [Federhen et al., 2014] and tools output them in different ways. With standard output definitions, the use of strain classification on the pipeline is straight forward.

Related in parts, a method called WEVOTE was developed in parallel and recently published [Metwally et al., 2016] where five classification tools were used to generate a merged taxonomic profile. Although the two methods present distinct ways of achieving better taxonomic profiling, they are not built for the same use case. WEVOTE relies on BLAST based tools and thereby is not suited for the large scale WMS applications, since the dataset sizes practically prohibit analyses via BLAST based approaches. Differently, MetaMeta was built accounting for high throughput data. Moreover, we supply an easy way to install tools and MetaMeta provides a complete pipeline which can configure databases and run classification tools with an integration module at the end, where WEVOTE provides only the integration method. As a result a comparison among the pipelines is hard to perform and interpret since they both use a different set of tools and databases.

In conclusion, MetaMeta is an easy way to execute and generate improved taxonomic profiles for WMS samples with multiple tool support. We believe the method can be very useful for researchers that are dealing with multiple metagenomic samples and want to standardize their analysis. The MetaMeta pipeline was built in a way to facilitate the execution in many computational environments using Snakemake and BioConda. That diminishes the burden of installing and configuring multiple tools. The pipeline also gives control over the storage of the results and has an easy set of parameters which makes it possible to obtain more precise or sensitive results. MetaMeta was coded in a standardized manner, allowing easy expansion to more tools, also collectively in the MetaMeta git repository (https://github.com/pirovc/metameta). We believe that the final profile could be even further improved with novel tools configured into the pipeline.

Chapter 5

Summary and Conclusion

Metagenomics is changing the way researchers understand and study life on earth. Before the exploration of microbial communities on a large scale, most of the biological concepts were based on eukaryotic organisms and their mechanical interactions. In an analogy proposed by [Robbins et al., 2016], classical biology and the classical macroscale eukaryotic realm is to classical physics the same way metagenomics and the microscale microbial realm is to quantum physics. Many foundations of the classical fields do not apply to their quantum and microscale equivalent realms. Metagenomics is providing the means to understand this vast and mostly unknown microbial world which comprises at least half of earth's biomass and most of its genetic diversity.

In this thesis I present three contributions with novel approaches in computational metagenomics, reference and taxonomy-based methods to analyze new data based on previously obtained data. Ganon and DUDes are independent tools for binning and profiling metagenomic samples, respectively, and MetaMeta is a pipeline to execute both categories of tools and improve their results. These contributions implement alternative methodologies to solve a primary question in metagenomics: which organisms are in the sample? Given the fast advances in genome sequencing and assembly, answering such a question is not as trivial as comparing the newly sequenced data with the already obtained data. Computer techniques and statistical methods should be applied to perform such tasks in a timely and efficient way. Simple comparisons between new and old data are not enough, given the characteristics of DNA molecules and how we obtain and read them. Reads are relatively short compared to genome sizes and contain errors as well as assembled reference sequences which are often incomplete. Bacteria of the same species have many similarities in their genomic content, and differences between them can be of just a few base pairs. Horizontal gene transfer is a phenomenon that changes how microorganisms interact and evolve, mixing their genomic content and playing an important role in the composition of high complexity environments. These and many others aspects make necessary computational methods to improve our ability to analyze this immense amount of data and provide precise and insightful results efficiently.

The astonishing data growth provided by high-throughput sequencing technologies keeps yielding many unforeseen possibilities for reference-based analysis. However, the massive amount of data which is increasing daily is becoming a real issue for current computational resources. Such data volume cannot be easily processed, limiting the exploration of its full potential. That was the main motivation to develop ganon. Using a variation of a well-studied probabilistic data structure, the bloom filter, together with a taxonomic clustering, ganon manages to quickly translate many reference sequences into a small and efficient searchable index. The clustering step brings similar sequences with close taxonomic classification together, reducing the number of unique entries on the filter. Since ganon's aim is to provide taxonomic binning of sequences, the clustering was taxonomic-oriented, but also extendable to assembly level — which is comparable to strain level [Federhen et al., 2014]. Ganon's index structure also allows updates, a functionality not available in most taxonomic binning tools hitherto. Those updates take a fraction of time necessary to build the complete index and are extremely useful to keep up to date with the data repositories. Ganon classifies reads based on the k-mer counting lemma, assuring that only reads with a minimum number of k-mer matches against the reference sequences will be classified, reducing spurious matches. On top of that, we apply a progressive filter which keeps only the top hits, resulting in a reduced number of false positives in the final assignments. Together with multiple filter support in a user defined hierarchy, false positive rate control on building step, optional lowest common ancestor to solve ambiguities, among others, ganon is a complete tool for sequence classification developed for metagenomics. Improvements in building and update performance are crucial for current data volumes and improvements in precision can produce more accurate follow-up analysis.

Resolution in community profiling is crucial to fully characterize high complexity environments, especially for pathogen detection, where precise identification at strain level is fundamental. Such environments may contain hundreds to thousands of species which can exchange genetic material among them. As a result, many strains of the same species with very similar genomes can be present in the community at the same time. Ideally, taxonomic profilers should be able to detect strains in high resolution. In practice, such resolution is not trivial given that short fragments of sequences match equally among similar genome sequences. A common solution for such entanglement is to select the lowest common ancestor of the multiple identified strains. This approach increases the precision on higher taxonomic levels but drastically decreases sensitivity at the desired low levels, which is not ideal. DUDes and the deepest uncommon descendent technique were developed to improve strain identification. By evaluating the presence of each taxonomic group in a top-down manner, from the root node until the leaf nodes of the taxonomic tree, DUDes provides more robust and statistically significant results. In cases where low taxonomic resolution (e.g. species, strain) are not possible due to multiple matches, DUDes evaluates the chances of each strain or taxonomic group to be present and reports a list of probable candidates. By comparing all assignments of every taxonomic groups at the same level, DUDes also accounts for the uneven distribution of sequences on the taxonomic tree. DUDes was one of the participant tools of the first CAMI Challenge [Sczyrba et al., 2017] and ranked among the top three tools when considering the average of precision and recall for taxonomic profiling.

Binning and taxonomic profiling methods both characterize metagenomes in different ways, by sequence and by sample, respectively. Taxonomic profiling can also be seen as an extension of binning methods, since first it is necessary to infer the origin of each read in a sample to later summarize the contents of the community. However, many binning tools do not profile communities and are specialized only on sequence classification. The translation among those two categories of tools can be beneficial but it is not a straightforward task. Genome lengths and multiple matches should be considered when profiling and estimating abundances. Merging results from both types of tools is also not trivial in an automated way due to lack of file standards. Those were some of the motivations to develop MetaMeta, a pipeline to execute both binning and taxonomic profiling tools and to merge their results. The unified result takes into account the abundance estimation and the availability of strains on each tool's database, ranking them and selecting the most probable candidates to provide a final profile. With a scoring technique, MetaMeta keeps most of the true positive identifications and discards false positives, achieving higher sensitivity with a small decrease in precision. Abundance estimation is also improved by balancing the output provided from each tool. In addition, MetaMeta is a complete pipeline to build indices and to parallelize the execution of several samples in large computational environments. MetaMeta not only facilitates the use of tools, but also standardizes their outputs and merges them into one final and improved profile, further generating plots for better visualization and comparisons.

Reference sequences and taxonomic classification were used throughout this work. This background data have a very important role in how metagenome analysis and classification tools are designed and how they work [Breitwieser et al., 2017]. More importantly, they also influence and have direct relation to any result, profile or classification obtained. Despite huge efforts towards sequencing more species [Gilbert et al., 2014, Lewin et al., 2018], we are far from having a comprehensive set of references of all living organisms or even one representative for each species, especially when considering the unknown microbial dark matter. The current taxonomic classification, especially for prokaryotes, has flaws and limitations. Those both factors should always be considered when using reference-based tools in real applications. Even though metagenomics is providing the means to explore the microbial world as never before, if those factors are overlooked, researchers are at risk of obtaining wrong or strongly biased results and their interpretation can lead to awry discoveries.

5.1 Outlook

Based on the lessons learned during the development of this thesis and considering the fast development of metagenomics, further improvements to the methods presented here are foreseeable. When DUDes was conceptualized and developed (2014-2015), current data repositories had less sequences than they have today. Analyzing the NCBI data growth depicted in Figure 2.1, the complete genome sequences from Archaea and Bacteria in RefSeq in June 2015 had a size of around 10 Gbp accounting for approximately 1500 different species. Indexing such dataset with the read mapper Bowtie2 [Langmead and Salzberg, 2012] could be done in a matter of hours. Roughly 2 years later, in September 2017, the same repository had 30 Gbp of data representing around 3000 unique species. The time necessary to build an index with this dataset is now shifted to days, a time frame which makes such task unpractical for many applications. Such increase in data volume is limiting traditional read mapping-based tools applications, as it is harder to keep such indices up-to-date with the current volume of data provided by public repositories. DUDes, a read mapper-based tool is affected by such issues. However, DUDes does not rely on a specific read mapper and could be integrated with the recently developed DREAM-Yara [Dadi et al., 2018]. That could be beneficial to make use of the most recent data releases, giving more sensitive and precise results. In addition, big data generates big indices which, in general, lead to slower mapping times. DUDes could have an improvement in analysis performance by using alternative pseudo-alignment methods like MetaKallisto [Schaeffer et al., 2017]. However, the alignment coverage calculation would have to be changed. Further, DUDes is currently using the old NCBI Taxonomy standards where every strain used to get a taxonomic identifier [Federhen et al., 2014] where nowadays only species get taxonomic identifiers. This is currently limiting the number of strains that can be classified and would require small changes to integrate the new definitions.

I foresee MetaMeta not only as an integration tool, but also as a complete pipeline to execute, standardize, benchmark, compare and visualize results. MetaMeta was initially developed, pre-configured and tested with six initial tools. Still, the pipeline was developed in a way that the addition of new methods is straight forward. Templates are provided for developers who want to include their tools into the pipeline in an open git repository. The more tools included into the pipeline, the more options for users to benefit from multiple methods in one single application. The final integrated results are very likely to improve in terms of sensitivity and precision with the addition of diverse approaches. Further, MetaMeta could also be linked to standard evaluation datasets and test cases, scoring and evaluating each new tool added as well as verifying the changes caused to the final profile. Integration with NCBI repositories for direct download of reference sequences would also provide more autonomy to the pipeline. Regarding the methodology, MetaMeta currently does not account for unmapped or unclassified reads, always normalizing abundances to 100%. This approach was necessary due to the fact that not every tool reports the number of reads classified. This functionality would be of great benefit to improve the precision of the abundance estimates, providing more realistic numbers when profiling metagenomes.

