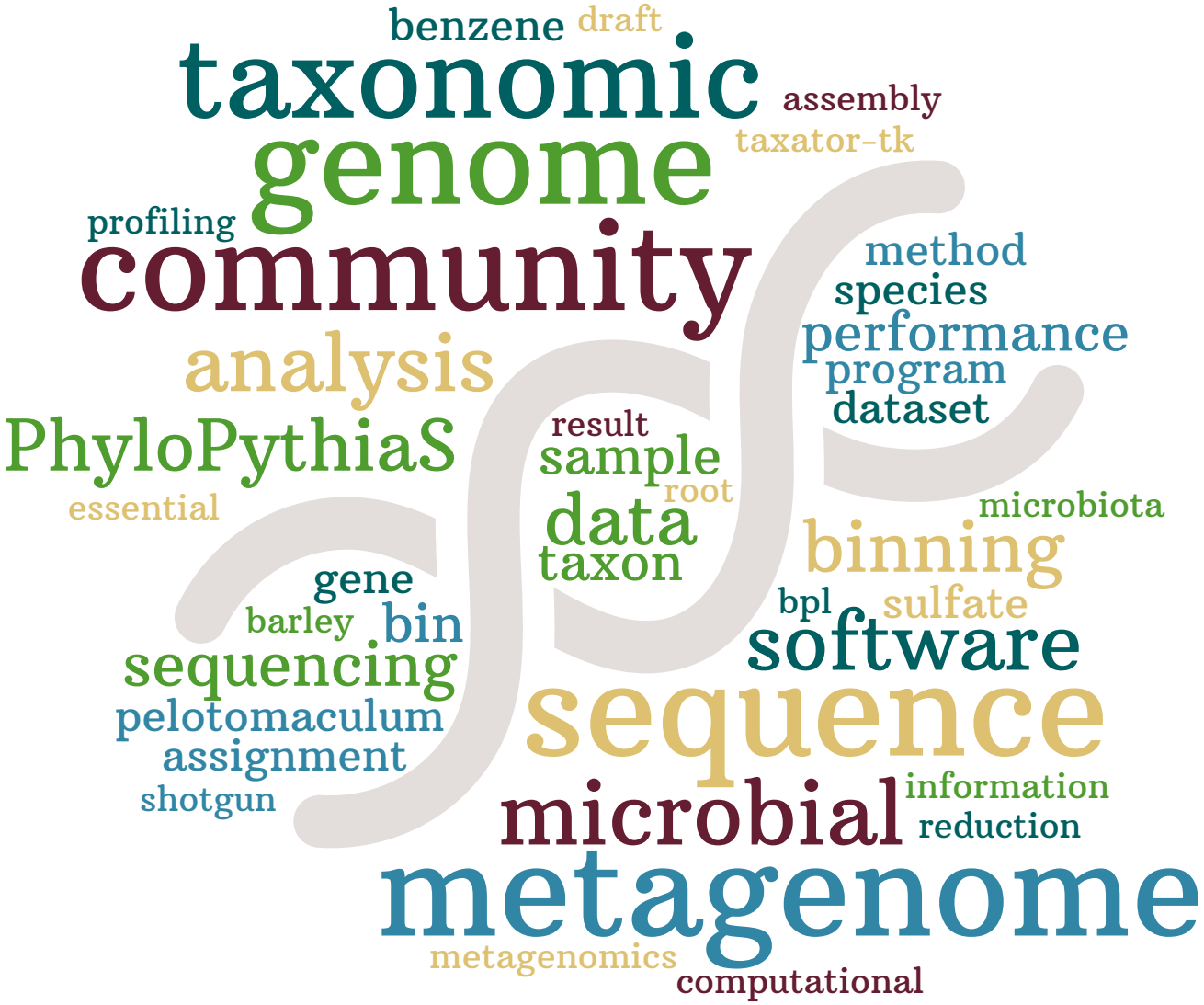

benzene
draft
taxonomic
assembly
taxator-tk
genome
profiling
community
method
species
performance
program
dataset
analysis
PhyloPythiaS
result
sample
root
essential
data
microbiota
taxon
binning
gene
barley
bin
bpl
sulfate
software
sequencing
sequence
pelotomaculum
assignment
shotgun
information
microbial
reduction
metagenome
metagenomics
computational

Cover image: tag cloud of publication abstracts
DOI: 10.1093/bioinformatics/btu745
DOI: 10.7717/peerj-cs.117
DOI: 10.1093/bib/bbs031
DOI: 10.1016/j.chom.2015.01.011
DOI: 10.7717/peerj.1603
DOI: 10.1093/femsec/fiw254

Computational Methods for Taxonomic Annotation and Genome Reconstruction in Metagenomics

Kumulative Dissertation

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

vorgelegt von

Johannes Dröge
aus Halle (Westf.)

Düsseldorf, 2017-03-31

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

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

Referent:	Prof. Dr. Alice C. McHardy
Koreferent:	Prof. Dr. Martin J. Lercher
Tag der mündlichen Prüfung:	2017-07-17

Selbstständigkeitserklärung

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

Düsseldorf , den ....
(Johannes Dröge)

Statement of authorship

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

Summary

Microbial communities can be found in almost every place, from biogas reactors over deep sea vents, the surface of plant leaves and roots, to the human body, which hosts a plethora of foreign cells in its digestion system. These communities may consist of thousands upon thousands of microorganisms, including bacteria, archaea, algae and fungi, which coexist within their habitats but which cannot simply be cultivated and studied due to their complex mutual dependencies and environmental requirements. Metagenomics is a field dedicated to the genetic analysis of such communities. The genes of their members enable their survival, for instance by making nutrients accessible, by neutralizing toxic compounds or by allowing symbiosis with other organisms. Through the use of nucleotide sequencing technologies, this genetic diversity can be explored and rendered usable, for instance in the form of new antibiotics or as enzymes in biotechnology. Apart from its considerable economic potential, metagenomic approaches lead to a fundamentally improved understanding of the microbial processes on earth. With current technology, it is not directly possible to sequence contiguous genomes from microbial communities. Instead, short sequences, called reads, are produced, which need to be assembled into genes and longer genome sequences using computer programs. Depending on the size and complexity of the metagenome, this task can be very difficult. This thesis describes two methods for assigning metagenomic sequences to taxonomic groups or genomes. The results can be used to analyze the genes, and the corresponding proteins and functions, within their phylogenetic and genetic context to gain better insight into the functioning of individual organisms and the microbial community.

Our first method, *taxator-tk*, assigns nucleotide sequences from metagenomes to corresponding taxa and approaches two challenges: the precise prediction of taxa and the application to datasets, which are constantly growing due to the rapid progress in DNA sequencing. Since annotation methods such as *taxator-tk*, which require similarity to known genomes, spend a considerable part of their runtime for sequence comparison, our algorithm exploits the underlying phylogenetic structure for similar gene sequences to efficiently calculate the taxonomic assignment. The same phylogenetic principles are used to achieve a

high assignment precision.

The second method in this thesis helps researchers to reconstruct individual genomes. It is a statistical classification model for metagenome data, for which we outline several direct and follow-up applications. These include classification of nucleotide sequences to individual genomes, *de-novo* calculation of genome clusters in metagenomes, *in-silico* sample enrichment for genomes and quality checking of reconstructed genomes. We published the method as a software library named *MGLEX* for integration into other programs to enable the efficient use of the data for reconstructing genomes in different scenarios.

Presumably, metagenomics will continue to play an important role in microbial research, and may partially obviate the sequencing of cloned strain genomes. This trend is supported by the rapid development of DNA sequencing technologies, which is progressing towards faster sequencing and longer reads. The presented methods supplement the existing set of bioinformatics tools for acquiring knowledge from metagenomes. By reducing metagenomes to individual genomes, one can apply traditional algorithms from genomics, for instance to reconstruct metabolic pathways, and one can link data from transcriptomic and proteomic experiments. Therefore, there is much interest in genome reconstruction methods, like the ones presented in this thesis.

Zusammenfassung

Mikrobielle Gemeinschaften existieren praktisch überall, in Biogas-Anlagen, heißen Quellen am Meeresgrund, auf der Oberfläche von Pflanzenblättern und -wurzeln und auch im menschlichen Körper, welcher z. B. im Verdauungstrakt an genetisch fremden Zellen ein Vielfaches seiner selbst beherbergt. Sie können aus Abertausenden von Mikroorganismen, wie Bakterien, Archäen, Algen und Pilzen, bestehen, die innerhalb ihrer Umgebung koexistieren und auf Grund ihrer komplexen wechselseitigen Abhängigkeiten und speziellen Umgebungsanforderungen nicht ohne Weiteres isoliert, kultiviert und untersucht werden können. Das Feld der Metagenomik widmet sich der genetischen Analyse dieser Gemeinschaften. Die Gene ihrer Mitglieder sichern ihnen das Überleben, indem sie unter anderem Nahrung verwertbar machen, Gifte neutralisieren oder Symbiosen mit anderen Organismen ermöglichen. Durch die Technik der Gensequenzierung kann man diesen genetischen Reichtum untersuchen und für Anwendungen nutzbar machen, z. B. in Form von neuen Antibiotika oder als Enzyme in der Biotechnologie. Abgesehen von dem großen ökonomischen Potential ermöglicht die Metagenomik ein fundamental besseres Verständnis der mikrobiologischen Prozesse auf unserer Erde.

Auf direktem Weg können nach heutigem technischen Stand noch keine zusammenhängenden Genome der mikrobiellen Gemeinschaften sequenziert werden. Vielmehr ergeben sich viele kurze DNA-Abschnitte, sogenannte Reads, die durch Computerprogramme zu Gen- und längeren Genom-Sequenzen zusammengesetzt werden müssen, was sich je nach Größe und Komplexität des Metagenoms als sehr schwierig erweisen kann. Diese Doktorarbeit beschreibt zwei Methoden, die das Ziel verfolgen, metagenomische Sequenzen bestimmten taxonomischen Gruppen oder Genomen zuzuordnen. Dadurch können die Gene bzw. ihre zugehörigen Proteine und Funktionen im phylogenetischen und genetischen Kontextes analysieren werden, um so ein besseres Verständnis der Funktionsweise der Organismen und der mikrobiellen Gemeinschaft zu erlangen.

Das erste Methode, *taxator-tk*, weist Nukleotidsequenzen aus Metagenomen bestimmten Taxa zu und begegnet dabei zwei Herausforderungen: zum einen der präzisen Vorhersage und zum anderen der Anwendbarkeit auf Datensätzen,

deren Größe mit dem rapiden Fortschritt der DNA-Sequenzierung stetig ansteigt. Annotationsmethoden wie *taxator-tk*, die auf Ähnlichkeit zu bereits bekannten Genomen setzen, benötigen einen beträchtlichen Teil ihrer Laufzeit für die Berechnung der Sequenzähnlichkeiten. Daher nutzt unser Algorithmus die zugrunde liegende phylogenetische Struktur ähnlicher Gensequenzen zur effizienten Berechnung einer taxonomischen Vorhersage. Durch die Anwendung der gleichen phylogenetischen Prinzipien erreicht er eine hohe Präzision der Vorhersagen.
Die zweite in dieser Arbeit vorgestellte Methode unterstützt Forscher bei der Rekonstruktion einzelner Genome. Es handelt sich um ein statistisches Klassifikationsmodell für Metagenomdaten, für das zahlreiche direkte und weitergehende Anwendungsmöglichkeiten skizziert werden. Diese umfassen die Klassifizierung von Nukleotidsequenzen nach Genomen, die *de-novo*-Berechnung von Genom-Clustern, die *in-silico* Anreicherung von Genomsequenzdaten und die Qualitätskontrolle rekonstruierter Genome. Die Methode wurde als Software-Bibliothek namens *MGLEX* zur Verwendung in anderen Programmen veröffentlicht und ermöglicht dadurch eine effiziente Datenverwertung bei der Rekonstruktion von Genomen in unterschiedlichen Situationen.
Es ist zu erwarten, dass die Metagenomik eine wichtige Rolle in der mikrobiologischen Forschung spielen und zunehmend in Konkurrenz zur Genomsequenzierung geklonter Stämme treten wird. Diese Prognose wird auch durch die rasante Entwicklung der DNA-Sequenziertechniken getragen, die eine immer schnellere Sequenzierung immer längerer Reads ermöglichen. Die hier vorgestellten Methoden ergänzen das Repertoire vorhandener Bioinformatik-Werkzeuge zur Gewinnung von Erkenntnissen aus Metagenomen. Die Reduzierung von Metagenomen auf einzelne Genome ermöglicht sowohl die Anwendung klassischer Algorithmen der Genomik, z. B. zur Rekonstruktion von Stoffwechselpfaden, als auch die Verknüpfung mit experimentellen Daten der Transkriptomik und Proteomik. Daher sind Verfahren zur Rekonstruktion einzelner Genome, wie sie in dieser Arbeit vorgestellt werden, von großem generellem Interesse.

Danksagung

Zahlreiche Personen haben mich im Verlauf meiner Promotion begleitet und unterstützt. Ihnen gebührt mein vollster Dank, auch wenn ich an dieser Stelle nicht alle erwähnen kann. Ohne die Hilfe, die wissenschaftliche Expertise und die Ideen von Alice McHardy wäre diese Arbeit nicht möglich gewesen. Außerdem bedanke ich mich bei meinen beiden Koautoren Alexander Schönhuth und Ivan Gregor, die mich in der Konzeption und Ausführung meiner Publikationen unterstützt haben. Unersetzlich war für mich über die gesamte Zeit meiner Promotion auch die Unterstützung durch meine Freundin Diana Rodriguez und meine Familie. Durch die gemeinsame Arbeit in unserer Forschungsgruppe habe ich mich immer bestärkt gefühlt, wofür ich mich bei meinen zahlreichen Kollegen bedanke. Hervorheben will ich Aaron Weimann und David Lähnemann für die vielen fachlichen Diskussionen, die gemeinsamen Erlebnisse und ihren fortwährenden Einsatz für ein gutes Betriebsklima. Für ihre andauernde Hilfe in organisatorischen Angelegenheiten und ihre wohlwollende Art möchte ich mich zuletzt auch bei Angela Rennwanz bedanken.

Contents

List of Figures

List of Tables

Chapter 1

Synopsis

1.1 Metagenomics

Metagenomics is a more recent variant of genomics which pursues medical or ecological questions at the scale of microbial communities using nucleotide sequencing. In contrast to microbial genomics, which is focused on single strains traditionally grown in lab cultures before genome sequencing, the metagenomic approach applies direct sampling from a natural ecosystem without cultivation. Microbes form so-called communities in their micro-environment because they interact, for instance by symbiosis (e.g. sharing metabolites) or competition (e.g. for food). Such a community may consist of hundreds or thousands of different species, which are connected by complex interactions (Berry & Widder, 2014; Fuhrman, Cram & Needham, 2015). It is the principal interest of microbial ecology to understand these interaction networks, which make it difficult to isolate and grow the organisms on culture medium because the specific cultivation conditions cannot be reproduced (Riesenfeld, Schloss & Handelsman, 2004; Stewart, 2012). However, by extracting and sequencing environmental DNA directly after sampling, one can capture the genomes of all community members, although in a highly fractional and usually incomplete form. One could say that current metagenomics trades the species-level resolution and the completeness of very few

genomes for a higher level view on the genes in a community. The metagenome, a term coined in the early 2000 (Rondon et al., 2000; Riesenfeld, Schloss & Handelsman, 2004; Tyson et al., 2004), stands for all the genes in a microbial community. These genes determine the ecological functions of the community members through the proteins they encode. Metagenome sequencing can thus collect new environmental genes and discover protein functions with potential use in medicine and biotechnology, and provides a way to understand the microbial interactions within diverse ecosystems. It has been used to study many different environments (Figure 1.1).