Ganon innovates the way of indexing references sequences by using an interleaved bloom filter, allowing fast index and classification. To make these indices efficient, a clustering step is necessary. In the current implementation, ganon uses a taxonomy-oriented clustering method based on sequence sizes. Each cluster contains sequences related to a certain taxon and the sum of those sequences cannot surpass a defined threshold. The choice of this threshold is limited by the biggest reference sequence in the set. Implementations are already underway to allow sliced references, which would increase the number of clusters but at the same time reduce the overall k-mer content among them, providing a smaller and more efficient index. Compression of such indices is also being evaluated and could be beneficial, since interleaved bloom filters are in many cases in our applications sparse structures. Gapped k-mers and reduction of the reference k-mer usage are currently being evaluated and could provide an even smaller index, speeding up classification without harming sensitivity. With some improvements and changes on ganon's underlying data structure I can foresee the application of the tool as an approximate read mapper, where a read would not only be assigned to a taxon or assembly reference, but also to the approximate position in the genome sequence. In addition, a sequence-based instead of a taxonomic-oriented clustering would make ganon's index more succinct, due to the further reduction of unique k-mer for each cluster. Such implementation is currently possible but would only make sense when used together with a purely sequence-based taxonomy.

The aforementioned sequence-based taxonomy could be beneficial for all other methods presented in this thesis. As a general outlook for the field of metagenomics, I consider essential to make a bigger effort towards a restructured view of the taxonomy for prokaryotes, built purely based on whole genome sequences and their similarities. This change would avoid misclassifications given the dichotomy between the way taxonomy is built and the way analysis are commonly done. An entirely sequence-based taxonomy was already proposed [Varghese et al., 2015, Parks et al., 2018] with advantages over the current polyphasic and mostly marker gene based approaches [Chun and Rainey, 2014, Whitman, 2015, Garrity, 2016]. Such a taxonomy would allow tools to perform analyses in a complete genomic-centric fashion. However current methods are mostly taxonomy dependent and it would be difficult to introduce such a change without a community agreement on a new classification. Such an effort will only be definitive and replace state-of-the-art taxonomy once researchers as a community adopt new standards and connect the new taxonomy to the already obtained sequences currently available.

The availability of reference genome sequences has a huge impact on how most methods work and how their final results are. Computational biologists working in metagenomics, analyzing data or developing methods should keep in mind: what would change if all strains had reference sequences available and what would change if just half of the current references were there. Yet, those issues are not easy to evaluate and account for. For reference-based tools it is important to be up-to-date with the most references available to give a better view of the diversity already known and convert it into discoveries. Also, approaches to account for the unbalanced distribution of sequences along the tree of life are crucial. In addition, methods should make it clear to the final user what kind of references are available and were used to generate a specific result. In MetaMeta I developed the concept of database profiles, where it is possible to track what every tool is capable of classifying in terms of strain and taxonomic content. That way, MetaMeta evaluates each tool based on what they can achieve. At the same time, it is clear what each tool uses as background data, providing better understanding of the final results. A step further in this direction would be a way to provide a linkage between samples and databases, keeping track of every change in the database and updating the analysis results whenever new reference sequences are added.

Reference-free and function-based metagenomics are fields in active development due to many recent advances in binning methods and metagenome-assembled genomes. Reference-free methods bypass the nomenclature and classification issues of reference-based methods. In such analyses, names or labels are not important but organism groups which perform similar functions are. The function-based analysis dissociates metagenomics from labels and provides a true sequence-based classification. This way, environments would be classified by their ability to perform a number of functions, independently of which strains are present or not. Those analysis, recently named as Metagenomics 2.0 [Watson, 2018] can be linked with the early days of metagenomics, where many expectations were created around the technology which would provide understanding of microbial communities as a whole and not as a collection of independent organisms.

Appendix

A. Appendix for Chapter 3

Methods

Tools and parameters

Diginorm (normalize-by-median.py) - khmer (1.0.1) - screed (0.7): -k 20 -C 20 -N 4 -x 8e+09

Bowtie (2.2.4): –fast –no-unal -k 60 -p 8

DUDes (v0.05):

HMP: -m 50 -a 0.000005 -l species -t 36
64-organism: -t 12
STEC: -m 50 -a 0.000005 -l strain
Tara: default

GOTTCHA (1.0b):

HMP: –minCov 0.000005 –cCov 0.000005 -t 36
64-organism: -t 12

kraken (0.10.4-beta):

HMP: –threads 36
64-organism: –threads 12

MetaPhlAn (2.2.0):

HMP: –no_map –ignore_viruses –ignore_eukaryotes –nproc 36
64-organism: –no_map –ignore_viruses –ignore_eukaryotes –nproc 12

bin generation

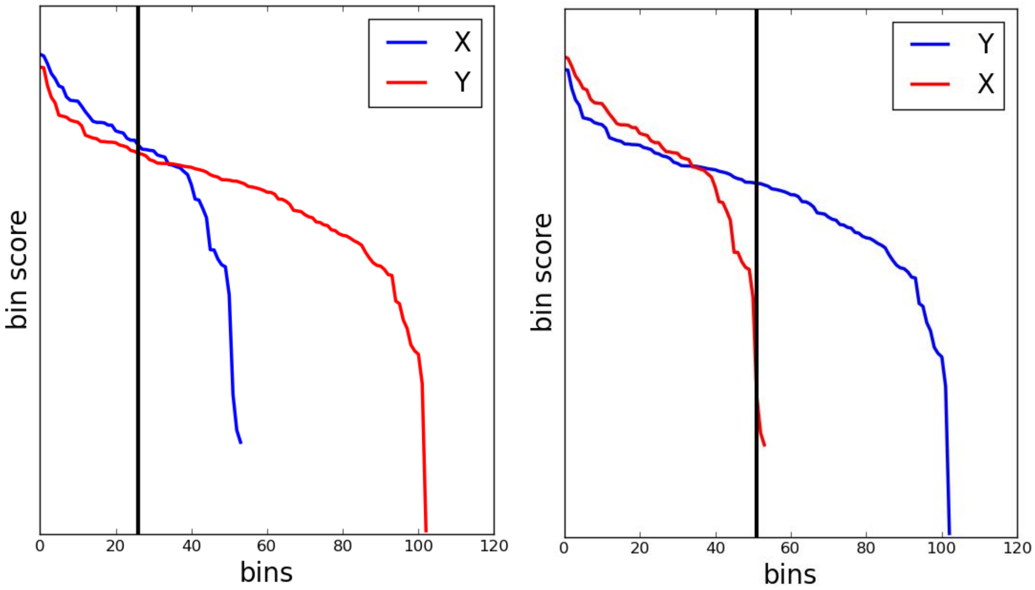


Figure 5.1: Difference in cutoff selection. The comparison of a node X against node Y can have a different cutoff value from the comparison of the same node Y against X. In both figures, the blue line represents the bins of the main node of the comparison. When X is compared against Y (a), the cutoff (black line) is 27 (top 50% bins). Although, when comparing Y against X (b) the cutoff is 52

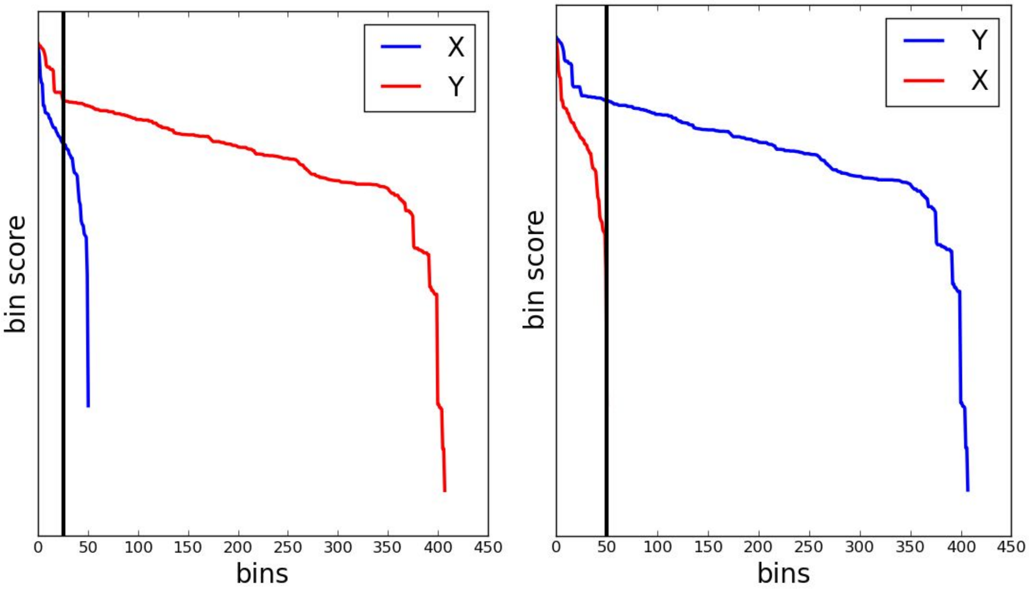


Figure 5.2: Cutoff selection without enough bins. In both figures, the blue line represents the bins of the main node of the comparison. In the first comparison of X against Y (a), the cutoff value is 26 (black line). In the second comparison of Y against X (b) the cutoff should be 204 (50% from total bins of Y) but X has not enough bins. In this scenario, the cutoff value is reset to 51 (total number of X's bins).

p-value estimation

Each node i will have a set of p-values pv_i, one for each comparison against j nodes in a taxonomic level, calculated as:

$$pv_{ij} = \frac{\sum_{n=1}^{perm} \mathbb{1}(rnd_n \geq obs_{ij})}{perm} \tag{5.1}$$

where $perm$ is the total number of permutations and $\mathbb{1}$ is the indicator function. Every random permutation rnd that achieves higher or equal difference of means than the observed difference obs is counted and the total sum is divided by the number of permutations. This function will provide a one-sided (right-tailed) p-value for each pair of nodes.

Experiments

Sample description	Accession	# Reads	Reference
HMP Illumina staggered mock	SRR172903	7.932.819	[Turnbaugh et al., 2007]
64-organism Bacteria and Archaea	SRR606249	109.629.496	[Shakya et al., 2013]
STEC sample 1122	ERR262939	1.758.352	[Loman et al., 2013]
STEC sample 1253	ERR260476	9.949.304	
STEC sample 2535	ERR260478	664.514	
STEC sample 2638	ERR262943	1.009.732	
Tara oceans sample 078 mesopelagic zone	ERR599159	288.321.845	[Sunagawa et al., 2015]

Table 5.1: Detailed information about the datasets used in this project

Results

HMP mock community

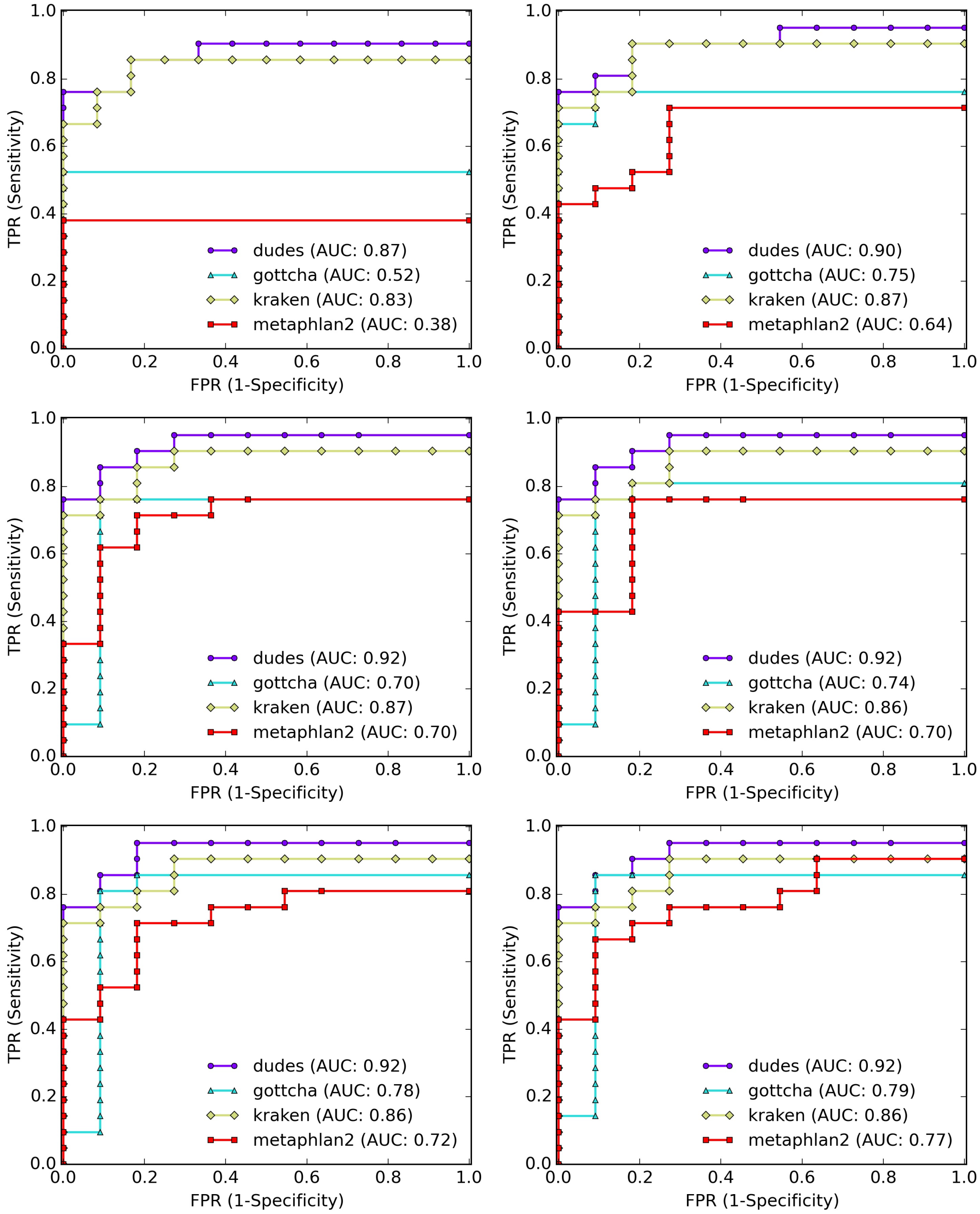


Figure 5.3: ROC curves comparing results based on six sub-sets (1%, 10%, 25%, 50%, 75% and 100%) of the normalized HMP Illumina set of reads

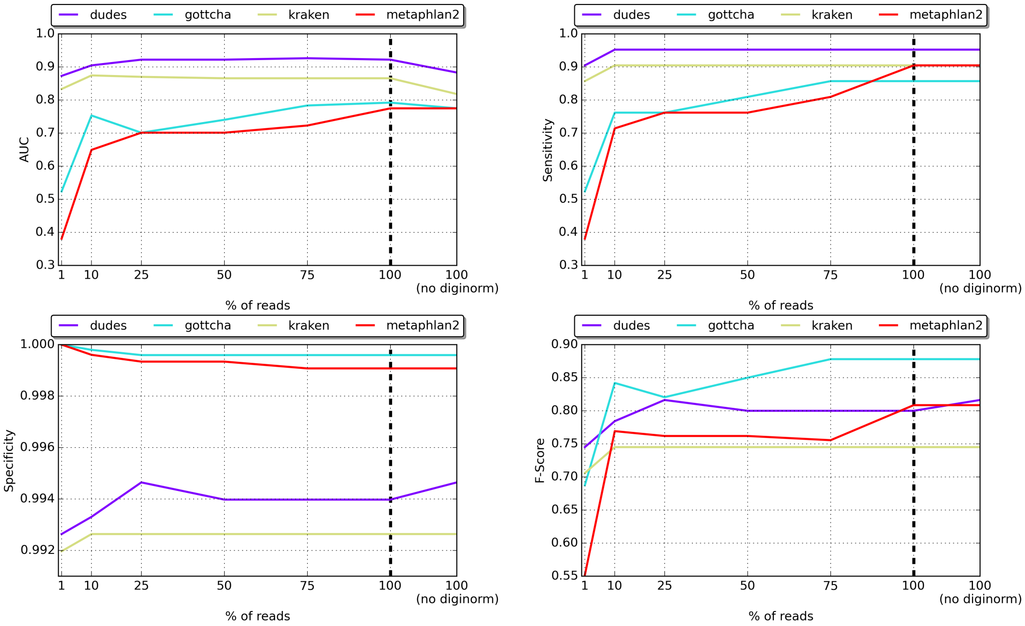


Figure 5.4: Results for AUC, Sensitivity, Specificity and F-Score values. 7 data points were evaluated for each tool (1%, 10%, 25%, 50%, 75%, 100% and 100% raw [without digital normalization]) representing the percentage of reads in each sub-set

Sub-set	# Reads	# Matches
1%	65.075	951.812
10%	650.758	9.374.136
25%	1.626.896	23.454.747
50%	3.253.792	46.846.450
75%	4.880.688	70.246.246
100%	6.507.585	93.678.968
100% no diginorm	7.932.819	117.358.571

Table 5.2: Number of reads and matches from the read mapping for each subset of the normalized Illumina HMP staggered mock community

Shiga-toxigenic *Escherichia coli* (STEC)

Samples	2535		2638	
Starting tax. level (taxid)	# candidate strains	STEC relative abundance	# candidate strains	STEC relative abundance
root (1)	47	10%	37	26%
s.kingdom (2)	49	10%	37	26%
phylum (1224)	14	19%	8	31%
class (1236)	12	19%	8	31%
order (91347)	10	19%	7	31%
family (543)	10	19%	7	31%
genus (561)	2	19%	2	31%
species (562)	1	19%	1	31%

Table 5.3: DUDes' STEC identification in two samples with different starting taxonomic levels. "# candidate strains" stands for all strains of all identified organisms with relative abundance above 0.01%

B. Appendix for Chapter 4

Implementation

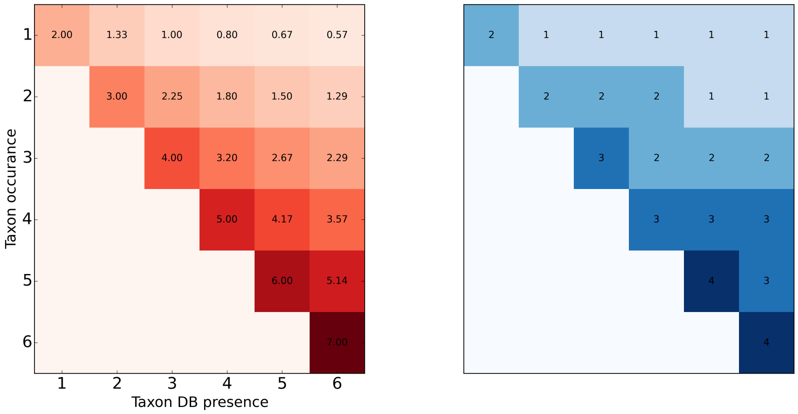


Figure 5.5: Score and bin matrices: Left: Matrix with an example of calculated scores for 6 tools. Right: matrix showing the division of the scores into 4 bins

File formats

MetaMeta accepts BioBoxes format directly (https://github.com/bioboxes/rfc/tree/master/data-format) or a .tsv file in the following format:

Profiling: rank, taxon name or taxid, abundance

Example:

genus	Methanospirillum	0.0029
genus	Thermus	0.0029
genus	568394	0.0029
species	Arthrobacter sp. FB24	0.0835
species	195	0.0582
species	Mycoplasma gallisepticum	0.0536

Binning: readid, taxon name or taxid, lenght of sequence assigned

Example:

<table>
<tr><td>M2|S1|R140</td><td>354</td><td>201</td></tr>
<tr><td>M2|S1|R142</td><td>195</td><td>201</td></tr>
<tr><td>M2|S1|R145</td><td>457425</td><td>201</td></tr>
<tr><td>M2|S1|R146</td><td>562</td><td>201</td></tr>
<tr><td>M2|S1|R147</td><td>1245471</td><td>201</td></tr>
<tr><td>M2|S1|R150</td><td>354</td><td>201</td></tr>
</table>