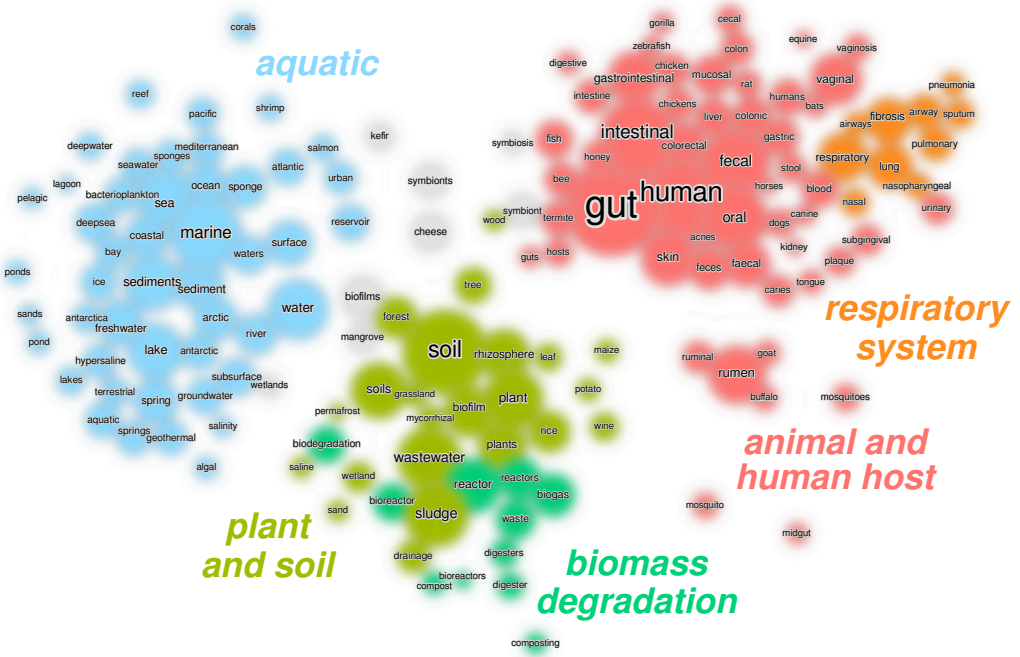


Figure 1.1: Microbial environments extracted from 10,043 publication titles (2011-2017) positioned by cooccurence in publication titles. The articles were selected by topic and the corresponding metadata downloaded from Europe PMC (europepmc.org). The titles were then reduced to environment-related words and these were grouped by the number of cooccurences using Gephi (gephi.org) with a force-directed layout and subsequent annotation of clusters. Three major clusters emerge, relating to aquatic environments, soil and plant biomass degradation and (human) host-related environments.

Early metagenomic studies have impressively demonstrated the potential of this new approach. For instance, new antibiotics and antibiotic resistance genes were identified (Gillespie et al., 2002; Riesenfeld, Goodman & Handelsman, 2004). An ocean survey (Venter et al., 2004) revealed hundreds of new rhodopsin-like genes in seawater environments (rhodopsin is an essential protein to sensing light) among over 1.2 million novel genes. In the following, numerous micro-environments were explored to provide a census of genes and species, many of them previously unknown. For the various sites in and on the human body, which represent well-studied environments due to medical applications, the resulting data provided new insight into the interactions between the human host and its so-called microbiome. For instance abnormal microbial colonization of the gut was observed with chronic inflammation (Qin et al., 2010). Although most investigations have focused on the bacteria, the best known domain in the microbial tree of life, metagenomics has also been used to study the genes of archaea, microscopic eukaryotes, viruses and genetic elements like plasmids (Hugenholtz & Tyson, 2008; Cuvelier et al., 2010; Garrett et al., 2010), which helped to broaden the view on the global genetic repertoire of life and its evolution.

1.1.1 DNA sequencing

Past and present progress in the field of metagenomics is tightly coupled to the development of next-generation sequencing technologies (NGS). While earlier studies were based on the Sanger sequencing technology (Wommack, Bhavsar & Ravel, 2008), the underlying chemistry has been subject to many improvements, such as the engineering of highly parallel reaction and detection procedures. This has led to an considerable drop in overall time and cost of nucleotide sequencing (Dröge & McHardy, 2012). The first sequencing approaches in metagenomics targeted well studied single genes, predominantly the bacterial and archaeal gene of the ribosomal 16S subunit (Quince, Curtis & Sloan, 2008; Hamady & Knight, 2009), which is a good taxonmic marker because it contains both conserved and divergent regions. In this context, sequences identity thresholds were applied to define operational taxonomic units (OTUs) as an approximate species replacement. The variable regions were amplified in a polymerase chain reaction (PCR)

before sequencing and are therefore called amplicons. Using this selective approach reduced the amount of target DNA from millions of bases per genome to a few hundreds while giving estimates of genetic species diversity. Amplicon sequencing is still in use and represents a cost-effective way to study the taxonomic composition and taxon abundances. However, it cannot be used to discover the functional potential unless the corresponding genome sequences are available for consideration. To target novel community genomes, universal sequencing primers initiate sequencing at random starting positions on the DNA strands. This approach is called shotgun sequencing due to the fact that the reads are more or less randomly scattered over the entire genome sequence. With a sufficient number of reads, metagenomic shotgun sequencing can cover most genes and continues to evolve together with next-generation sequencing platforms, but also with respect to experimental protocols and data analysis methods. A major limitation of current sequencing technologies is the length of the primary sequencing products (reads). In particular, the currently dominating Illumina sequencing platform produces reads which are still much shorter than typical genes (Dröge & McHardy, 2012) so that overlapping reads are typically assembled to form longer contiguous sequences (contigs) (Miller, Koren & Sutton, 2010). New technologies such as PacBio and Oxford Nanopore sequencing yield longer reads but have larger error rates and higher costs compared to Illumina, which limits their current use in metagenomics (Goodwin, McPherson & McCombie, 2016).

Metagenomic studies have highlighted the advantages of metagenomic over the traditional sequencing approach using isolated and cultured strains. The genomes of environmental microorganisms were found to be much more genetically diverse than those of corresponding lab strains (Tyson et al., 2004; Handelsman, 2004), which essentially represent clones of a single cell. Researchers also become more aware of the fact that genetic data collections are strongly biased towards taxa which are easily grown in lab cultures and which are of medical relevance, leaving many black spots in the microbial tree of life (Tyson et al., 2004; Wu et al., 2009). Using the exploratory metagenomics approach, there is no need to narrow the focus on certain species and to hypothesize about the role of these organisms in their environment beforehand. The bird's eye view on the genes helps to identify mutual dependencies, such as pathways that are connected between different

genomes (Ponomarova & Patil, 2015), and to associate new functions and new species. Apart from this, direct sequencing also creates new problems. Some sequencing platforms introduce a bias related to the nucleotide composition (Dohm et al., 2008), which may affect the analysis. In general, it is difficult to distinguish sequencing errors from natural genetic variation, which, in some cases, could lead to wrong conclusions such as inflated microbial diversity estimates (Quince et al., 2009; Kunin et al., 2010). Another problem with this sequence heterogeneity is that longer genome sequences often fail to assemble due to the natural and artificial nucleotide variations in the reads (Melsted & Pritchard, 2011; Pell et al., 2012). Typical metagenome data therefore contain many incomplete genes whose origin and functional role needs to be determined.

1.1.2 The role of computer programs

Today's genomic data are ubiquitous and abundant due to high-throughput nucleotide sequencing. Consequently, the data generation marks a starting point of knowledge discovery, making modern metagenomics in large part a data-driven science in which algorithms have replaced lab techniques to sort and analyze genetic material. Metagenome data are large (because they represent many genomes) and require extensive processing to deal with the phylogenetic and genetic diversity in the sample. It is convenient to divide the downstream processing of raw sequencing data into three consecutive steps which are illustrated in Figure 1.2: (a) sequence processing specific to the sequencing platform and often performed by proprietary software; (b) metagenome analysis and reduction to non-redundant draft genome sequences; (c) algorithms to study the individual genomes and how they interact. Step (a) applies not only to metagenomics but to all sciences using nucleotide sequencing and, from a practical perspective, decouples downstream algorithms from the specifics of sequencing technology and its development. The work presented in this thesis contributes to step (b), to prepare the data for use in downstream algorithms in step (c), which are tailored to the biological questions.

An important step following nucleotide sequencing is the assembly of overlapping

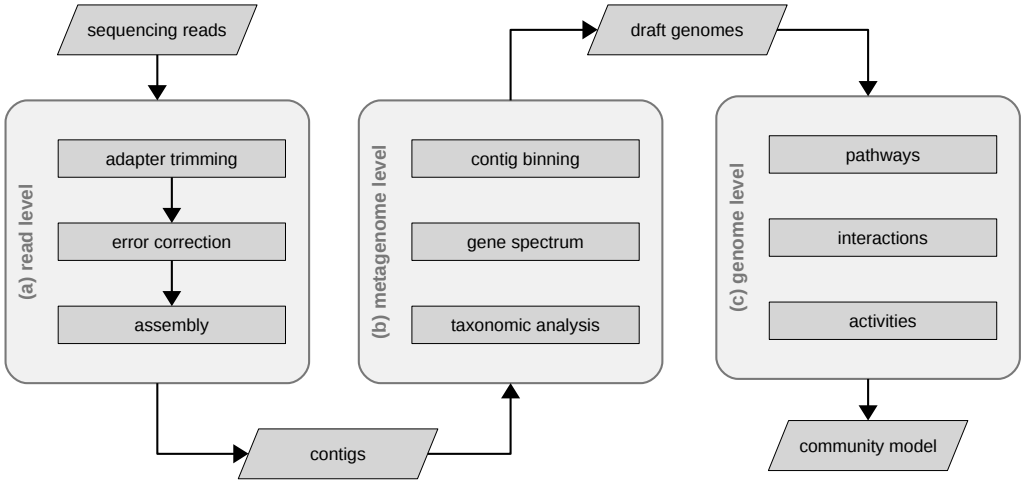


Figure 1.2: Major steps in metagenome data processing. Typical processing consists of three consecutive levels: (a) read processing (b) contig analysis and binning and (c) the analysis at the genome level.

reads into longer contigs. For this, many reads must be sequenced to cover the corresponding genome positions. In current Illumina sequencing protocols, pairs of reads are typically linked in the experimental library preparation (Goodwin, McPherson & McCombie, 2016) to capture their relative orientation and approximate distance (insert size). This information helps to construct longer contigs, because otherwise repetitive regions or homologous genes which are longer than the read length cannot be distinguished if they cause loops in the assembly graph (Ghurye, Cepeda-Espinoza & Pop, 2016). When the read coverage drops for intermediate regions, the corresponding genomes also break into multiple shorter contigs. Existing assemblers for isolate genome assembly, which has been available for a long time (Sutton et al., 1995; Huang & Madan, 1999), has been adjusted to assemble metagenomes (Ghurye, Cepeda-Espinoza & Pop, 2016). Metagenome assemblers must cope with the natural genetic variance of strains compared to clonal DNA and must also take into account that, due to different abundances in the sample, the number of genome copies varies considerably among the species or strains, resulting in a large range of read coverages. The assembly of reads for complex communities is considered an algorithmic challenge, but often reduces

the amount of data considerably and produces a fraction of longer contigs which represent full or partial genes. Assembly is therefore a reasonable first step towards recovering the full genome sequence of environmental microbes. In the workflow Figure 1.2, the assembly bridges steps (a) and (b) because the input sequencing reads have a length and error profile which is specific to the sequencing platform but the output contigs represent generic sequences with most errors removed.

Genomic methods frequently operate on complete genome sequences, for instance inferring functional models for specific organisms (Price, Reed & Palsson, 2004). Gene regions are identified, their corresponding protein sequences determined and hypothetical pathways constructed. To do similar in metagenomics, contigs are often grouped to form hypothetical draft genomes, called genome bins. The binning process tries to reconstruct the genomes and solves a problem which, at first, appears very similar to that of metagenome assembly. However, contig binning is usually independent of the sequencing platform (it makes no use of sequencing quality) and considers information which assembly programs ignore (e.g. gene annotations). Both steps can be iterated in a feedback cycle (Figure 1.3) to improve the quality of the resulting genomes (Albertsen et al., 2013). Metagenome binning connects step (b) and (c) in Figure 1.2 because it reduces the data to individual genomes. This thesis presents algorithms related to the binning problem which I, in collaboration with my colleagues, developed and published during my doctoral studies.

1.1.3 Community transcriptomics, proteomics and metabolomics

Nucleotide gene sequences can only tell about potential functions of an organism but there may be much more to discover. For instance, we are interested in seeing genes which are actively expressed and to understand how the gene expression is regulated within the community. The proteins, for which the genes code, are the acting agents in any organism, so it is most important to determine the functional role of proteins, how they interact, and which metabolites they

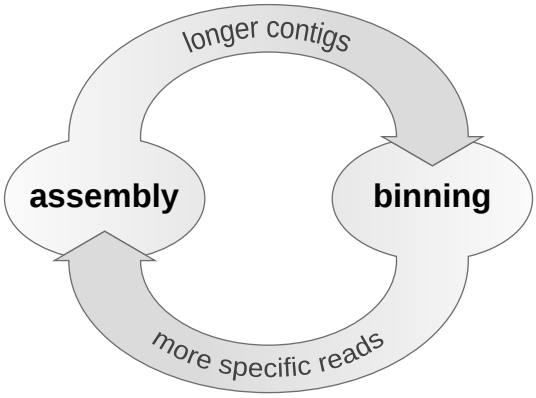


Figure 1.3: Assembly and binning cycle for genome reconstruction in metagenomes. Longer contigs yield better preliminary genome bins and when collecting the reads within a bin, these are more specific to the genome and lead to better assembly.

target and mediate. Corresponding experimental techniques for transcriptome, proteome and metabolome analysis are being adapted and applied to microbial communities (Turnbaugh & Gordon, 2008; Aguiar-Pulido et al., 2016). Such data representing cellular activity are most informative when they can be linked to the corresponding gene sequences and genomes so that their regulation and coupling can be studied in detail. The genomes bins derived by metagenome binning can form the basis to build models which can integrate information from other experiments, for instance measuring the current state of a community in terms of genome activity, micro-evolution or population dynamics.