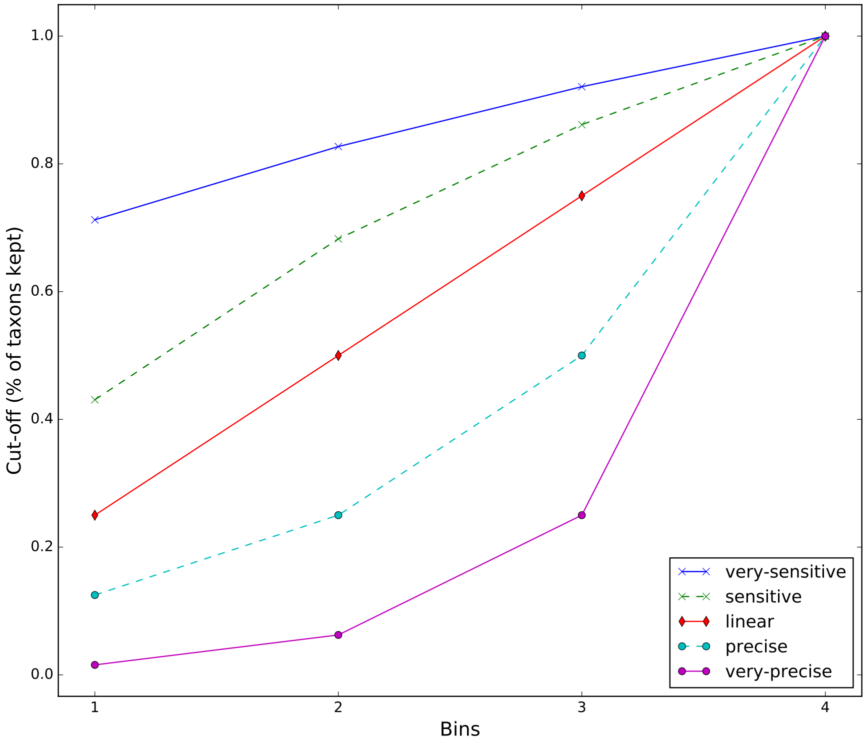


Figure 5.6: Example of cut-off values for 4 bins in each mode

Mode functions

The mode parameter can be selected among 5 different functions, that would generate more precise or sensitive results (Figure 5.6). Each *bin* will have a cut-off value C defined as:

Very-sensitive: $C_{bin} = log(bin + 3)/log(maxbins + 3)$
Sensitive: $C_{bin} = log(bin + 1)/log(maxbins + 1)$
Linear: $C_{bin} = bin/maxbins$
Precise: $C_{bin} = 2^{bin}/2^{maxbins}$
Very-precise: $C_{bin} = 4^{bin}/4^{maxbins}$
where $maxbins$ is the total number of bins.

Results

Databases

Computer specifications

The main evaluations were performed with MetaMeta v1.1 on a x86 cluster consisting of of a total of 1000 cores and roughly 3.5 TB RAM. The sub-sampling evaluations on CAMI

Table 5.4: MetaMeta pre-configured databases

Tool	Archaea + Bacteria (v1)	Custom
CLARK	Yes (https://doi.org/10.5281/zenodo.819305)	Yes
DUDes	Yes (https://doi.org/10.5281/zenodo.819343)	Yes
GOTTCHA	Yes (https://doi.org/10.5281/zenodo.819341)	No
kaiju	Yes (https://doi.org/10.5281/zenodo.819425)	Yes
kraken	Yes (https://doi.org/10.5281/zenodo.819363)	Yes
mOTUs	Yes (https://doi.org/10.5281/zenodo.819365)	No

data were performed with MetaMeta v1.0 on: 60 CPUs x Intel(R) Xeon(R) CPU E7-4890 v2 @ 2.80GHz, 1056 GB RAM, Debian GNU/Linux 8.4, 2.8 TB SSD.

Datasets and Parameters

MetaMeta pipeline was executed with all 6 pre-configured tools using the archaea and bacteria database (Table 5.4).

All CAMI toy sets (low, medium and high complexity) were obtained from https://data.cami-challenge.org/

148 stool samples from HMP were obtained at: http://hmpdacc.org/

List of analyzed samples: SRS011061, SRS011134, SRS011239, SRS011271, SRS011302, SRS011405, SRS011452, SRS011529, SRS011586, SRS012273, SRS012902, SRS013158, SRS013215, SRS013476, SRS013521, SRS013687, SRS013800, SRS013951, SRS014235, SRS014287, SRS014313, SRS014459, SRS014613, SRS014683, SRS014923, SRS014979, SRS015065, SRS015133, SRS015190, SRS015217, SRS015264, SRS015369, SRS015578, SRS015663, SRS015782, SRS015794, SRS015854, SRS015960, SRS016018, SRS016056, SRS016095, SRS016203, SRS016267, SRS016335, SRS016495, SRS016517, SRS016585, SRS016753, SRS016954, SRS016989, SRS017103, SRS017191, SRS017247, SRS017307, SRS017433, SRS017521, SRS017701, SRS017821, SRS018133, SRS018313, SRS018351, SRS018427, SRS018575, SRS018656, SRS018817, SRS019030, SRS019161, SRS019267, SRS019397, SRS019582, SRS019601, SRS019685, SRS019787, SRS019910, SRS019968, SRS020233, SRS020328, SRS020869, SRS021484, SRS021948, SRS022071, SRS022137, SRS022524, SRS022609, SRS022713, SRS023346, SRS023526, SRS023583, SRS023829, SRS023914, SRS023971, SRS024009, SRS024075, SRS024132, SRS024265, SRS024331, SRS024388, SRS024435, SRS024549, SRS024625, SRS042284, SRS042628, SRS043001, SRS043411, SRS043701, SRS045004, SRS045645, SRS045713, SRS047014, SRS047044, SRS048164, SRS048870, SRS049164, SRS049712, SRS049900, SRS049959, SRS049995, SRS050299, SRS050422, SRS050752, SRS050925, SRS051031, SRS051882, SRS052027, SRS052697, SRS053214, SRS053335, SRS053398, SRS054590, SRS054956, SRS055982, SRS056259, SRS056519, SRS057478, SRS057717, SRS058723, SRS058770, SRS062427, SRS063040, SRS063985, SRS064276, SRS064557, SRS064645, SRS065504, SRS075398, SRS077730, SRS078176

The sample SRS023176 couldn't be analyzed due to inconsistent read pairs.

Results

Table 5.5: MetaMeta (v1.1) parameters used for the CAMI and HMP data. Default parameters were used when not stated below.

	Default	CAMI low/med./high	HMP
trimming	0	-	-
desiredminlen	70	-	-
subsample	0	-	-
mode	linear	-	sensitive
cutoff	0.0001	-	0.00001
bins	4	-	-
ranks	species	-	-

Table 5.6: MetaMeta (v1.0) parameters used for the sub-sampled CAMI data. Default parameters were used when not stated below. N/A: not applicable

	Default	CAMI 1%	CAMI 5%	CAMI 10%	CAMI 16.6%	CAMI 25%	CAMI 50%	CAMI 100%
trimming	0	1	1	1	1	1	1	1
desiredminlen	70	-	-	-	-	-	-	-
strictness	0.8	-	-	-	-	-	-	-
errorcorr	0	-	-	-	-	-	-	-
subsample	0	1	1	1	1	1	1	-
samplesize	1	0.01	0.05	0.1	-	0.25	0.5	N/A
replacement	0	-	-	-	-	1	1	N/A
mode	linear	precise	precise	precise	precise	precise	precise	precise
cutoff	0.0001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001	0.00001
bins	4	3	3	3	3	3	3	3
ranks	species	-	-	-	-	-	-	-

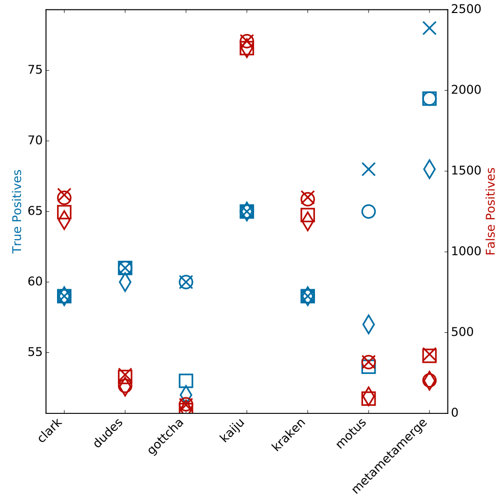


Figure 5.7: **True and False Positives - CAMI medium complexity set** In blue (left y axis): True Positives. In red (right y axis): False Positives. Results at species level. Each marker represents one out of four samples from the CAMI medium complexity set.

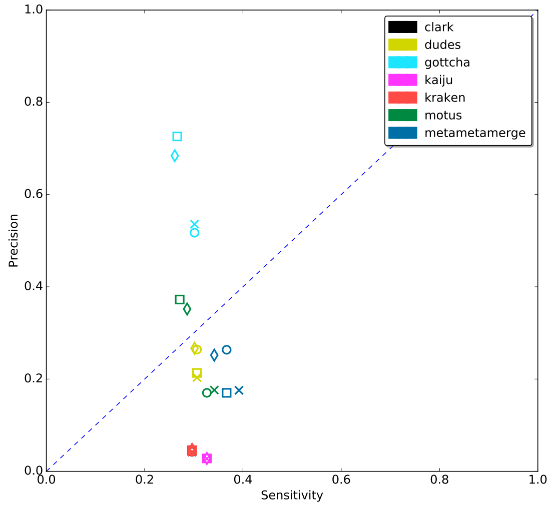


Figure 5.8: **Precision and Sensitivity - CAMI medium complexity set** Results at species level. Each marker represents one out of four samples from the CAMI medium complexity set.

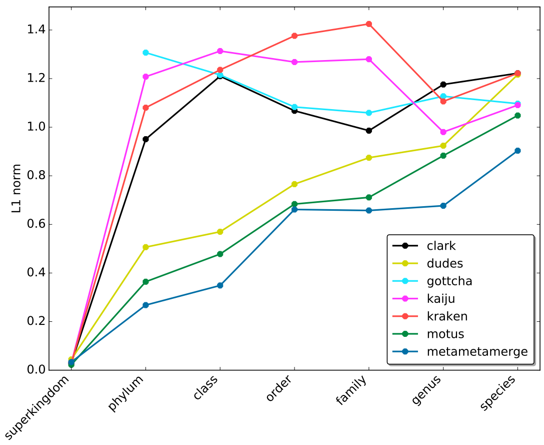


Figure 5.9: L_1 **norm error** Mean of the L_1 norm measure at each taxonomic level for four samples from the medium complexity CAMI set.

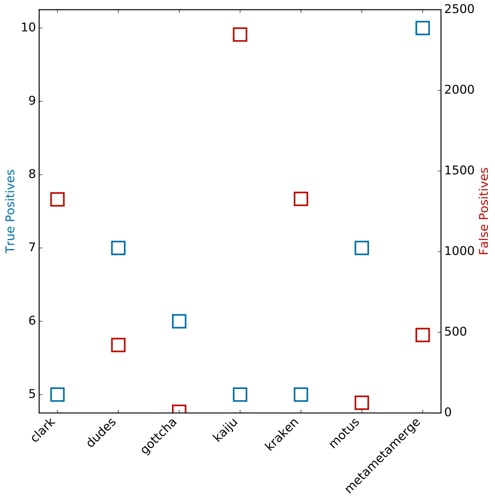


Figure 5.10: **True and False Positives - CAMI low complexity set** In blue (left y axis): True Positives. In red (right y axis): False Positives. Results at species level.

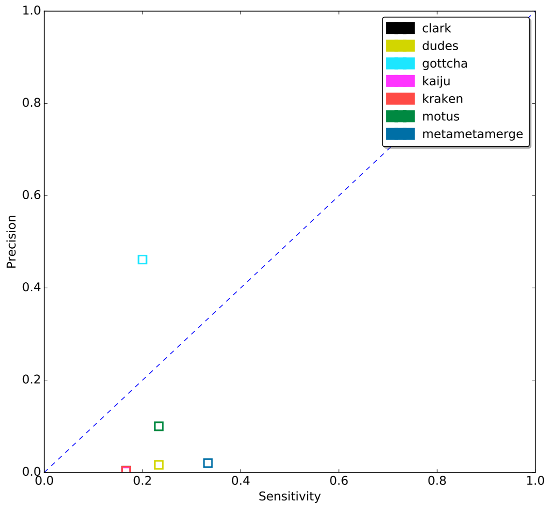


Figure 5.11: **Precision and Sensitivity - CAMI low complexity set** Results at species level.

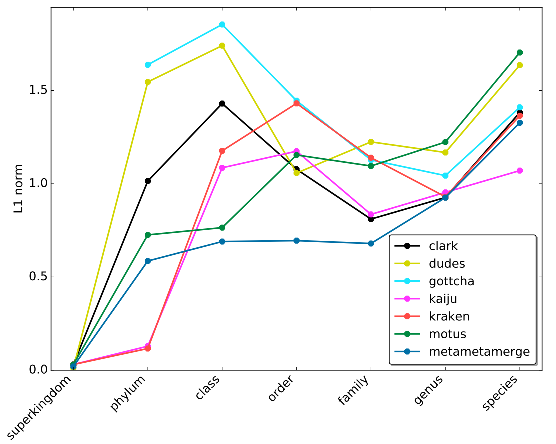


Figure 5.12: L_1 **norm error** L_1 norm measure at each taxonomic level for one sample from the low complexity CAMI set.

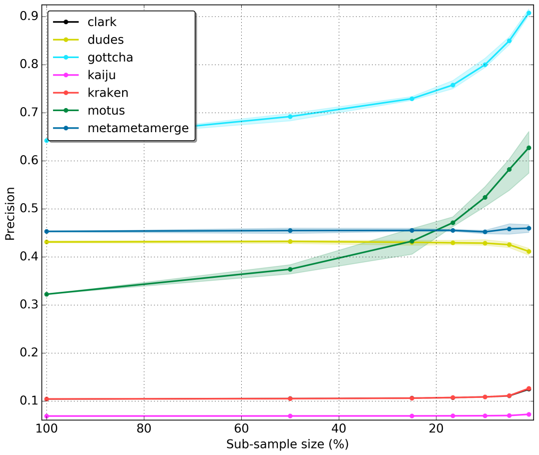


Figure 5.13: **Sub-sampling** Precision at species level for one randomly selected CAMI high complexity sample. Each sub-sample was executed five times. Lines represent the mean and the area around it the maximum and minimum achieved values. The evaluated sample sizes are: 100%, 50%, 25%, 16.6%, 10%, 5%, 1%. 16.6% is the exact division among 6 tools, using the the whole sample. Sub-samples above that value were taken with replacement and below without replacement.

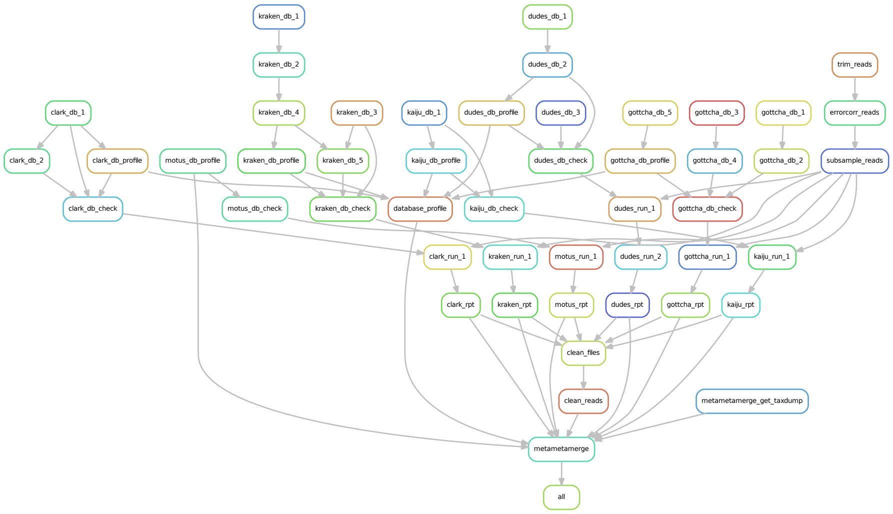


Figure 5.14: **Rulegraph** Overview of the rules and their dependencies on MetaMeta pipeline.

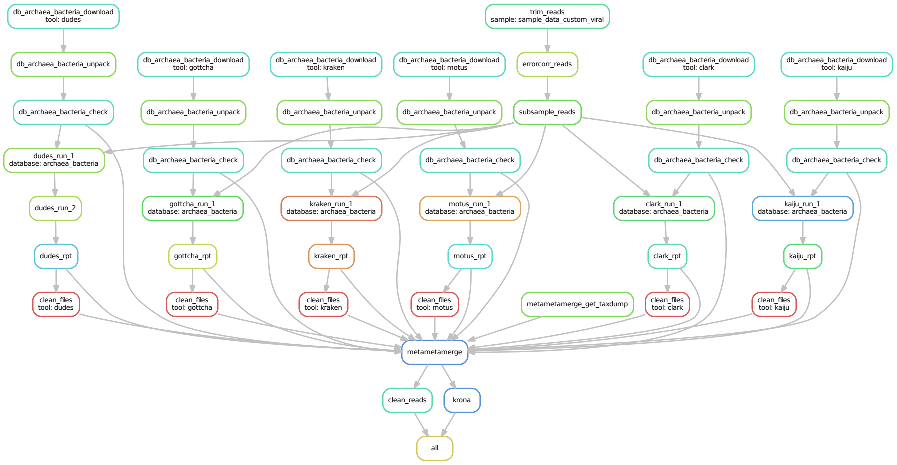

db_archaea_bacteria_download
tool: dudes
trim_reads
sample: sample_data_custom_viral
db_archaea_bacteria_unpack
db_archaea_bacteria_download
tool: gottcha
db_archaea_bacteria_download
tool: kraken
db_archaea_bacteria_download
tool: motus
errorcorr_reads
db_archaea_bacteria_download
tool: clark
db_archaea_bacteria_download
tool: kaiju
db_archaea_bacteria_check
db_archaea_bacteria_unpack
db_archaea_bacteria_unpack
db_archaea_bacteria_unpack
subsample_reads
db_archaea_bacteria_unpack
db_archaea_bacteria_unpack
dudes_run_1
database: archaea_bacteria
db_archaea_bacteria_check
db_archaea_bacteria_check
db_archaea_bacteria_check
db_archaea_bacteria_check
db_archaea_bacteria_check
dudes_run_2
gottcha_run_1
database: archaea_bacteria
kraken_run_1
database: archaea_bacteria
motus_run_1
database: archaea_bacteria
clark_run_1
database: archaea_bacteria
kaiju_run_1
database: archaea_bacteria
dudes_rpt
gottcha_rpt
kraken_rpt
motus_rpt
clark_rpt
kaiju_rpt
clean_files
tool: dudes
clean_files
tool: gottcha
clean_files
tool: kraken
clean_files
tool: motus
metametamerge_get_taxdump
clean_files
tool: clark
clean_files
tool: kaiju
metametamerge
clean_reads
krona
all

Figure 5.15: **DAG - pre-configured database** Directed acyclic graph of the MetaMeta pipeline for one sample, one database (pre-configured) and six tools.

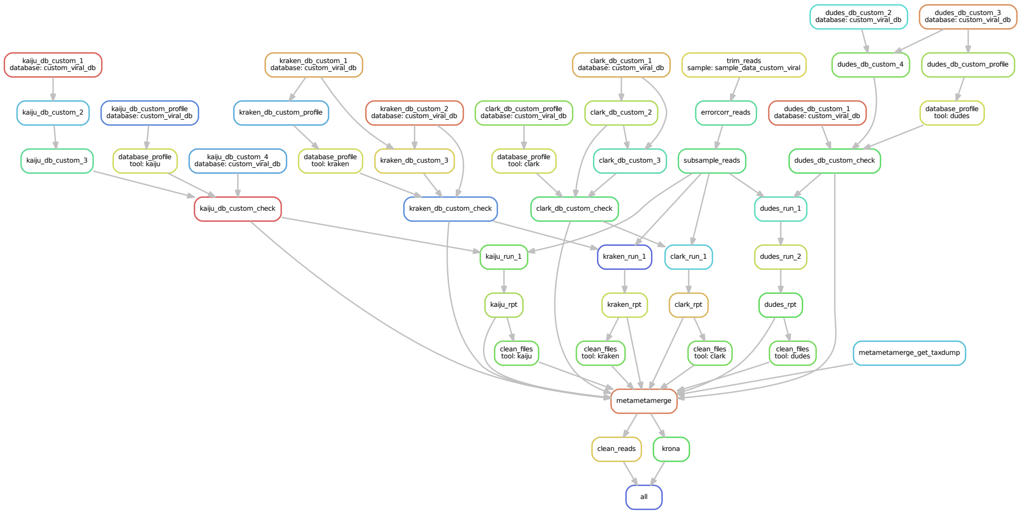


Figure 5.16: **DAG - custom database** Directed acyclic graph of the MetaMeta pipeline for one sample, one database (custom) and 4 tools.

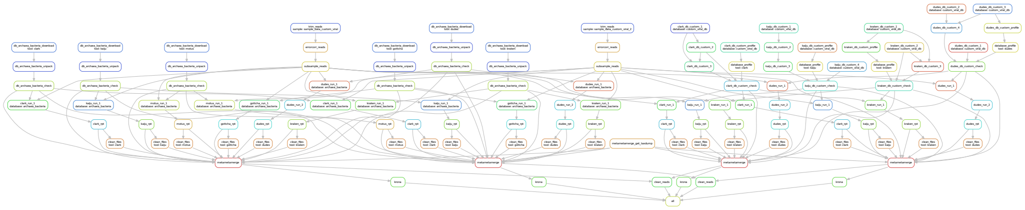


Figure 5.17: **DAG - multiple samples** Directed acyclic graph of the MetaMeta pipeline for two samples, two databases (pre-configured and custom) and six tools.