1.2 Metagenome binning

Functional screenings of metagenomes (Ufarté, Potocki-Veronese & Laville, 2015) aim to identify novel enzymes with biotechnological and medical applications. Though, when studying protein-coding genes and their regulation in more detail, it is often beneficial to look at the corresponding genomes to understand the genomic context. One way to collect cells and to retrieve a full genome sequence is by sampling from the environment followed by cultivation and sequencing, alter-

natively using enrichment cultures (Dong et al., 2017) or single-cell sequencing (Woyke et al., 2009, 2010). However, it can be difficult to extract specific organisms if there are hundreds or thousands of distinct species, subspecies or OTUs in a metagenomic sample (Woyke et al., 2009, 2010; Hess et al., 2011). Furthermore, the cultivation conditions required to produce clone libraries may be unknown, and environmental sequencing of extracted cells with small amounts of DNA is still in its infancy (Mende et al., 2016; Yu et al., 2017). For these reasons, *in-silico* metagenomic methods provide a solid alternative. Metagenome sequence binning is the algorithmic equivalent for reconstructing individual genomes from shotgun metagenome sequence data. Broadly speaking, a genome bin is a set of sequences, usually assembled contigs, which together present the sequenced part of a specific community genome. Capturing these partial genomes allows studying taxa on the level of genes and their associated functions. Genome binning aims to recover full genomes whereas taxonomic binning refers to the assignment of contigs to broader taxonomic groups. For an extensive introduction to metagenome binning, see the review article (Dröge & McHardy, 2012) in appendix C.

1.2.1 Binning methodology

Binning represents a machine learning procedure in which class labels (genomes or taxa) are assigned to data points (contigs) (see Hastie, Tibshirani & Friedman (2001) chapter 1, for a comprehensive introduction to these concepts). Most of the different algorithmic approaches to infer genome bins are either a form of data clustering or classification, including combinations of both approaches. Clustering is a so-called unsupervised method, which does not directly take into account external information like available genome sequences. The strength of clustering is that it can group any data to explore their intrinsic structure, being able to group contigs of genomes which have never been seen before. In contrast, classification algorithms utilize categorized (labeled) data, for instance large genome sequence collections, to assign sequences to genome bins. They are said to operate in a supervised manner. By the use of prior knowledge they can be very efficient but a major drawback is the difficulty to handle novel genomes. Clustering and classification methods give complementary results and it is common to combine

them, for instance classifying genome bins after clustering or initializing clusters using classification labels (Imelfort et al., 2014).

1.2.2 Sequence information for binning

Binning methods can also be categorized by the kind of information they use. Both clustering or classification methods for binning operate on so-called features derived from reads or contigs. These properties inform about genome membership and discriminate contigs of different genomes. Microbial genomes sequences expose characteristic frequencies of short nucleotide motifs (Karlin, Mrazek & Campbell, 1997) which are often used in binning and refered to as the genome or nucleotide composition. The combined relative frequency of guanine and cytosine (GC-content) is a simple way to represent nucleotide composition, and an evolutionary trait of genomes that has long been used to characterize different species. For instance, many Actinobacteria expose a high GC-content. Most methods, however, use short nucleotide motifs consisting of 4 to 7 bases called k-mers (k stands for the number of bases). Alternative formulations may use Hidden Markow Models (HMMs) to describe nucleotide composition (Brady & Salzberg, 2009). The second major feature type for binning is read coverage, the amount of sequencing reads for each assembled contig. Since contigs are constructed by stacking (aligning) overlapping reads, each nucleotide position of a resulting contig must be covered by at least a single read, but typically many more. Following random shotgun sequencing with universal primers, the expected number of reads covering a single position is approximately proportional to the genome copy number in the sequenced sample (Lander & Waterman, 1988), with a constant factor which depends on the total sequencing effort. Thus contig coverage helps to discriminate genomes with distinct sample abundances, but cannot differentiate between equally abundant genomes. It is therefore desirable to generate multiple metagenome samples of a community for which the genome copy numbers vary differently. This way, each genome has a unique set of genome abundances. Recent studies have shown that genome abundances represent a very informative feature type to obtain genome bins for complex metagenomes, if many varying samples are available (Albertsen et al., 2013; Alneberg et al., 2014). Sometimes,

binning programs may also employ assembly information such as associated contigs or scaffolds linked by paired reads (Lu et al., 2016), but such information, if available, is more frequently used to assess the binning quality (Patil et al., 2011) or to refine genome bins (Alneberg et al., 2014).

There is a specific class of homology-based classifiers, and an example of such a method is described in Section 1.3.1. These methods employ a two-step procedure, first identifying potential homologs for a contig, for instance by alignment to reference sequences, and second determining a corresponding evolutionary neighborhood. This neighborhood is usually reported by taxonomy, so that each contig is annotated with a taxonomic path. A grouping of contigs by taxa then provides a form of binning but higher-level taxon bins mix contigs from several genomes, if the sample contains more than a single member of this group. Hence, taxonomic classification using sequence similarity can only provide a partial solution to the binning problem. However, such annotation also informs about the taxonomic sample composition and diversity, similar to a 16S gene analysis, and may furthermore be used as secondary features for clustering, for instance to initialize genome clusters (Imelfort et al., 2014) or to train a classification model with sample data (Gregor et al., 2016; Dong et al., 2017). The probabilistic binning framework presented in Section 1.3.2 makes full use of taxonomic annotations similar to the use of nucleotide composition and contig coverage.

1.2.3 Overview of binning software

Binning programs emerged and evolved together with metagenome sequencing and assembly protocols, so that their focus changed accordingly. Recent programs for complex communities target longer contigs (1 kb or more) but some programs were also designed to bin raw sequencing reads (Vinh et al., 2015; Ulyantsev et al., 2016), for instance by comparison to genome sequence collections or nucleotide composition. Since the latter is unstable for short sequences due to low number of counts (McHardy et al., 2007), these programs are inherently limited to simple communities and community members with related genome sequences to compare to. Most newer binning programs with applications to

complex metagenomes, which are listed in Table 1.1, operate on co-assembled contigs, which are constructed using multiple sequenced samples of a microbial community.

Table 1.1: Contig binning programs with type (taxon or genome bins), methodology and release dates starting from the year 2011 up to the year 2016. This is a non-exhaustive list with rough methodology descriptions. Some programs employ additional sequence information in post-processing procedures which may be omitted here. A recent overview of binning methods can be found in (Sedlar, Kupkova & Provaznik, 2017).

Program	Type	Technique	Sequence information	Published/ updated	License
PhyloPythiaS	taxon	Structured Support Vector Machine (SVM)	5-mers	(2011)/ (2012)	proprietary
MetaWatt 3.x	genome	Heuristic thresholds	4-mers, differential coverage	(2012)/ (2015)	AFL
CONCOCT	genome	Gaussian mixture clustering	4-mers, differential coverage	(2014)/ (2015)	BSD
GroopM	genome	Biclustering	4-mers, differential coverage	(2014)/ (2016)	GPL
MaxBin 2.0	genome	Expectation-Maximization (EM) clustering	4-mers, differential coverage	(2014)/ (2016)	BSD
MetaBAT	genome	Distance-based clustering	4-mers, differential coverage	(2015)/ (2016)	proprietary

Program	Type	Technique	Sequence information	Published/ updated	License
PhyloPythiaS+	taxon	Structured Support Vector Machine (SVM)	5-mers	(2016)/ (2014)	proprietary
MyCC	genome	Stochastic neighbor embedding	4-mers, differential coverage	(2016)/ (2015)	proprietary
COCACOLA	genome	Gaussian mixture clustering	4-mers, differential coverage, genome co-alignment, paired reads	(2016)/ (2016)	GPL

1.2.4 Binning performance considerations

Binning methods are best judged in the context of their use cases. Clearly, an optimal binning would mean to obtain a single bin for each genome in the community. Suboptimal solutions contain either multiple smaller bins for a genome or bins with mixed contigs of different genomes. While the objective is clear, it is impossible to obtain perfect genomes for real metagenome data if there is not enough information to discriminate the contigs, especially shorter ones. All of the increasing number of binning methods typically produce suboptimal bins, and there is no consensus in the metagenomics community on the performance metrics for assessing the bins obtained by different methods and with different benchmark datasets. Initiatives such as the Critical Assessment of Metagenome Interpretation (CAMI) (Sczyrba et al., 2017) work towards establishing a common understanding to judge metagenome binning. Different views on the binning quality are valid as this depends on downstream processing and on the specific

research questions. For instance, the estimation of community structure might only require the construction of precise small-sized bins whereas a hypothetical pathway reconstruction for certain genomes might tolerate excess genes in the corresponding genome bins and discard all of the remaining bins.

Multiple factors, such as the number and abundance of taxa, their phylogenetic structure, availability of reference genome sequences and computing resources have an impact on binning performance. Binning algorithms are sensitive to the type of community, for example, taxonomic sequence classification methods rely on external genome sequences and, as a direct consequence, suffer from the uneven coverage of the tree of life by the reference genomes. Thus, poorly studied environments such as a deep sea vent community are likely too exotic for classifiers which only use public genome sequences. In contrast, communities such as the human gut microbiota are well suited to the classification approach because there are abundant genome data for these microbes. Another reason why binning methods perform differently may be rigid assumptions, for instance standard algorithm parameters which are optimized to give good results in specific scenarios tested and intended by the authors.

The broad range of applications involving many different microbial habitats, custom experimental techniques and heterogeneous sequencing platforms makes it difficult to define a state of the art for binning. Nonetheless, general trends can be observed. Recent works which have presented genome bins derived from complex metagenomes often applied clustering in concatenated and transformed feature spaces (Imelfort et al., 2014; Alneberg et al., 2014; Kang et al., 2015; Lin & Liao, 2016), which integrate several types of features including nucleotide composition and contig coverage for multiple samples. Nevertheless, deriving high-quality draft genomes today still relies on manual analysis and processing of genome bins (Albertsen et al., 2013; Eren et al., 2015).

1.3 Methods in this thesis

1.3.1 Taxonomic annotation of metagenomes (*taxator-tk*)

The method article in Section 3 describes a high-performance tool for taxonomic annotation of metagenomes using phylogenetic principles. The procedure splits the input sequences (contigs) into smaller separate homology regions (segments), to which it applies a newly developed realignment placement algorithm (RPA) for taxonomic classification of these regions. This algorithm calculates pairwise alignment scores to estimate the phylogenetic distances and simultaneously approximates a corresponding tree structure. The alignments are non-exhaustive and are stopped once a good taxon estimate has been determined or if no phylogenetic signal can be found in the input. In a final merging step, the subregion predictions are combined for the full sequence to minimize the error of the predicted taxon. The corresponding computer program *taxatork-tk* is implemented in C++ and utilizes parallel computation.

Introduction

In metagenomics, we study microbial communities from natural environments without obtaining cultures. Using sequencing followed by computational analyses, we can estimate the abundances of taxa, known as taxonomic profiling, and characterize their metabolic potentials by sorting nucleotide sequences into genome bins (binning) and predicting proteins therein. Taxonomic profiling is conceptually different from taxonomic binning because it only requires (partial) genes, which are taxonomically informative, and which can be obtained using amplicon sequencing whereas binning needs to deal with all parts of a genome. Universal marker genes used for profiling are usually classified by phylogenetic placement, which considers a gene reference tree of the corresponding gene as a proxy for the species phylogeny. Random genome regions, as obtained by shotgun sequencing, typically lack such reference trees. Therefore, a taxonomy is used instead and query sequences are compared to reference genomes, which are

annotated with corresponding taxa. Such comparison can be done based on direct sequence matching or based on nucleotide sequence composition, for instance k-mers, which also allows recovering draft genomes from deep-branching lineages. However, sequence matching by alignment is more accurate, in particular for sequences shorter than 1 kb. Corresponding algorithms use alignment scores and threshold parameters to quickly determine an evolutionary neighborhood of a query but lack a well-motivated evolutionary framework. Calculating de-novo gene trees for every query in the metagenome is computationally too demanding for large metagenome samples. The software *taxator-tk* extends the traditional score-based approach by approximating phylogenetic gene trees using a linear number of pairwise alignments and thereby provides more accurate taxonomic assignments without requiring conservation threshold parameters.

Methods

The workflow for the taxonomic assignment of a query sequence consists of three parts (Figure 1.4): (a) a local alignment search for homologs, (b) the core assignment algorithm and (c) a post-processing step to merge subregion annotations. The initial search can be run by different aligners and using different reference sequence collections. Based on the resulting local alignments, each query sequence is split into distinct subregions (segments), omitting parts which have no similarity to any reference. This step reduces the overall number of positions for further alignments and accounts for genome arrangements. Each segment, along with its homologous reference sequences, is processed by the core algorithm to predict a taxon. The final merging step considers all segment predictions of a query sequence and determines the final taxon for assignment.

The core realignment placement algorithm (RPA) (Figure 1.5) assigns a taxon Q to a query segment q using a limited number of pairwise alignments among q and its homologous segments obtained by local alignment to reference sequences. It aims to identify a set of segments which form a monophyletic group or subtree in the corresponding phylogeny. First, the most similar segment s is aligned to the query q and all other segments in the set (pass 1). An outgroup segment o is determined as the first sequence with distance larger than $distance(s, q)$. The

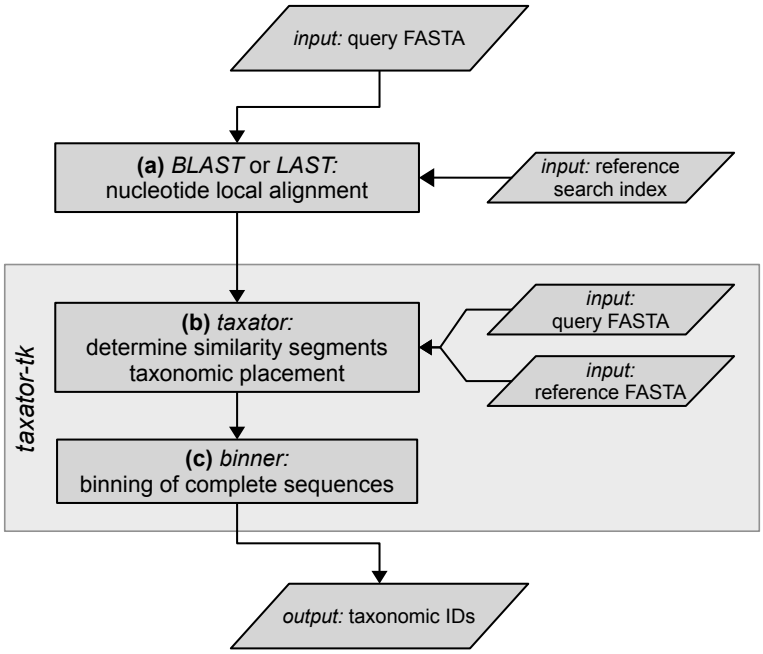


Figure 1.4: Workflow diagram for the taxonomic assignment of a nucleotide query sequence with : (a) Homology search for query sequence in reference collection using local alignment; (b) program taxator splits the query into distinct segments and determines a taxon ID for each; (c) program binner determines a consensus taxon ID for the entire query from the segment predictions.

taxa of all segments with distance smaller or equal to $distance(s, o)$ are added to the neighborhood set M. Then, all segments are aligned to the outgroup segment o (pass 2), again adding taxa with distances smaller than $distance(o, q)$ to M. We assign the least common ancestor (LCA) of all taxa in M to segment q. The segments in M form a subtree among all available segment taxa. Sometimes, if no outgroup can be found or if the taxa in M are very diverse, the algorithm terminates and the predicted taxon is the taxonomy root, meaning unassigned. The RPA requires approximately $2n$ alignments, where n is the number of reference segments.

Results

We evaluated the performance of taxonomic assignment with *taxator-tk* for different datasets: (a) 7176 16S rRNA genes, (b) simulated short sequences of length 100, 500 and 1000 bp, (c) simulated contigs for a synthetic microbial community and two public benchmark datasets and (d) contigs of a microbial community from cow rumen. When possible, we applied cross-validation and evaluated different taxonomic distances between sample and reference taxa. In all cases, the reference data were a diverse collection of full and partial genome sequences with taxonomic annotation. As expected, performance for 16S marker genes was best because it contained a clear phylogenetic signal. In practice, such sequences are best classified using phylogenetic placement because it makes use of reference phylogenies. The second evaluation with nucleotide sequences resembling individual reads, which were sampled from 1729 different species, showed that precision was high even for short sequences, but about 10% lower on average than for 16S data. The recall increased with the length of the sequences. Therefore, it is recommended to assemble reads prior to assignment with *taxator-tk*. For the validation with assembled contigs, we compared our results to other state-of-the-art assignment methods: *CARMA*, *MEGAN*, *Kraken* (all similarity-based) and PhyloPythiaS (composition-based). For the newly simulated community consisting of 49 different species and the two benchmark datasets, *taxator-tk* misassigned substantially fewer contigs at species and genus levels, resulting in a much better precision but a reduced recall. *PhyloPythiaS*, a classifier based on nucleotide composition (k-mers), had the best recall in a specific usage scenario. For the 319 Mb cow rumen dataset, *taxator-tk* was most consistent in assigning 2 kb subsequences to taxonomic bins, which confirmed the previous results on simulated contigs. In summary, *taxator-tk* also predicted the most realistic number of taxa in the samples compared to the other programs. Considering the runtime, it was slower than *Kraken* and *MEGAN*, due to additional computations, but faster than *CARMA* due to the efficient and parallel implementation: we processed $\sim$ 6 Gb per day using 10 CPU cores, including the initial local alignment step. The segmentation procedure of *taxator-tk* accounted for a 30 % decrease of the overall runtime and the program scaled approximately linearly with the input data size.

Discussion

For all compared methods, the bin precision decreased with the bin size. Throughout all validation experiments, we could show that *taxator-tk* was the most precise method in assigning metagenome nucleotide sequences to corresponding taxa among the compared methods (an example shown in Figure 1.6), which also resulted in the most realistic number of taxa. However, it assigned fewer data overall than other methods. This trade-off is a direct implication of the algorithm design, which is tailored towards minimization of errors. Therefore, it can confidently assign a core of sequences, for instance to train a model using nucleotide composition or to estimate taxon abundances. The use of unstructured reference data allows assigning across all domains of life, in contrast to most methods using specific gene families. From a methodological point of view, we presented an alternative phylogenetic inference algorithm which runs in linear time with respect to the number of homologs, and which applies to any nucleotide sequences with no need to select algorithm parameters. Besides taxonomic annotation of metagenomes, it can be applied to any DNA or RNA sequence, for instance to detect contamination in isolate sequencing data.

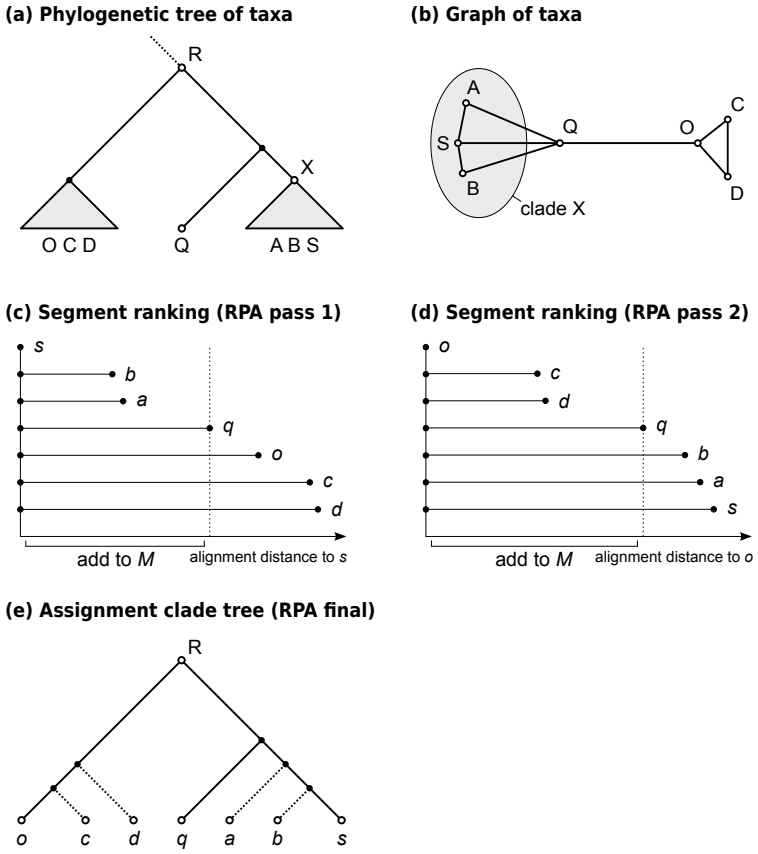


Figure 1.5: Realignment placement algorithm (RPA) for labeling a query segment q with a taxon ID. (a) Underlying taxonomy with query taxon Q and reference taxa A, B, C, D, O and S which is approximated by the query segment alignment. (b) Approximate graph representing pairwise distances between the taxa. The subgraph for clade X is highlighted. (c, d) The two alignment passes which add segment taxa to an empty set M. Segment s is the segment with the smallest local alignment score (distance) to q in the initial similarity search. (c) First, all segments are aligned to segment s. The resulting distances are ordered and the taxa with equal or smaller distances than *distance(s,q)* are added to M. The outgroup segment o is the next most similar segment to s after q, with *distance(o,s) distance(s,q)*. (d) All segments are aligned to o. From the ranked distances, taxa with distances smaller than *distance(o,q)* are also added to M. Thus, M includes all the nearest evolutionary neighbors for the query segment q (the taxa corresponding to segments a, b, c, d, o and s). The taxon ID assigned to q is the lowest common ancestor of taxa in M. (e) Partially resolved segment subtree at node R, which is implied by distances obtained in (c) and (d), where the exact position of some segments (a, b, c and d with dashed branches) is left unresolved by the RPA

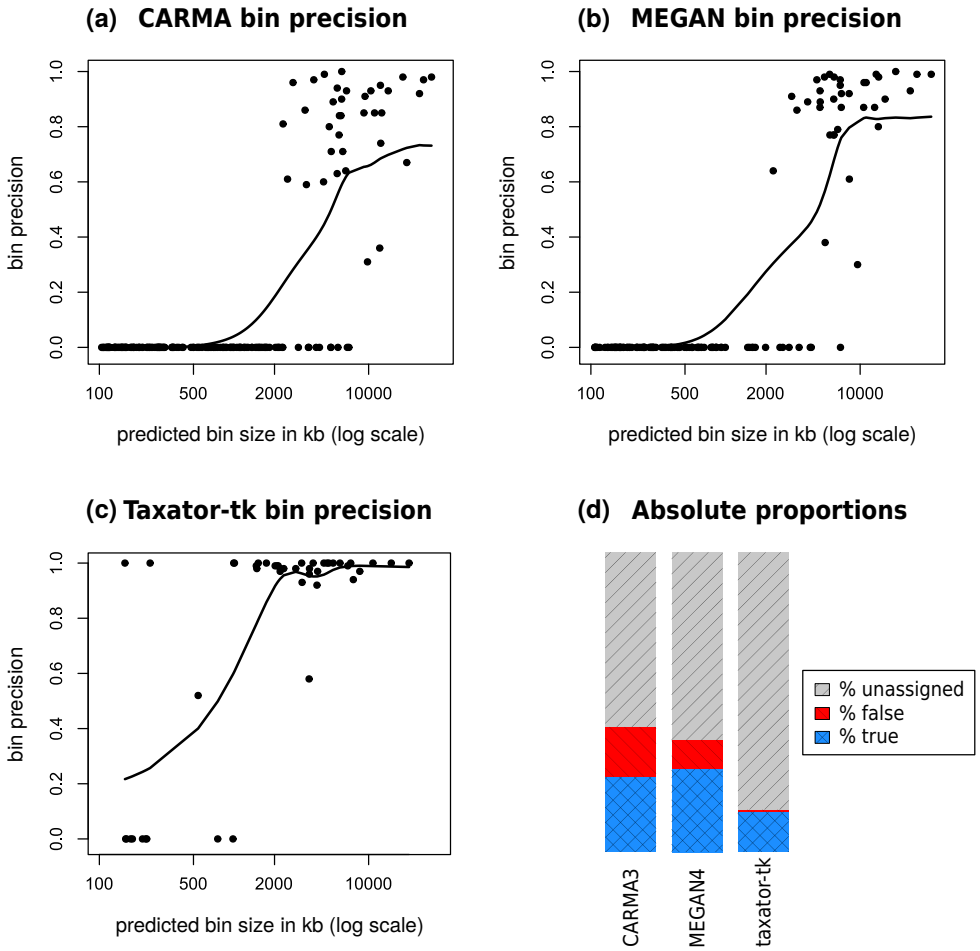


Figure 1.6: Family-level bin precision for the simulated metagenome sample with 49 species (simArt49e). (a-c) Each family bin's assignment precision related to logarithmic bin size for seven cross-validation experiments with simArt49e. The results of the single experiments were added to assess the taxonomic assignment performance across a range of evolutionary distances between the query and the reference sequences, excluding the least abundant bins (1% of total bp). We calculated the precision values for (a) CARMA3, (b) MEGAN4 and (c) taxator-tk, counting assignments to lower-ranking taxa at the family level, and added a smoothed k-nearest-neighbor estimate of the mean precision in R using wapply (width=0.3) followed by smooth.spline (df=10). CARMA3 and MEGAN4 incorrectly identified many small taxonomic bins, substantially more than taxator-tk. (d) gives the amount of correct, false and undetermined family-level assignments for the different classifiers with simArt49e.

1.3.2 A probabilistic model for genome recovery (*MGLEX*)

The corresponding article in Section 4 describes a probabilistic model for use in metagenome binning. Such likelihood models are at the core of many popular algorithms, including sequence classification and clustering. While some models exist as fixed parts of contig clustering programs, we developed a new modular, stand-alone and reusable model using a large set of input features. This model is based on parameterized submodels for which maximum likelihood (ML) parameter estimates can be inferred. Besides classification and clustering, we demonstrate alternative applications such as sample size reduction and visualization. The method is available as an open-source Python library and command line program called MGLEX.

Introduction

Shotgun sequencing of a microbial community bypasses the need to obtain pure cultures and thus enables novel insights into ecosystems, in particular for those genomes that are inaccessible by cultivation. Since current metagenome assemblies are oftentimes highly fragmented, a process called binning sorts assembled sequences (contigs) according to the underlying genomes. Various programs were written to bin metagenomes, using different methodologies and sequence features. These comprise classification and clustering by consideration of k-mer distributions (nucleotide composition), sequence similarity (homology) and assembly read coverage (genome copy number). Coverage information can be very powerful for separating genomes, if multiple samples with varying genome copies are sequenced and co-assembled. However, with a limited number of samples, it remains difficult to reconstuct high-quality bins down to the strain level. Here, we propose a model for metagenome binning, using probabilities to represent natural uncertainty. The model aggregates explicit submodels for read coverage, nucleotide composition and contig similarity to reference sequences (via taxonomic annotation). This design incorporates knowledge about the feature generation process in each submodel, which leads to a robust fit when few data are available. In

contrast, other methods frequently apply a data-driven transform before clustering with a single, e.g. Gaussian, model. Our implementation *MGLEX* does not represent an automatic binning solution but a flexible framework for genome recovery.

Methods

A classification model is trained to distinguish data of different classes. In probabilistic modeling, training means to determine the model parameters (θ) from example data for a set of different classes. Here, classes correspond to different genomes which make part of a metagenome and the data to be classified are contigs. Hence, we need to provide training sequences for each genome before we can classify unknown contigs.

Let $1 \leq i \leq D$ be an index referring to D contigs resulting from a shotgun metagenomic experiment. For the i^{th} contig, we define a joint likelihood for genome bin g (Equation 1.1), which is a weighted product over M independent submodels likelihoods for the different feature types. For the k^{th} submodel, $\boldsymbol{\Theta_k}$ is the corresponding parameter vector, $\boldsymbol{F_{i,k}}$ the feature vector of the i^{th} contig and α_k defines the contribution of the respective submodel or feature type. β is a free scaling parameter to adjust the smoothness of the aggregate likelihood distribution over the genome bins (bin posterior).

$$\mathcal{L}(\boldsymbol{\Theta_g} \mid \mathbf{F_i}) = \left(\prod_{k=1}^{M} \mathcal{L}(\boldsymbol{\Theta_{gk}} \mid \boldsymbol{F_{ik}})^{\alpha_k} \right)^{\beta} \tag{1.1}$$

The model assumes statistical independence of the submodel features. All model parameters are determined from training data, $\boldsymbol{\Theta}$ using submodel ML estimation, $\boldsymbol{\alpha}$ using the inverse standard deviations of the class log-likelihood distributions (Figure 1.7) and β by mean squared error (MSE) minimization (Figure 1.8).