Bibliography

J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, nov 2014. ISSN 1548-7091. doi: 10.1038/nmeth.3103. URL http://www.nature.com/articles/nmeth.3103.

D. A. Antonopoulos, R. Assaf, R. K. Aziz, T. Brettin, C. Bun, N. Conrad, J. J. Davis, E. M. Dietrich, T. Disz, S. Gerdes, R. W. Kenyon, D. Machi, C. Mao, D. E. Murphy-Olson, E. K. Nordberg, G. J. Olsen, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, J. Santerre, M. Shukla, R. L. Stevens, M. VanOeffelen, V. Vonstein, A. S. Warren, A. R. Wattam, F. Xia, and H. Yoo. PATRIC as a unique resource for studying antimicrobial resistance. *Briefings in Bioinformatics*, jul 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx083. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx083.

S. Ardui, A. Ameur, J. R. Vermeesch, and M. S. Hestand. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, 46(5):2159–2168, mar 2018. ISSN 0305-1048. doi: 10.1093/nar/gky066. URL https://academic.oup.com/nar/article/46/5/2159/4833218.

D. A. Baltrus. Divorcing Strain Classification from Species Names. *Trends in Microbiology*, 24(6):431–439, jun 2016. ISSN 0966842X. doi: 10.1016/j.tim.2016.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0966842X16000408.

M. Balvočiūtė and D. H. Huson. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics*, 18(S2):114, mar 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3501-4. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3501-4.

A. L. Bazinet and M. P. Cummings. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92, 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-92. URL http://www.biomedcentral.com/1471-2105/13/92.

P. Belmann, J. Dröge, A. Bremges, A. C. McHardy, A. Sczyrba, and M. D. Barton. Bioboxes: standardised containers for interchangeable bioinformatics software. *GigaScience*, 4(1):47, dec 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0087-0. URL https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0087-0.

D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36–D42, nov 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1195. URL http://academic.oup.com/nar/article/41/D1/D36/1068219/GenBank.

D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K. D. Pruitt, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 46(D1):D41–D47, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1094. URL http://academic.oup.com/nar/article/46/D1/D41/4621329.

A. Blanco-Míguez, F. Fdez-Riverola, B. Sánchez, and A. Lourenço. Resources and tools for the high-throughput, multi-omic study of intestinal microbiota. *Briefings in Bioinformatics*, nov 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx156. URL http://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbx156/4665692.

B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970. ISSN 00010782. doi: 10.1145/362686.362692. URL http://portal.acm.org/citation.cfm?doid=362686.362692.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, aug 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu170. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170.

F. P. Breitwieser and S. L. Salzberg. KrakenHLL: Confident and fast metagenomics classification using unique k-mer counts. *bioRxiv*, page 262956, 2018. doi: 10.1101/262956. URL https://www.biorxiv.org/content/early/2018/02/09/262956.

F. P. Breitwieser, J. Lu, and S. L. Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, sep 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx120. URL http://academic.oup.com/bib/article/doi/10.1093/bib/bbx120/4210288/A-review-of-methods-and-databases-for-metagenomic.

C. Brown, A. Howe, and Q. Zhang. A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv*, mar 2012. URL http://arxiv.org/abs/1203.4802.

J. Chun and F. A. Rainey. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 64(Pt 2):316–324, feb 2014. ISSN 1466-5026. doi: 10.1099/ijs.0.054171-0. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.054171-0.

B. Codenotti, G. De Marco, M. Leoncini, M. Montangero, and M. Santini. Approximation algorithms for a hierarchically structured bin packing problem. *Information Processing Letters*, 89(5):215–221, mar 2004. ISSN 00200190. doi: 10.1016/j.ipl.2003.12.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0020019003005301.

N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 43(D1):D6–D17, jan 2015. ISSN 1362-4962. doi: 10.1093/nar/gku1130. URL https://academic.oup.com/nar/article/43/D1/D6/2438293.

N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1):D7–D19, jan 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1290. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1290.

T. H. Dadi, B. Y. Renard, L. H. Wieler, T. Semmler, and K. Reinert. SLIMM: species level identification of microorganisms from metagenomes. *PeerJ*, 5:e3138, mar 2017. ISSN 2167-8359. doi: 10.7717/peerj.3138. URL https://peerj.com/articles/3138.

T. H. Dadi, E. Siragusa, V. C. Piro, A. Andrusch, E. Seiler, B. Y. Renard, and K. Reinert. DREAM-Yara: an exact read mapper for very large databases with short update time. *Bioinformatics*, 34(17):i766–i772, sep 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty567. URL https://academic.oup.com/bioinformatics/article/34/17/i766/5093228.

E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, and N. C. Kyrpides. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology*, 1(4):15032, feb 2016. ISSN 2058-5276. doi: 10.1038/nmicrobiol.2015.32. URL http://www.nature.com/articles/nmicrobiol201532.

S. Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136—-D143, jan 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr1178. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr1178.

S. Federhen, K. Clark, T. Barrett, H. Parkinson, J. Ostell, Y. Kodama, J. Mashima, Y. Nakamura, G. Cochrane, and I. Karsch-Mizrachi. Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Standards in Genomic Sciences*, 9(3):1275–1277, jan 2014. ISSN 1944-3277. doi: 10.4056/sigs.4851102. URL http://www.standardsingenomics.org/index.php/sigen/article/view/sigs.4851102.

S. Federhen, R. Rossello-Mora, H.-P. Klenk, B. J. Tindall, K. T. Konstantinidis, W. B. Whitman, D. Brown, D. Labeda, D. Ussery, G. M. Garrity, R. R. Colwell, N. Hasan, J. Graf, A. Parte, P. Yarza, B. Goldberg, H. Sichtig, I. Karsch-Mizrachi, K. Clark, R. McVeigh, K. D. Pruitt, T. Tatusova, R. Falk, S. Turner, T. Madden, P. Kitts, A. Kimchi, W. Klimke, R. Agarwala, M. DiCuccio, and J. Ostell. Meeting report: GenBank microbial genomic taxonomy workshop (12–13 May, 2015). *Standards in Genomic Sciences*, 11(1):15, dec 2016. ISSN 1944-3277. doi: 10.1186/s40793-016-0134-1. URL http://www.standardsingenomics.com/content/11/1/15.

O. E. Francis, M. Bendall, S. Manimaran, C. Hong, N. L. Clement, E. Castro-Nallar, Q. Snell, G. B. Schaalje, M. J. Clement, K. A. Crandall, and W. E. Johnson. Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Research*, 23(10):1721–1729, oct 2013. ISSN 1088-9051. doi: 10.1101/gr.150151.112. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.150151.112.

T. A. K. Freitas, P.-E. P.-E. Li, M. B. Scholz, and P. S. G. Chain. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Research*, 43(10):e69–e69, may 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv180. URL http://academic.oup.com/nar/article/43/10/e69/2409024/Accurate-readbased-metagenome-characterization.

W. F. Fricke and D. a. Rasko. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics*, 15(1):49–55, jan 2014. ISSN 1471-0056. doi: 10.1038/nrg3624. URL http://www.nature.com/articles/nrg3624.

J. L. Gardy and N. J. Loman. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nature Reviews Genetics*, 19(1):9–20, nov 2017. ISSN 1471-0056. doi: 10.1038/nrg.2017.88. URL http://www.nature.com/doifinder/10.1038/nrg.2017.88.

G. M. Garrity. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? *Journal of Clinical Microbiology*, 54(8):1956–1963, aug 2016. ISSN 0095-1137. doi: 10.1128/JCM.00200-16. URL http://jcm.asm.org/lookup/doi/10.1128/JCM.00200-16.

J. A. Gilbert, J. K. Jansson, and R. Knight. The Earth Microbiome project: successes and aspirations. *BMC Biology*, 12(1):69, dec 2014. ISSN 1741-7007. doi: 10.1186/s12915-014-0069-1. URL http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-014-0069-1.

G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(November):1–6, nov 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.02224. URL http://journal.frontiersin.org/article/10.3389/fmicb.2017.02224/full.

J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, may 2014. ISSN 02776715. doi: 10.1002/sim.6082. URL http://doi.wiley.com/10.1002/sim.6082.

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, jun 2016. ISSN 1471-0056. doi: 10.1038/nrg.2016.49. URL http://www.nature.com/articles/nrg.2016.49.

W. W. Greenwald, N. Klitgord, V. Seguritan, S. Yooseph, J. C. Venter, C. Garner, K. E. Nelson, and W. Li. Utilization of defined microbial communities enables effective evaluation of meta-genomic assemblies. *BMC Genomics*, 18(1):296, dec 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3679-5. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3679-5.

B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, and J. Köster. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15(7):475–476, jul 2018. ISSN 1548-7091. doi: 10.1038/s41592-018-0046-7. URL http://www.nature.com/articles/s41592-018-0046-7.

D. H. Haft, M. DiCuccio, A. Badretdin, V. Brover, V. Chetvernin, K. O'Neill, W. Li, F. Chitsaz, M. K. Derbyshire, N. R. Gonzales, M. Gwadz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, C. Zheng, F. Thibaud-Nissen, L. Y. Geer, A. Marchler-Bauer, and K. D. Pruitt. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, 46(D1):D851–D860, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1068. URL http://academic.oup.com/nar/article/46/D1/D851/4588110.

J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, oct 1998. ISSN 10745521. doi: 10.1016/S1074-5521(98)90108-9. URL http://linkinghub.elsevier.com/retrieve/pii/S1074552198901089.

A. Howe and P. S. G. Chain. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Frontiers in Microbiology*, 6, jul 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00678. URL http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00678/abstract.

D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, feb 2007. ISSN 1088-9051. doi: 10.1101/gr.5969107. URL http://www.genome.org/cgi/doi/10.1101/gr.5969107.

C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, M. G. Giglio, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. Bonazzi, J. Paul Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. G. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. J. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, S. Kinder Haake, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, N. B. King, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C. A. Lozupone, R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavromatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya,

J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. Pop, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, B. A. Methé, K. E. Nelson, J. F. Petrosino, G. M. Weinstock, R. K. Wilson, and O. White. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, jun 2012. ISSN 0028-0836. doi: 10.1038/nature11234. URL http://www.nature.com/articles/nature11234.

P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. *Mathematical Foundations of Computer Science 1991*, pages 240–248, 1991. ISSN 3540543457. URL http://www.springerlink.com/index/p58155n8012x0477.pdf.

I. Karsch-Mizrachi, T. Takagi, and G. Cochrane. OUP accepted manuscript. *Nucleic Acids Research*, 46(D1):D48–D51, 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx1097. URL https://academic.oup.com/nar/article/46/D1/D48/4668651.