We integrate different submodels $\mathcal{L}(\boldsymbol{\Theta_k} \mid \boldsymbol{F_{i,k}})$ according to distinct input feature types:

- a Poisson model for absolute read coverage considering multiple samples

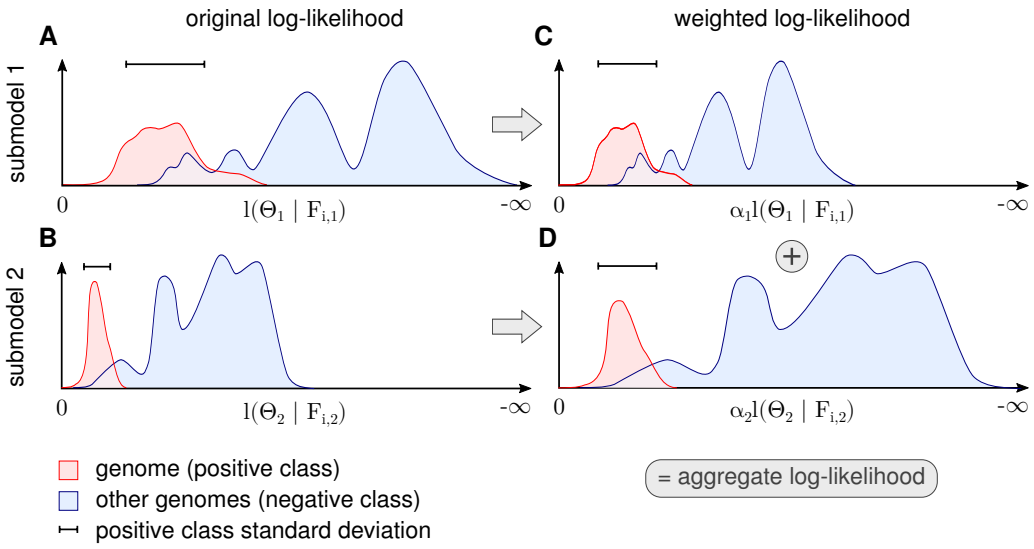


Figure 1.7: Procedure for determination of α_k for each submodel. The figure shows a schematic for a single genome and two submodels. The genome's contig log-likelihood distribution is scaled to a standard deviation of one before adding the term in the aggregate model.

- a Binomial model for relative read coverage considering multiple samples
- a frequency model for k-mers
- a set of layered frequency models for taxonomic annotation of contigs

The layered frequency model is an adjustment of the standard frequency model for hierarchical labels because the taxonomy represents a tree-like structure (Figure 1.9). The listed submodels are kept simple and make feature independence assumptions to simplify calculations.

We simulated a metagenome (400 genomes with strain heterogeneity) and created short contigs (1 kb) to validate and demonstrate the aggregate model. Differential abundances were produced by simulating Illumina reads (150 bp) for a primary lognormal and three secondary abundance distributions and by mapping the resulting reads to the contigs, introducing typical biases but omitting the actual read assembly. For each genome, we obtained 300 kb of contig data and calculated the read coverage, 5-mer frequencies and taxonomic annotations as features for the model.

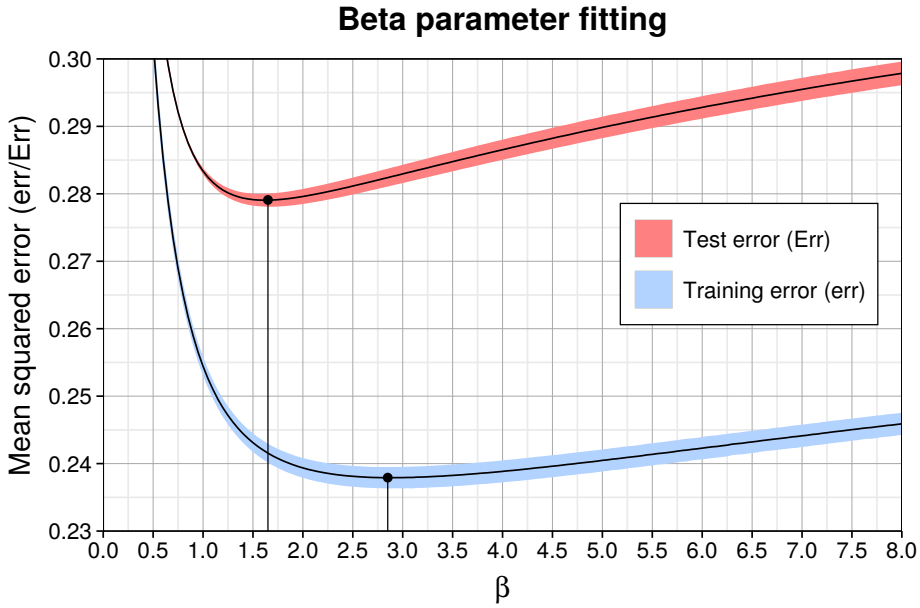


Figure 1.8: Model training (err) and test error (Err) as a function of β for the complete aggregate model including all submodels and feature types. The solid curve shows the average and the colored shading the standard deviation of the three partitions in cross-validation. The corresponding optimal values for β are marked by black dots and vertical lines. The minimum average training error is 0.238 ($\beta = 2.85$) and test error is 0.279 at $\beta = 1.65$.

Results

Using the simulated metagenome, we applied three-fold cross-validation and checked how well the model classified contigs to the most likely genome (ML) with different combinations of input features. Genome abundance turned out to be the weakest single feature type while taxonomic annotation from local alignment to reference genome sequences was the strongest. However, the aggregation of submodels according to Equation 1.1 yielded better performance in all cases. In summary, about 68% of contig pairs, which were not used for model training, were classified to the same genome using the full set of available submodels. Considering species-level bins, this value increased to 79%, which showed that the model had difficulties to distinguish strains of the same species using the differential abundance values stemming from only four samples in our simulation.

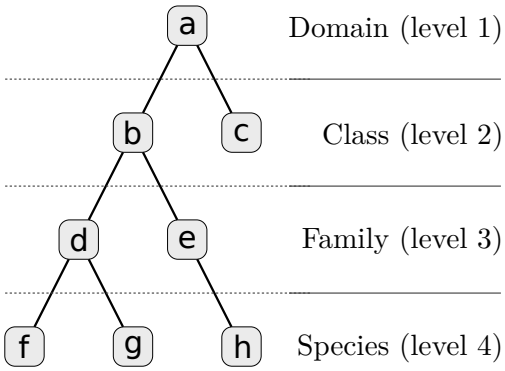


Figure 1.9: Taxonomy stucture simplified to four levels and eight nodes. A full taxonomy may consist of thousands of nodes. Each taxonomy level uses a frequency model which is assumed independent of the remaining levels.

The error decreased further when applying soft (not ML) classification, fitting the parameter β (Figure 1.8), because each contig could then belong to several genomes with varying class posterior probability. When the model was used to refine the genome bins from two popular automatic binning programs, the quality (adjusted Rand index) improved for both of these programs.

We demonstrated alternative model applications besides classification. Using the likelihood distributions in the training data, we calculated p-values, which indicates how extreme a particular contig likelihood is with respect to the training data. With sufficient training data (100 kb in our example), we used the p-value to enrich a metagenome sample *in-silico* for a specific genome, so that irrelevant contigs were removed and the overall sample size was reduced. On average, a critical p-value of 2.5% led to a sample size reduction of 95%. Such a filtering step may be useful for a more focused analysis or to apply a method with otherwise prohibitive runtime. As a second application example, we derived a probabilistic measure to quantify the similarity between any two genomes or genome bins. The quantity is based on a relative mixture likelihood and may be used to cluster bins hierarchically and to analyze the similarity structure of genome bins (Figure 1.10). In particular, the method indicates whether the resolution of individual bins is justified with respect to the model and contig data.

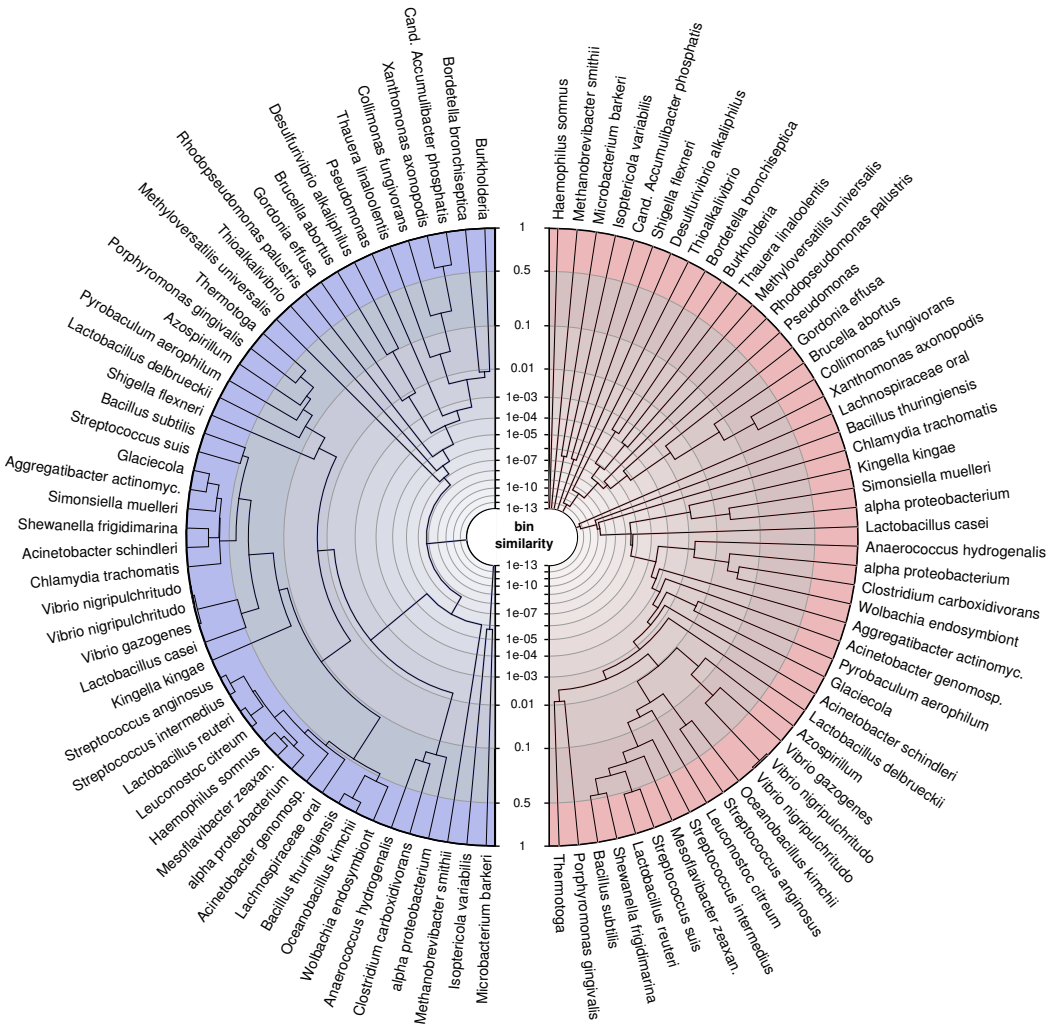


Figure 1.10: Average linkage clustering of a random subset of 50 out of 400 genomes using probabilistic distances to analyze bin resolution. This example compares the left (blue) tree, which was constructed only with nucleotide composition and taxonomic annotations, with the right (red) tree, which uses all available features. The tip labels were shortened to fit into the figure. The similarity axis is scaled logarithmically to focus on values close to one. Bins which are more than 50% similar branch in the outermost ring whereas highly dissimilar bins branch close to the center.

Discussion

We described an aggregate likelihood model with applications in metagenome binning, for instance classification, genome enrichment and visualization. It builds on specific submodels, each responsible for different feature types. The modular design helps to improve the model and to compute and interpret the results. In comparison to previous methods, we added two new submodels. The first is a binomial model for relative differential read coverage over multiple samples to account for systematic read mapping biases and the second is a layered frequency model for taxonomic annotation, which allows considering external knowledge from reference sequences for sequence binning. We also proposed a new weighting scheme to combine the information of several submodels. The reference implementation called *MGLEX* in its current state lacks support for parallel computations, which will be added later. As the runtime for all submodel ML parameter estimations and sequence classification is linear in the number of contigs, embedding it into clustering algorithms such as the Expectation Maximization (EM) or Markov Chain Monte Carlo (MCMC) algorithms is also feasible. We hope to continue developing the open-source package *MGLEX* as a flexible framework for metagenome analysis and binning, to be integrated into programs and workflows.

1.3.3 Further works

The published methods in this thesis were validated using data simulations and sampling from reference genome sequences. Nevertheless, their use must be shown when applied to a variety of real metagenomes. The program *taxator-tk* was subsequently applied in two metagenome studies in completely different settings. For the publication by Bulgarelli et al. (2015), taxonomic profiles were generated for metagenome contigs to study complex microbial communities associated with plant roots (rhizosphere). The taxonomic profiles where shown to be consistent with profiles based on independent 16S amplicon sequencing for the same communities. Furthermore, *taxator-tk* was able to discover members of clades, for instance Archaea and Cyanobacteria, which the 16S primers seemed to have missed in the amplification step. Such biases for the primers used to amplify regions of the 16S gene were also independently confirmed (Eloe-Fadrosh et al., 2016). The taxonomic profiles based on shotgun metagenome data were also not influenced by 16S copy number variations in the corresponding genomes, unlike the amplicon profiles. In a second study of a benzene-degrading enrichment community (Dong et al., 2017), *taxator-tk* was applied to derive bin-specific sequence data to train a full model for the composition-based classifier PhyloPythiaS (Patil et al., 2011), so that the genomes of four species could be recovered, two of them with over 97% completeness. Thereby, we used the same logic to define the model and to seed the genome bins with training data as in the program PhyloPythiaS+ (Gregor et al., 2016), but we replaced the homology search based on marker genes with *taxator-tk*, which offered a better coverage of genomic reference for this task. The completeness and potential contamination levels of the derived genomes were checked independently, based on single-copy marker genes and the near-complete genomes were then used to study benzene degradation pathways by linkage to metabolomic experiments and to propose a benzene oxidation pathway with direct sulfate reduction.

Working with metagenomic data and comparing the results of different binning programs, for instance in (Dröge, Gregor & McHardy, 2014), we observed that the current metagenome analysis toolbox features many programs for similar problems giving different results. One possible explanation is that metagenomics

is an interdisciplinary field with contributions from biotechnology, ecology and medicine, each with a different focus on ecosystems and data (see Figure 1.1). As a result, metagenomics lacks a systematic and cross-discipline view on software for data processing and analysis. To improve the situation, the Critical Assessment of Metagenomic Interpretation (CAMI) challenge (http://cami-challenge.org) compared computer programs for metagenome analysis, such as metagenome assembly, taxonomic profiling and genome binning. As part of my thesis work, I contributed both by taking part in the conception and implementation of the binning evaluation framework as well as by submitting *taxator-tk* for comparison (Sczyrba et al., 2017).

1.4 Summary of results

The taxonomic annotation program *taxator-tk* was shown to obtain very high precision on a number of synthetic and real metagenomes by applying phylogenetic principles. It requires similar reference genome sequences to calculate a phylogenetic neighborhood for annotation. In its initial stage, the provided example workflow has the option to use two different search programs, but the local aligner is exchangeable in order to adapt to sequence data which stem from different experimental procedures. Within the core algorithm (RPA), which is based on pairwise alignment of partial sequences (segments), *taxator-tk* neither relies on exact scores from the local aligner nor on a complete set of retrieved homologs and there are no related parameters to be set. The RPA was recently adapted to amino acid sequences, so that direct protein alignment can be used for the similarity search without the need to back-translate similarity matches to the nucleotide level. For example, some alternative local alignment programs for identification of similar sequences have been presented lately, which claim to improve the search time by fast protein alignment with a reduced alphabet (Zhao, Tang & Ye, 2011; Huson & Xie, 2014; Hauswedell, Singer & Reinert, 2014; Buchfink, Xie & Huson, 2014). Another advantage of *taxator-tk* is its independence of curated reference data, in contrast to the standard procedures in phylogenetic analysis using precomputed HMMs or gene families. This comes at the cost of an increased computation time for de-novo phylogenetic structure detection but enables *taxator-tk* to be applied in less frequent, non-standard situations, for example to analyze communities with eukaryotic content, like algae or fungi.

The probabilistic model for metagenome binning and its software implementation *MGLEX* make use of many available sequence features to classify contigs to genomes or genomes bins, and we exemplified alternative applications such as genome enrichment and bin analysis. We could also show on benchmark data that the application of the model improved on the results from recent automatic binning procedures, which confirmed our initial incentive to make better use of the available data to recover individual genomes. The model itself is very generic so that it can, in theory, also be applied to non-metagenomic datasets. We designed

MGLEX as a subroutine for use in other software to maximize the benefits resulting from future improvements. It should be integrated into more user-friendly applications for genome recovery.

In the conception stage of both methods, we considered that the algorithms scale with large datasets and that they solve well-defined problems. As a commitment to open science, we released the program source codes to the public and used simple and well specified data formats wherever possible. The software ought to be flexible enough to keep pace with the future progress both in experimental protocols and sequencing technologies.

The two methods in this thesis extend available software for analyzing metagenomes. From a methodological perspective, these methods cover several algorithmic fields including sequence alignment, phylogenetics and probabilistic modeling. Each of the articles published in the course of the thesis follows the track to improve on the understanding of metagenomic data. While the binning review (Dröge & McHardy, 2012), see Appendix C, gave an extensive introduction to the different metagenome binning and analysis approaches, the first method article in Section 3, (Dröge, Gregor & McHardy, 2014) presented the program *taxator-tk*, which enables precise taxonomic annotation of entire metagenomes by fast calculation of phylogenetic neighborhoods. The second method article in Section 4, (Dröge, Schönhuth & McHardy, 2017) proposed a statistical classification framework to recover genomes from shotgun-sequenced metagenomes. Applied studies used *taxator-tk* and demonstrated its utility to inform about taxonomic composition (Bulgarelli et al., 2015) and to reconstruct near-complete genomes for a simple community (Dong et al., 2017). Finally, a comprehensive comparison of metagenome processing software was conducted as a challenge (Sczyrba et al., 2017) to improve on the overall interpretation of metagenome studies.

1.5 Conclusions and outlook

Metagenomics as a discipline has matured in the course of this thesis, with regard to nucleotide sequencing, metagenome assembly and computational analysis. For instance, paired read insert libraries and long-read technologies allow assembling larger fractions of metagenomes. The subsequent assembly of metagenomes, which differs from isolate genome assembly, is considered an important task and led to the development of dedicated algorithms. The interest in medical applications has been continuously increasing in metagenomics so that analyzing the human gut microbiome and its impact on human health has become a common procedure. Several large-scale projects reflect an increased interest in different areas, for instance the Human Microbiome Project (http://hmpdacc.org), MetaHIT (http://www.metahit.eu) or the Earth Microbiome Project (http://www.earthmicrobiome.org), which all seek to collect data using standardized protocols and analysis methods.

For the analysis of metagenomics data, the impact of algorithms on the overall conclusions may not be underestimated, as most of the data is directly or indirectly produced by computer programs. Each specific procedure may be sensitive to the applied software pipeline and the results may, for example, differ in the number and abundance of OTUs, the quality of assembled genome sequences and the robustness to particular experimental details such as sequencing errors. For the multitude of methods which have been developed over the past years, including the methods presented in this thesis, it is still to determine under which conditions they should, or should not, be applied and how they compare to other methods which claim to solve similar problems. Therefore, in addition to developing new methods, rigorous testing is required to provide a more complete picture of the metagenomic software landscape for the scientific community. In the course of the CAMI initiative, we noted that software accessibility represents an important factor, among others like code quality and program (re-)usability, in order to enable systematic testing and reproducibility of results. Future compliance of academic software with these criteria will therefore be an important factor for a better assessment of programs and their results.

This thesis and the methods presented here contribute to the field by providing some base-level metagenome analysis tools. They implement new theoretical approaches and are accessible for evaluation and application as open-source. Both *taxator-tk* and *MGLEX* are also suitable to assess the quality of metagenome assemblies and binning from various environments. In the near future, the aim will be to recover high-quality genomes in an automatic way. This target may soon be reached, not only by algorithmic improvements in metagenomics but also by combination with new experimental techniques and further progress in sequencing technology. For instance, single-molecule sequencing can eliminate problems in metagenome processing, which are associated with the short read length. Single cell sequencing is another complementary technique which allows assembling genomes from very limited numbers of microbial cells (Lasken & McLean, 2014; Gawad, Koh & Quake, 2016), which need to be isolated but not grown in medium. The combination of data from single-cell and metagenome sequencing can improve genome reconstructions (Mende et al., 2016; Bremges et al., 2016).

Chapter 2

Report of Publications

This chapter lists the publications to which I contributed in the course of this thesis. My attribution to the individual works is reported as percentage estimates (5% ranges) and a short description of the contributions.

2.1 Central publications

These are the publications of the developed methods on which this cumulative thesis is based.

Title	*Taxator-tk*: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods
Journal	Bioinformatics
Published	10 November 2014
Authors	Johannes Dröge, Ivan Gregor, Alice C. McHardy
DOI	10.1093/bioinformatics/btu745
Contributions	Designed method, developed software, designed experiments, conducted experiments, wrote manuscript
Attribution	71% to 75%

Title	A Probabilistic Model to Recover Genomes in Shotgun Metagenomics
Journal	PeerJ
Published	17 December 2016 (preprint), 22 May 2017
Authors	Johannes Dröge, Alexander Schönhuth, Alice C. McHardy
DOI	10.7717/peerj-cs.117
Contributions	Designed method, developed software, designed experiments, conducted experiments, wrote manuscript
Attribution	86% to 90%

2.2 Related publications

These publications relate to metagenome binning or for which the developed software was applied to analyze metagenomes.

Title	Taxonomic binning of metagenome samples generated by next-generation sequencing technologies
Journal	Briefings in Bioinformatics
Published	31 July 2012
Authors	Johannes Dröge, Alice C. McHardy
DOI	10.1093/bib/bbs031
Contributions	Collected information, wrote manuscript
Attribution	50%

Title	Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley
Journal	Cell Host & Microbe
Published	11 March 2015
Authors	Davide Bulgarelli, Ruben Garrido-Oter, Philipp C. Münch, Aaron Weimann, Johannes Dröge, Yao Pan, Alice C. McHardy, Paul Schulze-Lefert

DOI	10.1016/j.chom.2015.01.011
Contributions	Assembled metagenome, analyzed metagenome, added to manuscript, reviewed manuscript
Attribution	1% to 5%

Title	PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes
Journal	PeerJ
Published	8 February 2016
Authors	Ivan Gregor, Johannes Dröge, Melanie Schirmer, Christopher Quince, Alice C. McHardy
DOI	10.7717/peerj.1603
Contributions	Contributed to method design, reviewed manuscript
Attribution	1% to 5%

Title	Reconstructing metabolic pathways of a member of the genus Pelotomaculum suggesting its potential to oxidize benzene to carbon dioxide with direct reduction of sulfate.
Journal	FEMS Microbiology Ecology
Published	23 December 2016
Authors	Xiyang Dong, Johannes Dröge, Christine von Toerne, Sviatlana Marozava, Alice C. McHardy, Rainer U. Meckenstock
DOI	10.1093/femsec/fiw254
Contributions	Annotated metagenome, created phylogenetic trees of genes, added to manuscript, reviewed manuscript
Attribution	16% to 20%

Title	Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software

Journal	Nature Methods
Published	12 June 2017 (preprint), 2 October 2017
Authors	Alexander Sczyrba, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, Stephan Majda, Jessika Fiedler, Eik Dahms, Andreas Bremges, Adrian Fritz, Ruben Garrido-Oter, Tue Sparholt Jørgensen, Nicole Shapiro, Philip D Blood, Alexey Gurevich, Yang Bai, Dmitrij Turaev, Matthew Z DeMaere, Rayan Chikhi, Niranjan Nagarajan, Christopher Quince, Fernando Meyer, Monika Balvočiūtė, Lars Hestbjerg Hansen, Søren J Sørensen, Burton K H Chia, Bertrand Denis, Jeff L Froula, Zhong Wang, Robert Egan, Dongwan Don Kang, Jeffrey J Cook, Charles Deltel, Michael Beckstette, Claire Lemaitre, Pierre Peterlongo, Guillaume Rizk, Dominique Lavenier, Yu-Wei Wu, Steven W Singer, Chirag Jain, Marc Strous, Heiner Klingenberg, Peter Meinicke, Michael D Barton, Thomas Lingner, Hsin-Hung Lin, Yu-Chieh Liao, Genivaldo Gueiros Z Silva, Daniel A Cuevas, Robert A Edwards, Surya Saha, Vitor C Piro, Bernhard Y Renard, Mihai Pop, Hans-Peter Klenk, Markus Göker, Nikos C Kyrpides, Tanja Woyke, Julia A Vorholt, Paul Schulze-Lefert, Edward M Rubin, Aaron E Darling, Thomas Rattei, Alice C McHardy
DOI	10.1038/nmeth.4458
Contributions	Contributed to data simulation design, curated reference data, contributed to evaluation methods, contributed to framework design, added to manuscript, reviewed manuscript
Attribution	1% to 5%

2.3 Other publications

Title	Bioboxes: standardised containers for interchangeable bioinformatics software.
Journal	GigaScience
Published	15 October 2015
Authors	Peter Belmann, Johannes Dröge, Andreas Bremges, Alice C. McHardy, Alexander Sczyrba, Michael D. Barton
DOI	10.1186/s13742-015-0087-0
Contributions	Contributed to design, created template containers, reviewed manuscript
Attribution	6% to 10%

Chapter 3

Taxator-tk: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods

J. Dröge[1,2], **I. Gregor**[1,2] and **A. C. McHardy**[1,3*]

[1]Department for Algorithmic Bioinformatics, Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany

[2]Max-Planck Research Group for Computational Genomics and Epidemiology, Max-Planck Institute for Informatics, University Campus E1 4, 66123 Saarbrücken, Germany

[3]Computational Biology of Infection Research, Helmholtz Center for Infection Research, Inhoffenstraße 7, 38124 Braunschweig, Germany

---

This is an author-produced version of an article accepted for publication in *Bioinformatics* following peer review. This version has been

adapted to the thesis layout. The original open-access article is accessible by DOI 10.1093/bioinformatics/btu745.

3.1 Abstract

3.1.1 Motivation

Metagenomics characterizes microbial communities by random shotgun sequencing of DNA isolated directly from an environment of interest. An essential step in computational metagenome analysis is taxonomic sequence assignment, which allows identifying the sequenced community members and reconstructing taxonomic bins with sequence data for the individual taxa. For the massive datasets generated by next-generation sequencing technologies, this cannot be performed with de-novo phylogenetic inference methods. We describe an algorithm and the accompanying software, taxator-tk, which performs taxonomic sequence assignment by fast approximate determination of evolutionary neighbors from sequence similarities.

3.1.2 Results

Taxator-tk was precise in its taxonomic assignment across all ranks and taxa for a range of evolutionary distances and for short as well as for long sequences. In addition to the taxonomic binning of metagenomes, it is well suited for profiling microbial communities from metagenome samples because it identifies bacterial, archaeal and eukaryotic community members without being affected by varying primer binding strengths, as in marker gene amplification, or copy number variations of marker genes across different taxa. Taxator-tk has an efficient, parallelized implementation that allows the assignment of 6 Gbp of sequence data per day on a standard multiprocessor system with ten CPU cores and microbial RefSeq as the genomic reference data.

3.1.3 Availability

Taxator-tk source and binary program files are publicly available at http://algbio.cs.uni-duesseldorf.de/software/.

3.2 Introduction

Metagenomics allows us to study microbial communities from natural environments without the need to obtain pure cultures of the individual member species (Hugenholtz, 2002; Riesenfeld, Schloss & Handelsman, 2004). The shotgun sequencing of microbial community DNA with current techniques generates reads that range from less than 100 to several thousand nucleotides (Dröge & McHardy, 2012; Klumpp, Fouts & Sozhamannan, 2012). By computational analyses of metagenome sequence samples, we can estimate the abundances of different taxa for the sampled communities, known as taxonomic profiling, characterize their functional and metabolic potential based on the predicted proteins and resolve the contributions of individual taxa to the latter by reconstructing "bins" of unassembled or assembled sequences that originate from the same taxon.

A taxonomic profile of a microbial community can be inferred by either targeted amplification and sequencing of taxonomic marker genes or from metagenome shotgun datasets (Lindner & Renard, 2013; Sunagawa et al., 2013; Silva et al., 2014). Most metagenome profiling methods classify reads based on predefined taxon-specific (Segata et al., 2012) or "universal" marker genes (Darling et al., 2014), or directly estimate a taxonomic profile for the underlying microbial community from their k-mer composition (Koslicki, Foucart & Rosen, 2013). Frequently used phylogenetic placement programs within such frameworks are pplacer (Matsen, Kodner & Armbrust, 2010) or EPA/RAxML (Berger, Krompass & Stamatakis, 2011), which both operate in a probabilistic framework to place a query gene sequence in a pre-computed reference phylogeny of a particular gene family. If this gene tree is an approximate representation of the respective species tree – or reference taxonomy – this can be used to assign a taxon identifier (ID) to the query sequence (Matsen, Kodner & Armbrust, 2010; Stark et

al., 2010). Taxon abundances are then derived from the individual read counts or gene frequencies within each taxonomic group.

Binning methods place the sequences of a shotgun metagenome sample into bins representing the different taxa of the sampled microbial community. If a bin represents a low-ranking taxon, such as species, then the set of reads or contigs of an individual taxonomic bin serves as a draft-genome reconstruction for a community member (Pope et al., 2011). Binning methods are either based on clustering or classification. Clustering methods group sequences into bins without consideration of external reference sequences or taxonomic information. Instead, bins are inferred based on similarities in GC content, oligomer frequencies, the abundance of genes or contig coverage within one or multiple samples (Baran & Halperin, 2012; Carr, Shen-Orr & Borenstein, 2013; Albertsen et al., 2013; Alneberg et al., 2014), or by using a combination of these (Iverson et al., 2012). This allows draft genome recovery from deep lineages for sequences of sufficient length. Taxonomic binning, like profiling, uses the resemblance of a sequence to known taxa in either global sequence composition or local sequence similarity to assign a taxon ID. For the human gut microbiome, extensive genome sequencing of isolate cultures allowed species-level taxonomic binning for a substantial portion (approx. 40%) of a metagenome sample (Schloissnig et al., 2013) by mapping the reads to isolate genome sequences, which exist for many abundant species [Sunagawa et al. (2013). However, this procedure is not suitable for environments in which most species are from deep-branching lineages without available reference genome sequences. Taxonomic binning of these requires more sophisticated similarity-based or composition-based taxonomic assignment methods (McHardy et al., 2007; Brady & Salzberg, 2011; Huson et al., 2011). Taxonomic binning by sequence composition also allows draft genome recovery from deep-branching lineages, based on limited amounts of sequences for the individual taxa (McHardy et al., 2007). Composition-based programs achieve linear classification times regarding metagenome sample size, while similarity-based binning methods require considerably more computational resources for sequence similarity searches in large reference sequence collections. Programs with a focus on processing large amounts of raw sequencing reads, such as Kraken (Wood & Salzberg, 2014), implement the fastest search strategies. Similarity-based programs are more accurate

for the assignment of sequences shorter than 1 kbp (Patil et al., 2011).