P. J. Keeling, F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. a. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. a. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyhrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. a. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz, K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. B. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E. Umale, D. Vaulot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone, and A. Z. Worden. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, 12(6):e1001889, jun 2014. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001889. URL http://dx.plos.org/10.1371/journal.pbio.1001889.

D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, dec 2016. ISSN 1088-9051. doi: 10.1101/gr.210641.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.210641.116.

I. Klymiuk, C. Högenauer, B. Halwachs, G. G. Thallinger, W. F. Fricke, and C. Steininger. A physicians' wish list for the clinical application of intestinal metagenomics. *PLoS medicine*, 11(4):e1001627, apr 2014. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001627. URL http://dx.plos.org/10.1371/journal.pmed.1001627.

C. U. Köser, M. J. Ellington, E. J. P. Cartwright, S. H. Gillespie, N. M. Brown, M. Farrington, M. T. G. Holden, G. Dougan, S. D. Bentley, J. Parkhill, and S. J. Peacock. Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. *PLoS Pathogens*, 8(8):e1002824, aug 2012. ISSN 1553-7374. doi: 10.1371/journal.ppat.1002824. URL http://dx.plos.org/10.1371/journal.ppat.1002824.

J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, oct 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480.

M. Kuhring, P. W. Dabrowski, V. C. Piro, A. Nitsche, and B. Y. Renard. SuRankCo: supervised ranking of contigs in de novo assemblies. *BMC Bioinformatics*, 16(1):240, dec 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0644-7. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0644-7.

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/articles/nmeth.1923.

S. Lee, H. Min, and S. Yoon. Will solid-state drives accelerate your bioinformatics? In-depth profiling, performance analysis and beyond. *Briefings in Bioinformatics*, 17(4):713–727, jul 2016. ISSN 1467-5463. doi: 10.1093/bib/bbv073. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv073.

V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, feb 1966. URL https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, and G. Zhang. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, apr 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720115115. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1720115115.

D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10):1674–1676, may 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv033. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv033.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352.

J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, N. Pons, E. Le Chatelier, J.-M. Batto, S. Kennedy, F. Haimet, Y. Winogradski, E. Pelletier, D. LePaslier, F. Artiguenave, T. Bruls, J. Weissenbach, K. Turner, J. Parkhill, M. Antolin, F. Casellas, N. Borruel, E. Varela, A. Torrejon, G. Denariaz, M. Derrien, J. E. T. van Hylckama Vlieg, P. Viega, R. Oozeer, J. Knoll, M. Rescigno, C. Brechot, C. M'Rini, A. Mérieux, T. Yamada, S. Tims, E. G. Zoetendal, M. Kleerebezem, W. M. de Vos, A. Cultrone, M. Leclerc, C. Juste, E. Guedon, C. Delorme, S. Layec, G. Khaci, M. van de Guchte, G. Vandemeulebrouck, A. Jamet, R. Dervyn, N. Sanchez, H. Blottière, E. Maguin, P. Renault, J. Tap, D. R. Mende, P. Bork, and J. Wang. An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32(8):834–841, jul 2014. ISSN 1087-0156. doi: 10.1038/nbt.2942. URL http://www.nature.com/doifinder/10.1038/nbt.2942.

X. Li, S. A. Naser, A. Khaled, H. Hu, and X. Li. When old metagenomic data meet newly sequenced genomes, a case study. *PLOS ONE*, 13(6):e0198773, jun 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0198773. URL http://dx.plos.org/10.1371/journal.pone.0198773.

S. Lindgreen, K. L. Adair, and P. P. Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, 6(1):19233, may 2016. ISSN 2045-2322. doi: 10.1038/srep19233. URL http://www.nature.com/articles/srep19233.

M. S. Lindner and B. Y. Renard. Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS. *PLOS ONE*, 10(2):e0117711, feb 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0117711. URL http://dx.plos.org/10.1371/journal.pone.0117711.

K. J. Locey and J. T. Lennon. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, may 2016. ISSN 0027-8424. doi: 10.1073/pnas.1521291113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1521291113.

N. J. Loman, C. Constantinidou, M. Christner, H. Rohde, J. Z.-M. Chan, J. Quick, J. C. Weir, C. Quince, G. P. Smith, J. R. Betley, M. Aepfelbacher, and M. J. Pallen. A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4. *JAMA*, 309(14):1502, apr 2013. ISSN 0098-7484. doi: 10.1001/jama.2013.3231. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2013.3231.

S. S. Mande, M. H. Mohammed, and T. S. Ghosh. Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, 13(6):669–81, nov 2012.

ISSN 1477-4054. doi: 10.1093/bib/bbs054. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbs054.

J. R. Marchesi and J. Ravel. The vocabulary of microbiome research: a proposal. *Microbiome*, 3(1):31, dec 2015. ISSN 2049-2618. doi: 10.1186/s40168-015-0094-5. URL http://www.microbiomejournal.com/content/3/1/31.

A. B. R. McIntyre, R. Ounit, E. Afshinnekoo, R. J. Prill, E. Hénaff, N. Alexander, S. S. Minot, D. Danko, J. Foox, S. Ahsanuddin, S. Tighe, N. A. Hasan, P. Subramanian, K. Moffat, S. Levy, S. Lonardi, N. Greenfield, R. R. Colwell, G. L. Rosen, and C. E. Mason. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biology*, 18(1):182, dec 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1299-7. URL http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1299-7.

N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, feb 2008. ISSN 0006-3444. doi: 10.1093/biomet/asn007. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asn007.

P. Menzel, K. L. Ng, and A. Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7:11257, apr 2016. ISSN 2041-1723. doi: 10.1038/ncomms11257. URL http://www.nature.com/doifinder/10.1038/ncomms11257.

B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, J. H. Badger, A. T. Chinwalla, A. M. Earl, M. G. FitzGerald, R. S. Fulton, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. R. Bonazzi, P. Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, S. Kinder-Haake, N. B. King, R. Knight, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C. A. Lozupone, R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavrommatis, J. M.

McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Qing Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, G. M. Weinstock, R. K. Wilson, and O. White. A framework for human microbiome research. *Nature*, 486(7402):215–221, jun 2012. ISSN 0028-0836. doi: 10.1038/nature11209. URL http://www.nature.com/articles/nature11209.

A. A. Metwally, Y. Dai, P. W. Finn, and D. L. Perkins. WEVOTE: Weighted Voting Taxonomic Identification Method of Microbial Sequences. *PLOS ONE*, 11(9):e0163527, sep 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0163527. URL http://dx.plos.org/10.1371/journal.pone.0163527.

S. Mukherjee, R. Seshadri, N. J. Varghese, E. A. Eloe-Fadrosh, J. P. Meier-Kolthoff, M. Göker, R. C. Coates, M. Hadjithomas, G. A. Pavlopoulos, D. Paez-Espino, Y. Yoshikuni, A. Visel, W. B. Whitman, G. M. Garrity, J. A. Eisen, P. Hugenholtz, A. Pati, N. N. Ivanova, T. Woyke, H.-P. Klenk, and N. C. Kyrpides. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature Biotechnology*, 35(7):676–683, jun 2017. ISSN 1087-0156. doi: 10.1038/nbt.3886. URL http://www.nature.com/doifinder/10.1038/nbt.3886.

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/articles/nrg3367.

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–e155, nov 2012. ISSN 1362-4962. doi: 10.1093/nar/gks678. URL https://academic.oup.com/nar/article/40/20/e155/2414459.

D. J. Nasko, S. Koren, A. M. Phillippy, and T. J. Treangen. RefSeq database growth influences the accuracy of k-mer-based species identification. *bioRxiv*, pages 1–21, 2018. doi: 10.1101/304972. URL https://www.biorxiv.org/content/early/2018/04/19/304972.

S. Nayfach and K. S. Pollard. Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, 166(5):1103–1116, aug 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.08.007. URL http://dx.doi.org/10.1016/j.cell.2016.08.007.

S. I. Nikolenko, A. I. Korobeynikov, and M. a. Alekseyev. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, 14(Suppl 1):S7, 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-S1-S7. URL http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-S1-S7.

S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner. metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, 27(5):824–834, may 2017. ISSN 1088-9051. doi: 10.1101/gr.213959.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.213959.116.

N. D. Olson, T. J. Treangen, C. M. Hill, V. Cepeda-Espinoza, J. Ghurye, S. Koren, and M. Pop. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*, aug 2017. ISSN 1467-5463. doi: 10.1093/bib/bbx098. URL http://academic.oup.com/bib/article/doi/10.1093/bib/bbx098/4075034/Metagenomic-assembly-through-the-lens-of.

B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, 12(1):385, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-385. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-385.

A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and L. Iliopoulos. Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, 9:BBI.S12462, jan 2015. ISSN 1177-9322. doi: 10.4137/BBI.S12462. URL http://journals.sagepub.com/doi/10.4137/BBI.S12462.

R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236, dec 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1419-2. URL http://www.biomedcentral.com/1471-2164/16/236.

I. Pagani, K. Liolios, J. Jansson, I.-M. a. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research*, 40 (D1):D571–D579, jan 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr1100. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr1100.

M. J. Pallen. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*, 141(14):1856–1862, dec 2014. ISSN 0031-1820. doi: 10.1017/S0031182014000134. URL http://www.journals.cambridge.org/abstract{_}S0031182014000134.

D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542, nov 2017. ISSN 2058-5276. doi: 10.1038/s41564-017-0012-7. URL http://www.nature.com/articles/s41564-017-0012-7.

D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, aug 2018. ISSN 1087-0156. doi: 10.1038/nbt.4229. URL http://www.nature.com/doifinder/10.1038/nbt.4229.

M. A. Peabody, T. Van Rossum, R. Lo, and F. S. L. Brinkman. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, 16(1):362, dec 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0788-5. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0788-5.

Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, jun 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts174. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts174.

V. C. Piro, M. S. Lindner, and B. Y. Renard. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics*, 32(15):2272–2280, aug 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw150. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw150.

V. C. Piro, M. Matschkowski, and B. Y. Renard. MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome*, 5(1):101, dec 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0318-y. URL http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0318-y.

V. C. Piro, T. H. Dadi, E. Seiler, K. Reinert, and B. Y. Renard. Ganon: Continuously Up-To-Date With Database Growth for Precise Short Read Classification in Metagenomics. *bioRxiv*, 2018. doi: 10.1101/406017. URL https://www.biorxiv.org/content/early/2018/08/31/406017.

O. Prakash, M. Verma, P. Sharma, M. Kumar, K. Kumari, A. Singh, H. Kumari, S. Jit, S. K. Gupta, M. Khanna, and R. Lal. Polyphasic approach of bacterial classification — An overview of recent advances. *Indian Journal of Microbiology*, 47(2):98–108, jun 2007. ISSN 0046-8991. doi: 10.1007/s12088-007-0022-x. URL http://link.springer.com/10.1007/s12088-007-0022-x.