A common aim in taxonomic profiling and taxonomic binning is the identification of known taxa from a sample. A taxonomic profiler estimates a taxonomic abundance profile for the entire sample, which can be inferred by analyzing a smaller number of marker genes, though one needs to account for variations in gene copy numbers for taxon-specific markers (Lindner & Renard, 2013). Taxonomic binning assigns taxon IDs to a large portion of the sample sequences for subsequent functional and metabolic analyses of individual taxon bins. In addition, one can generate a taxonomic profile by quantifying the assigned reads, based on read counts or coverage for each individual bin.

From a methodological standpoint, the differences between the phylogenetic-placement-based methods for profiling and alignment-score-based methods for taxonomic binning and profiling, such as MEGAN (Huson et al., 2011), CARMA3 (Gerlach & Stoye, 2011) or SOrT-ITEMS (Monzoorul Haque et al., 2009) are that the latter lack a well-motivated evolutionary framework. However, they have the advantages of being computationally lightweight and applicable to arbitrary genes, which is a necessity for taxonomic binning. Phylogenetic-placement-based methods cannot currently be used for binning, because the de-novo inference of trees for gene families on a metagenome-wide scale is computationally too demanding, particularly for next-generation sequencing (NGS) data.

Our taxator toolkit (taxator-tk) is a software package for the taxonomic sequence assignment in shotgun metagenomics with applications to both profiling and binning. Conceptually, it lies between sequence similarity based programs which use local sequence alignment scores and those using trees. Taxator-tk extends the alignment score based approach by approximating phylogenetic gene trees and thereby provides more accurate taxonomic assignments, without assuming universal, rank or clade-specific gene conservation levels as parameters. We improve in terms of applicability to large data sets compared to phylogenetic methods by assigning genomic sequences without the computationally demanding steps of de-novo multiple sequence alignment (MSA) and tree inference. Taxator-tk determines a subset of homologs, which represent the approximate evolutionary neighbors for a query sequence, by a linear number of pairwise sequence comparisons

with regard to the number of considered homologs and then assigns a taxon ID using a reference taxonomy based on the taxonomic IDs of these neighbors. We have furthermore reduced the run-time by limiting the analysis to distinct homology-supported regions of the query sequence, which we termed query segmentation. Our open-source (GPLv3) software can be applied to arbitrary nucleotide sequences, such as reads, contigs, scaffolds and complete genomes sequences. It can be downloaded from http://algbio.cs.uni-duesseldorf.de/software/.

3.3 Methods

3.3.1 Taxator-tk's workflow for taxonomic assignment

The workflow for the taxonomic assignment of a nucleotide query sequence comprises three stages (Figure 3.1 a–c). The first stage uses a local sequence aligner to identify similar regions from a reference sequence collection, such as microbial RefSeq (mRefSeq) (Sayers et al., 2009). The implemented workflows currently use BLAST+ (Camacho et al., 2009) version 2.2.28+ using any of the blastn, megablast or tblastx algorithms and nucleotide LAST (Frith, Hamada & Horton, 2010) version 320. Other aligners can be used via conversion to a TAB-separated format, if found to be more appropriate. We discuss our choice of the aligner in the Supplementary Methods ("IX. Sequence Homology Search via Local Alignment"). At the beginning of the taxator algorithm in stage two, overlapping regions on the query, each defined by local alignment to a nucleotide reference sequence, are merged into larger subsequences called segments (Supplementary Fig. 1). These query segments are flanked by regions without similarity to any reference data (Supplementary Fig. 2) and are not considered further. This step reduces the overall number of positions in the following alignment computations and improves the taxonomic assignment of queries that have undergone genome rearrangements, resulting in a different order of these segments. The reference sequence regions corresponding to the local alignments are extended at both sides by the missing number of nucleotides to match to the corresponding query segment with respect to its length and we refer to these as reference segments. Each

independent set of homologous segments is the input to the core algorithm in the program taxator in stage two (Figure 3.1 b), which calculates independent taxon IDs for every corresponding query segment.

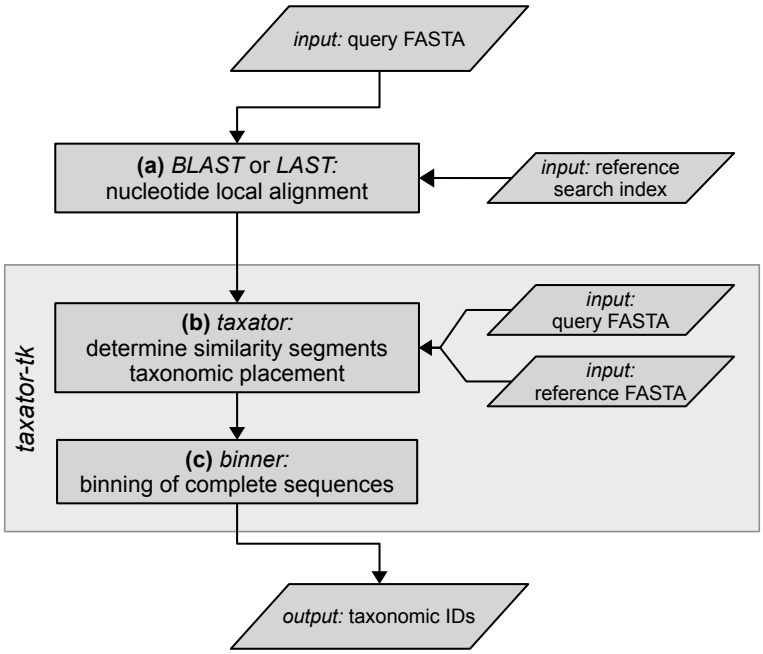


Figure 3.1: Workflow diagram for the taxonomic assignment of a nucleotide query sequence with taxator-tk. Taxonomic assignment with taxator-tk includes three steps. (a) Homology search for query sequence in reference collection using a nucleotide local alignment program. (b) Program taxator splits the query into distinct segments and determines a taxon ID for each using the corresponding homologs. (c) Program binner determines a taxon ID for the entire query based on the taxonomic assignments of the individual segments.

In the third stage (Figure 3.1 c), multiple segments belonging to the same query are considered and their IDs are combined in the program binner, to derive a consensus taxon ID. The corresponding algorithm weights the individual segment assignments by the number of identical bases to the closest reference sequence and assigns to the entire query the taxon ID supported by the majority (default = 70% identical bases) of weighted assignments with a minimum number of identical bases (default = 50 bp) (Supplementary Methods, "II. Consensus Binning Algorithm"). Binner has the optional parameters minimum sequence identity

and minimum sample abundance, but these were not applied in our analysis. If the taxonomic information is limited or contradictory, taxator and binner assign identifiers to higher ranking taxa in a conservative fashion to obtain the most reliable taxonomic assignments.

3.3.2 The taxonomic assignment algorithm (taxator)

The input to the algorithm is a segment q of the original query sequence from an (unknown) taxon Q and a set of homologous segments with known taxon IDs. The term "segment" refers to a gap-less subsequence of either the query or a reference sequence. Given that for the set of homologs we know the correct underlying species tree of taxa (Figure 3.2 a), we can see that for our query taxon Q, the closest evolutionary neighbors would be A, B and S. If we simply assign X, the parental taxon of A, B and S, as a taxon identifier, this would be inaccurate, as A, B and S are more closely related to each other than to Q. Instead, the correct taxonomic assignment would be a parent of X and Q, and of at least one additional outgroup taxon (O) in the reference tree, such that Q also becomes a descendant of the identified parent (R in Figure 3.2 a). If we therefore identify the taxa A, B, S and O in the reference tree, we can determine the taxon ID of R as the lowest common ancestor (LCA) of these taxa and assign it to Q (and q).

Assuming that the underlying segment tree for a set of homologs is similar to the species tree, a natural procedure to identify the segments corresponding to the leaf taxa within R among the homologs would be to construct a MSA for the segment and a phylogenetic tree with a corresponding subtree as in Figure 3.2 a. However, the computational effort for this approach is superlinear with respect to the number of homologs being compared and substantial for all the query segments in a large sample, even using fast techniques for MSA construction and tree inference. The taxator algorithm attempts to identify these segments with a linear number of pairwise segment comparisons. Let us consider an undirected graph in which nodes represent the segments (tree leaves) and edge lengths the evolutionary distances between pairs of segments within the underlying tree (Figure 3.2 b). In this graph, a monophyletic group in the species tree is a subgraph.

For all pairs of subgraph nodes, the following inequality is true, given that the segments have evolved with a constant rate of evolution (i.e. the segment tree is ultrametric): The distance between any two subgraph nodes is smaller than that to any other node outside the subgraph. The relationship becomes clearer when thinking of the evolutionary distance between two nodes as the divergence time from their most recent ancestor. Members of a monophyletic group derive from a single common ancestor and thus there is a maximum distance for all possible pairs. If one member's distance to an outside node is smaller than this maximum, both must share a more recent common ancestor and the corresponding group is not monophyletic by definition. The stated inequality can be used to augment an incomplete group or corresponding subgraph iteratively by taking an internal distance, ideally close to the maximum, as a threshold and adding outside nodes to the group which have a smaller distance to some internal node.

In this manner, taxator-tk searches for the leaf node taxa of clade R among all segments based on a linear number of sequence comparisons between the input segments and adds them to an empty working set M:

0. A ranking by alignment scores from the input local alignments is used at the beginning to identify the reference segment s that is most similar to the query *q*. The working set *M* is then augmented in two passes:
1. In the first pass, all segments are aligned to s using fast nucleotide alignment and the edit distance. These scores in the following serve to approximate the evolutionary distances in the underlying segment phylogeny. All segment taxa with a distance less than or equal to the threshold *distance(s,q)* are added to *M* (Figure 3.2 c).
2. The outgroup segment o is determined as the first segment for which *distance(s,o)* is larger than *distance(s,q)*. In the second pass, all segments are then aligned to o and segment taxa with distances smaller than or equal to *distance(o,q)* are added to *M* as well (Figure 3.2 d).

This procedure requires approximately $2n$ alignments, where n is the number of reference segments.

3. The resulting set M of taxa (implicit in the partially resolved tree in Figure 3.2 e) is used to determine the taxon ID for q, corresponding to the LCA of these taxa in a reference taxonomy, such as the NCBI taxonomy.

If no outgroup could be determined or if M is so diverse that the LCA corresponds to the taxonomy root, q is left unassigned. The algorithm requires at least two homologous segments (s and o) to determine a meaningful taxon ID. The taxa in M become more diverse if the alignment scores are inaccurate ultrametric distance estimates, if the species subtree's topology deviates from the respective part of the taxonomy or if the gene tree's topology deviates from the species tree, for instance due to varying rates of evolution or the inclusion of non-homologous segments in the analysis. The robustness of the algorithm in avoiding incorrect assignments under these circumstances relies on the number of taxa in M and the subsequent LCA operation. Further details relating to the robustness of the implementation are given in the Supplementary Methods, "I. Taxonomic Assignment of Sequence Segments".

3.3.3 Evaluation procedures

Before evaluating any method, we removed the smallest predicted bins (1%) as likely errors. We used the macro-precision and macro-recall as measures of assignment performance (Supplementary Methods, "Performance measures"). The macro-precision specifies the fraction of correct assignments per predicted bin (precision), averaged over all such bins, while the macro-recall measures the fraction of correctly recovered sequence data per truly existing bin (recall), averaged over all such bins. To account for strong differences in bin size, we also pooled the species, genus and family assignments, and reported the overall precision for these ranks as the total fraction of correct assignments. We tested the assignment performance of different methods using three simulated short read datasets, simulated 16S rRNA data, three simulated metagenome contig datasets and using assembled cow rumen metagenome contigs. For every simulated dataset, we performed seven cross-validation experiments (Supplementary Methods, "VII.

Cross-validation"). In each experiment, we simulated a specific minimum taxonomic distance between a query sequence and the reference sequences. For the first experiment, all reference data, including the species genome data from which the query had been sampled, were made available to the method for assigning a single query sequence as an idealized test case. In the other six scenarios, all reference data belonging to the species, genus, family, order, class or phylum of the query sequence, respectively, were made inaccessible for the method in leave-one-taxon-out cross-validation experiments. We summarized the sequence assignments from these experiments to characterize a method's assignment performance across the entire range of taxonomic distances. For evaluation with the cow rumen metagenome sample, for which no true taxonomic labels were known, we divided the assembled contigs into multiple sequence "chunks" and characterized the consistency of taxonomic assignments for chunks originating from same contig (Supplementary Methods, "VIII. Consistency Analysis").

3.4 Results

3.4.1 Evaluation with unassembled data

We first evaluated the performance of taxator-tk for classification of the most widely used taxonomic marker in bacterial diversity studies – the 16S rRNA gene (Supplementary Material, Supplementary Fig. 3). This served as a proof of concept, as taxator-tk classifies arbitrary sequence regions including taxonomic marker genes. We did not expect it to perform better than sophisticated phylogenetic models for this task, but wanted to confirm a satisfactory performance. The macro-precision for the taxonomic assignment of 7176 16S rRNA genes (Supplementary Fig. 4) was constantly above 92% (Supplementary Fig. 3a) in the combined cross-validation (Methods), using the whole-genome reference sequences in mRefSeq47 (Supplementary Fig. 5), not just the 16S genes. More precisely, the average error rate per bin (one minus precision) was 7.4% at the species level and 4.6% at the order level.

Next, we simulated 100,000 reads at 100, 500 and 1000 bp by subsampling ran-

domly from 1729 species in mRefSeq47 and evaluated taxator-tk with these three datasets using the (combined) cross-validation experiments. The performance was very similar for the different fragment sizes (Supplementary Fig. 6-8a). Overall, taxator-tk showed high precision in simulated read assignment: The macro-precision for all short read lengths remained above 74% and was 82–99% for the genus to kingdom ranks, about 10% lower on average than for the 16S data. This was still good for the assignment of short sequence fragments from arbitrary genomic regions compared to a marker gene. At genus level, the macro-recall was 19–23% (~33% genera recovered) if genome sequences of the same species as the query sequence were provided in the reference (Supplementary Fig. 6-8b) and as low as 5–7% (~16% genera recovered) otherwise (Supplementary Fig. 6-8c). The macro-recall depends on the availability of related reference data at the respective ranks. It decreases when removing reference data for cross-validation. For example, if all reference data at genus level are removed, then no correct assignments to the genus rank are possible. For lower taxonomic ranks, the macro-recall was also low due to the large number of sample taxa and their uneven representation caused by the taxonomic bias towards a few abundant phyla in mRefSeq47. The longer reads had a slightly higher macro-recall than the shorter ones. Since longer sequences yield better recall and because overlapping reads contain redundant information, leading to more alignment computations, we recommend applying taxator-tk to (partially) assembled data. For longer query sequences, we were more likely to find segments for processing and therefore to assign a larger portion of the sample.