C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844, 2017. ISSN 1087-0156. doi: 10.1038/nbt.3935. URL http://www.nature.com/doifinder/10.1038/nbt.3935.

D. Ramasamy, A. K. Mishra, J.-C. Lagier, R. Padhmanabhan, M. Rossi, E. Sentausa, D. Raoult, and P.-E. Fournier. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 64(Pt 2):384–391, feb 2014. ISSN 1466-5026. doi: 10.1099/ijs.0.057091-0. URL http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.057091-0.

M. Ravenhall, N. Škunca, F. Lassalle, and C. Dessimoz. Inferring Horizontal Gene Transfer. *PLOS Computational Biology*, 11(5):e1004095, may 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004095. URL http://dx.plos.org/10.1371/journal.pcbi.1004095.

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, aug 2015. ISSN 1527-8204. doi: 10.1146/annurev-genom-090413-025358. URL http://www.annualreviews.org/doi/10.1146/annurev-genom-090413-025358.

K. Reinert, T. H. Dadi, M. Ehrhardt, H. Hauswedell, S. Mehringer, R. Rahn, J. Kim, C. Pockrandt, J. Winkler, E. Siragusa, G. Urgese, and D. Weese. The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. *Journal of Biotechnology*, 261(July):157–168, nov 2017. ISSN 01681656. doi: 10.1016/j.jbiotec.2017.07.017. URL https://linkinghub.elsevier.com/retrieve/pii/S0168165617315420.

R. J. Robbins, L. Krishtalka, and J. C. Wooley. Advances in biodiversity: metagenomics and the unveiling of biological dark matter. *Standards in Genomic Sciences*, 11(1):69, dec 2016. ISSN 1944-3277. doi: 10.1186/s40793-016-0180-8. URL http://standardsingenomics.biomedcentral.com/articles/10.1186/s40793-016-0180-8.

R. Rosselló-Móra and R. Amann. Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, 38(4):209–216, jun 2015. ISSN 07232020. doi: 10.1016/j.syapm.2015.02.001. URL https://linkinghub.elsevier.com/retrieve/pii/S0723202015000223.

L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter. Pseudoalignment for metagenomic read assignment. *Bioinformatics*, 33(14):2082–2088, jul 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx106. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx106.

R. Schlaberg, C. Y. Chiu, S. Miller, G. W. Procop, and G. Weinstock. Validation of Metagenomic Next-Generation Sequencing Tests for Universal Pathogen Detection. *Archives of Pathology & Laboratory Medicine*, 141(6):776–786, jun 2017. ISSN 0003-9985. doi: 10.5858/arpa.2016-0539-RA. URL http://www.archivesofpathology.org/doi/10.5858/arpa.2016-0539-RA.

A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software.

Nature Methods, 14(11):1063–1071, oct 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4458. URL http://www.nature.com/doifinder/10.1038/nmeth.4458.

N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, jun 2012. ISSN 1548-7091. doi: 10.1038/nmeth.2066. URL http://www.nature.com/articles/nmeth.2066.

M. Shakya, C. Quince, J. H. Campbell, Z. K. Yang, C. W. Schadt, and M. Podar. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, 15(6):1882–1899, jun 2013. ISSN 14622912. doi: 10.1111/1462-2920.12086. URL http://doi.wiley.com/10.1111/1462-2920.12086.

J.-i. Sohn and J.-W. Nam. The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, oct 2016. ISSN 1467-5463. doi: 10.1093/bib/bbw096. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbw096.

R. D. Stewart, M. D. Auffret, A. Warr, A. H. Wiser, M. O. Press, K. W. Langford, I. Liachko, T. J. Snelling, R. J. Dewhurst, A. W. Walker, R. Roehe, and M. Watson. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*, 9(1):870, dec 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-03317-6. URL http://www.nature.com/articles/s41467-018-03317-6.

S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. a. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Doré, S. D. Ehrlich, A. Stamatakis, and P. Bork. Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12):1196–1199, dec 2013. ISSN 1548-7091. doi: 10.1038/nmeth.2693. URL http://www.nature.com/articles/nmeth.2693.

S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. D'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon. Structure and function of the global ocean microbiome. *Science*, 348(6237): 1261359–1261359, may 2015. ISSN 0036-8075. doi: 10.1126/science.1261359. URL http://www.sciencemag.org/content/348/6237/1261359.short.

S. H. Tausch, B. Strauch, A. Andrusch, T. P. Loka, M. S. Lindner, A. Nitsche, and B. Y. Renard. LiveKraken—real-time metagenomic classification of illumina data. *Bioinformatics*, jun 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty433. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty433/5026657.

M. Teraguchi, N. Yoshimura, H. Hashizume, S. Muraki, H. Yamada, A. Minamide, H. Oka, Y. Ishimoto, K. Nagata, R. Kagotani, N. Takiguchi, T. Akune, H. Kawaguchi, K. Nakamura, and M. Yoshida. Prevalence and distribution of intervertebral disc degeneration over the entire spine in a population-based cohort: the Wakayama Spine Study. *Osteoarthritis and Cartilage*, 22(1):104–110, jan 2014. ISSN 10634584. doi: 10.1016/j.joca.2013.10.019. URL http://linkinghub.elsevier.com/retrieve/pii/S1063458413010029.

D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10):902–903, oct 2015. ISSN 1548-7091. doi: 10.1038/nmeth.3589. URL http://www.nature.com/articles/nmeth.3589.

D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, and N. Segata. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research*, 27(4):626–638, apr 2017. ISSN 1088-9051. doi: 10.1101/gr.216242.116. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.216242.116.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, oct 2007. ISSN 0028-0836. doi: 10.1038/nature06244. URL http://www.nature.com/doifinder/10.1038/nature06244.

N. J. Varghese, S. Mukherjee, N. Ivanova, K. T. Konstantinidis, K. Mavrommatis, N. C. Kyrpides, and A. Pati. Microbial species delineation using whole genome sequences. *Nucleic Acids Research*, 43(14):6761–6771, aug 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv657. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv657.

J. Vollmers, S. Wiegand, and A.-K. Kaster. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLOS ONE*, 12(1):e0169662, jan 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0169662. URL http://dx.plos.org/10.1371/journal.pone.0169662.

M. Watson. Microbiome 2.0 - or what to do when you have no hits, feb 2018. URL https://naturemicrobiologycommunity.nature.com/users/83344-mick-watson/posts/30668-microbiome-2-0-or-what-to-do-when-you-have-no-hits-to-public-databases.

W. B. Whitman. Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Systematic and Applied Microbiology*, 38(4):217–222, jun 2015. ISSN 07232020. doi: 10.1016/j.syapm.2015.02.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0723202015000247.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, mar 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-3-r46. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46.

Y.-w. Wu, B. a. Simmons, and S. W. Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, feb 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv638. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv638.

M. L. Zepeda Mendoza, T. Sicheritz-Pontén, and M. T. P. Gilbert. Environmental genes and genomes: understanding the differences and challenges in the approaches and software for their analyses. *Briefings in Bioinformatics*, 16(5):745–758, sep 2015. ISSN 1467-5463. doi: 10.1093/bib/bbv001. URL https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv001.

Zusammenfassung

Metagenomik bietet die Möglichkeit, die große und noch weitgehend unbekannte Welt der Mikroben zu untersuchen. Diese machen mindestens die Hälfte der genetischen Vielfalt der Erde aus. Computergestützte Metagenomik ermöglicht diese Entdeckungen durch die Analyse von großen Datenmengen, die durch High-throughput-sequencing in einem schnellen Tempo generiert werden. Referenzbasierte Methoden werden üblicherweise verwendet, um Umweltproben basierend auf zuvor zusammengestellten Referenzsequenzen zu untersuchen, die oft mit einer taxonomischen Klassifikation verbunden sind. Den Ursprung jedes sequenzierten Fragments zu finden und die Umweltprobe als Gesamtes zu beschreiben ist das Hauptziel von Binningtools und taxonomischer Profilingtools.

In dieser Arbeit präsentiere ich drei Methoden der computergestützten Metagenomik. Kuratierter Referenzsequenzen und taxonomische Klassifikation werden zur Charakterisierung von Umweltproben verwendet. Das Hauptziel dieser Beiträge ist es, den Stand der Technik des taxonomischen Profiling und Binning mit schnellen, sensiblen und präzisen Methoden zu verbessern.

Zuerst stelle ich ganon vor, ein Tool zur Sequenzklassifizierung metagenomischer Daten, welches mit einer sehr großen Anzahl von Referenzsequenzen arbeitet. ganon bietet eine effiziente Methode zur Indexierung von Sequenzen und Aktualisierung dieser Indizes in sehr kurzer Zeit. Darüber hinaus führt ganon taxonomisches Binning mit stark verbesserter Genauigkeit im Vergleich zu den derzeit verfügbaren Methoden durch. Für ein generelles Profiling metagnomischer Daten und Bestandsschätzung stelle ich DUDes vor. Statt die in der Probe vorhanden Stämme nur basiert auf relativen Häufigkeiten vorherzusagen, identifiziert DUDes zuerst mögliche Kandidaten durch Vergleichen der Konfidenz der zugewiesenen Reads in jedem Knoten des Taxonomiebaumes auf eine iterative Top-Down-Weise. Diese Technik arbeitet in entgegengesetzter Richtung des kleinsten gemeinsamen Vorfahren-Ansatzes. Am Ende der Arbeit stelle ich MetaMeta vor, eine Pipeline zur Ausführung metagenomischer Analysetools und zur Integration ihrer Ergebnisse. MetaMeta ist gleichzeitig eine Methode zur Kombination und Verbesserung von Ergebnissen aus mehreren taxonomischen Binning- und Profiling-Tools, sowie eine Pipeline zum einfachen Ausführen von Tools und Analysieren von Umweltdaten. MetaMeta umfasst eine Datenbankgenerierung, Vorbereitungs-, Ausführungs- und Integrationsschritte, die eine einfache Installation, Visualisierung und Parallelisierung von Tools auf dem neuesten Stand der Technik ermöglichen. Mit den gleichen Eingabedaten liefert MetaMeta empfindlichere und zuverlässigere Ergebnisse, wobei das Vorhandensein jedes identifizierten Organismus' von mehreren Methoden unterstützt wird. Diese drei Projekte stellen neuen Methodiken und verbesserte Ergebnisse gegenüber ähnlichen Methoden vor und leisten einen wertvollen Beitrag zur Charakterisierung von Gemeinschaften auf Referenz- und Taxonomiebasierten Methoden.

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.

Vitor C. Piro, Berlin den 12. Oktober 2018