3.4.2 Evaluation with simulated metagenome contigs

For our tests on three simulated contig samples, we compared taxator-tk to CARMA3 and MEGAN4/5 using the same taxonomy and the same nucleotide alignments against mRefSeq54 (Supplementary Fig. 9). Additionally, we applied these three methods to two datasets using protein-level alignments which we inferred using BLAST+/tblastx. When doing so, we used the programs recommended parameter settings (Supplementary Methods, “X. Program Parameters and Versions”) and cross-validation, as before (Supplementary Methods, “V.

Cross-validation").

We created a simulated NGS metagenome dataset (simArt49e, composition in Supplementary Fig. 10) for our evaluation. This sample includes 49 equally abundant species (51 strains) and was created by Illumina paired read simulation with pIRS (Hu et al., 2012), followed by SOAPdenovo version 1.05 (Luo et al., 2012) assembly. Around 160 Mbp or 267,178 contigs remained after removal of 0.03% chimeric sequences. In the combined cross-validation with this dataset (Supplementary Fig. 11–13a), taxator-tk produced substantially fewer errors: Sequence assignments to species, genus and family were 91% correct for taxator-tk, compared to 52% for CARMA3 and 59% for MEGAN4. Accordingly, taxator-tk showed the highest macro-precision of all methods, e.g. 61% at the species level, compared to 3% (CARMA3) and 5% (MEGAN4). The low macro-precision observed for CARMA3 and MEGAN4 is largely due to the prediction of many small bins with many false assignments (Supplementary Methods, "V. Performance Measures"). The majority of assignments were to Bacteria, Archaea, or undetermined in the case of CARMA3, because we restricted the availability of similar reference sequences in each of the individual cross-validations, which we then jointly assessed.

When only the sequences from the corresponding species and genus were removed from the reference (new genus scenario, Supplementary Fig. 11–13d), taxator-tk was also the most precise, though it had a lower recall than the other methods (taxator-tk: 56% family macro-precision, 60% overall precision for species to family, 10% family macro-recall; CARMA3: 13%, 27% and 20%; MEGAN4: 22%, 27% and 31%). Differences in assignment precision were also evident in the number of predicted taxon bins: For instance, when simulating novel families (Supplementary Fig. 11–13e), many more species bins were predicted by CARMA3 (1672) and MEGAN4 (824) than by taxator-tk (65), with 49 species being present in the sample. Similarly, MEGAN4 predicted 69 orders, CARMA3 81 and taxator-tk 27, compared to the existing 32 orders in simArt49e (Figure 3.3). Overall, taxonomic assignments of taxator-tk were more rarely to false taxa at low ranks than with the other methods, and instead were to higher-ranking correct taxa. The other two methods assigned a substantial amount of sequence data in-

correctly to bins at the family level or below. This can be seriously misleading if the results were to be used to estimate species diversity or to reconstruct genomes. Therefore, taxator-tk is better suited for taxonomic profiling in addition to its primary task – the recovery of individual taxonomic sequence bins from shotgun datasets.

To investigate the reason for the observed differences between overall and macro-precision, which reflect variations in assignment precision for bins of different sizes, we plotted the per-bin precision at the family level in the combined cross-validation, as a function of predicted bin size with a k-nearest-neighbor (kNN) estimate of macro-precision (Figure 3.4; see Supplementary Fig. 14 for all ranks). Overall, the bins predicted by taxator-tk were smaller, more precise and much more likely to represent truly existing taxa than those predicted by the other programs although larger bins tended to be more accurate for all methods. CARMA3 and MEGAN4 predicted a substantial number of mostly smaller-sized incorrect bins. Although the size-dependent kNN precision curves at large bin sizes is unaffected by these small bins, the curves remained below 70% (CARMA3) or 80% (MEGAN4), whereas the taxator-tk curve reached almost 100%. For the smallest bins, taxator-tk's kNN precision was ~20% whereas bins below 500 kbp for CARMA3 and MEGAN4 were practically indistinguishable from noise. This shows that the high macro-precision with taxator-tk is not only due to a lower frequency of falsely predicted bins, but also due to a substantially higher precision for the large bins.

Next, we performed cross-validation on the FAMeS (Mavromatis et al., 2007) SimMC/AMD (~17 Mbp/7307 contigs) and SimHC/soil (~17 Mbp/7307 contigs) simulated metagenome datasets. These contigs were assembled from simulated Sanger (not NGS) reads and represent considerably smaller samples than those which are generated with the current NGS technologies (Dröge & McHardy, 2012). We also measured the methods' performance on these data for a direct comparison to previous works. As before, taxator-tk had the highest macro-precision and the most realistic number of predicted taxon bins (Supplementary Fig. 15, 16; Supplementary Methods "XII. FAMeS Cross-validation").

For the contig assignments of the composition-based program PhyloPythiaS

(Patil et al., 2011), we could not apply cross-validation, due to the computational effort of training many models. Therefore we adopted the published evaluation scenario from (Patil et al., 2011), in which all genome sequences of the SimMC genera were removed from the reference genome sequence before classifying the contigs. All programs were provided with the remaining sequenced genomes and an additional 100 kbp of reference data for each of the three dominant strains. The latter could be used by PhyloPythiaS to infer a corresponding species model, but were less helpful for the similarity-based classifiers. We generated assignments with taxator-tk, CARMA3 and MEGAN4/5 under equivalent conditions, once with nucleotide and once separately with protein local alignments, and compared them to both Kraken and the published PhyloPythiaS assignments (Supplementary Fig. 21). The performance and error distributions for the similarity-based programs (Supplementary Fig. 21c-d) were consistent with our previous evaluations with SimMC. MEGAN4 and MEGAN5 produced almost identical results. Using protein local alignments, we observed a moderate increase in overall species to family precision for MEGAN5 and CARMA3, while taxator-tk improved in macro-recall. Notably, taxator-tk showed the best macro-precision of all similarity-based programs and all ranks, regardless of which alignment kind was used. Kraken produced most errors and the lowest macro-precision, because it assigned almost exclusively at species level. This would make it generally unsuitable in situations where sequences of closely related genomes are unavailable. However, it had a comparatively high macro-recall up to the order level.

Assignment with PhyloPythiaS showed that composition-based classification, when supplied with limited amounts of additional training data from the relevant species, correctly assigned most data at the genus and family levels (species assignments were not assessed in the original publication), which were either rarely assigned by taxator-tk or mostly incorrectly assigned by CARMA3, MEGAN and Kraken. However, PhyloPythiaS predicted only 6 families compared to 29 underlying families, versus 43 (Kraken), 14/18 (taxator-tk), 50/32 (CARMA3) and 17/18 (MEGAN5) with nucleotide or protein alignments, respectively. PhyloPythiaS had the highest macro-recall. The macro-precision (~50% for genus, family and order level) was also higher than for Kraken (~4-13%), MEGAN (~7–

31%) or CARMA3 (~7–48%) but less than for taxator-tk (~32–68%). However, unlike for the other programs, the modeled taxa for PhyloPythiaS should be specified a priori to achieve optimal performance. It is therefore best applied when the taxonomic composition of a microbial community has already been determined and sufficient training data are available for the identified taxa.

3.4.3 Evaluation with real metagenome contigs

For microbial communities in many environments, only distantly related reference genome sequences are available. We analyzed a medium complex metagenome sample of such a microbial community from cow rumen (Hess et al., 2011) with taxator-tk, CARMA3, MEGAN4/5 and PhyloPythiaS (the general model with the 100 most abundant species among sequenced prokaryotes). We considered scaffolds to be less reliable than contigs, which we reconstructed by splitting the available scaffolds at gaps of more than 200 positions (A. Sczyrba, personal communication). We subsequently divided contigs longer than 10 kbp into sequence "chunks" of 2 kbp, resulting in a 319 Mbp dataset, which we used to assess the assignment consistency for chunks originating from the same contig. The chunk sequences were assigned with taxator-tk, CARMA3, MEGAN (given identical nucleotide/protein alignments), Kraken and PhyloPythiaS. As the standard of truth for each contig, we determined the taxon minimizing the inconsistency between all corresponding chunk assignments (Gregor et al., 2014, unpublished) for each method independently. A chunk assignment was considered consistent, if it was to the same taxon as the one for entire contig, and inconsistent otherwise. The consistency of a taxonomic bin is the fraction of chunk sequences with matching contig assignments and the macro-consistency is the consistency averaged over all predicted taxa, similar to the macro-precision.

In agreement with the results for the simulated metagenome datasets, the taxator-tk results were the most consistent among all tested methods, regardless of the alignment type (Supplementary Fig. 22): 76–89% macro-consistency at species to order level, in comparison to MEGAN (34–40%), CARMA (0-55%), Kraken (32–35%) and PhyloPythiaS (56–65%). The overall consistency (analogous to overall

precision) for species to family levels was 97/97% with taxator-tk, 39/48% with CARMA3, 62/64% with MEGAN (nucleotide/protein-level), 42% with Kraken and 82% with PhyloPythiaS. Likewise, taxator-tk assigned less data at species to family level, with a total of 13/12 Mbp being consistent compared to CARMA3 (8/26 Mbp), MEGAN (42/47 Mbp), Kraken (19 Mbp) or PhyloPythiaS (14 Mbp). The different methods again determined different numbers of taxa: CARMA3 predicted 572/611 genera with a macro-consistency of 53/31%, MEGAN 264/203 genera (34/37%), Kraken 661 (32%), PhyloPythiaS 33 (63%) and taxator-tk found 110/27 genera (76/81%). The high consistency values observed for taxator-tk indicate that it is a precise taxonomic classifier for real metagenomic contigs.

3.4.4 Run-time analyses

The run-time for the taxonomic metagenome assignment was measured as the time to find homologs and to assign taxon IDs to all sequences. We evaluated the run-times of all methods using the same set of alignments generated with either BLAST or LAST. Thus the run-time for the initial similarity search was identical for all methods. We determined the time for the taxonomic assignment of simArt49e for all methods when performing a cross-validation with families present in the test dataset removed from the reference data (Figure 3.3). This took two minutes with Kraken (single CPU core and ~100 GiB RAM), one hour for MEGAN4 (interactive mode), 6 hours for taxator-tk (~10 CPU cores) and almost a week for CARMA3 (~20 CPU cores). The parallelization of taxator-tk led to a linear decrease in time with the number of CPU cores for up to 15 cores, which became sublinear for 20 cores or more (Supplementary Fig. 23). To provide a more specific estimate of the throughput of taxator-tk, we aligned ~1 Gbp of cow rumen sequence data with BLAST against mRefSeq54 and assigned the data with taxator-tk on 10 CPU cores (AMD Opteron 6386 SE). We measured an average throughput of 5.9 Gbp per day for the combined alignment and taxonomic assignment steps with this dataset. We also determined how our implementation scaled for increasing input sequence lengths and reference exclusion scenarios (Supplementary Fig. 24a). The run-time scaled approximately linearly except when the same or very similar species were among the reference genome sequences.

In general, the greater the number of similar sequences in the reference data, the longer taxator-tk's run-time was for the alignment of longer sequence stretches with more homologs. Simultaneously, we investigated the impact of the query segmentation on taxator-tk's run-time (Supplementary Fig. 24b) and found that it reduced the total run-time by up to 30%.

3.5 Discussion

Taxator-tk is a taxonomic assignment software package which generates very precise taxonomic assignments with few errors for metagenome shotgun sequences. To provide a fair comparison, we invested extensive effort into ensuring that we evaluated all methods under identical conditions with the same reference sequences, test datasets and background taxonomies, using their recommended settings. We evaluated taxator-tk on 16S gene sequences, on simulated short reads, with simulated assembled contigs and with 2 kbp contig fragments from a real cow rumen metagenome. For each simulated sample, we evaluated a wide range of evolutionary distances between the query and reference sequences using leave-one-taxon-out cross-validation. Taxator-tk was the most precise of all tested methods with the most realistic number of identified taxa overall. This property was very pronounced for lower taxonomic ranks from species to family level. However, taxator-tk assigned fewer data overall than other methods from species to family. For the small assembled SimMC dataset, it assigned fewer data, particularly in comparison to the composition-based classifier PhyloPythiaS, when 100 kbp of data were provided for individual community members to train species-level models. For the real cow rumen dataset, taxator-tk was the most consistent in terms of classifying multiple pieces of one contig. Our results consistently indicate that taxator-tk's strength is its high precision of assignments, which allows us to confidently assign a core of sample sequences and thereby to infer the taxonomic composition of the community. In comparison to assignments based on marker genes, it has the advantages that it makes assignments across all domains of life and that corresponding abundance estimates from shotgun sequences are less affected by copy number variations of individual genes. Such shotgun estimates are

also unaffected by PCR primer amplification biases, unlike marker gene sequencing techniques, and do not require high-quality reference gene phylogenies for marker genes. We confirmed this by in depth analysis of six 15 Gbp shotgun samples from the barley rhizosphere, where we applied taxator-tk to characterize the taxonomic composition of Bacteria, Archaea and Eukaryotes, which correlated with results from 16S rRNA profiling and showed the most notable deviations for taxa known to be affected by primer biases or having multiple copies of the 16S rRNA gene (Bulgarelli et al., unpublished). To target draft genome reconstructions, the data assigned to individual taxonomic bins by taxator-tk can be used as training data for complementary approaches, such as composition-based methods, or as independent information in combination with recently proposed clustering methods using the abundance of genes or contigs across multiple samples.

From a methodological point of view, we have introduced a method for the fast approximation of the evolutionary neighborhood of a query sequence with a run-time that increases linearly with the number of homologs. In de-novo phylogenetic inference methods, the run-time increases at least log-linearly with the number of homologs or they rely on time-consuming optimizations of parameter-rich phylogenetic models, which generates excessive computational requirements for the analysis of Gbp-sized NGS samples. Our software provides an easy to use and scalable alternative to taxonomic classification of marker genes that is applicable to any nucleotide fragment. Unlike other similarity-based taxonomic classifiers for shotgun data, our algorithm handles different degrees of sequence conservation without preset or user-specified parameters such as alignment scores (overall or per gene family) and without being restricted to the analysis of a number of high-quality homologs with a minimal length. At the same time, the inferred evolutionary neighborhood is extended by the identification of an outgroup, leading to more precise taxonomic assignments, while regions without detectable taxonomic signal are instantly discarded. We post-process independent taxonomic assignments of query segments to infer an assignment for the entire query and do this using a majority vote algorithm with a few robust default parameters. This computationally lightweight step can be quickly repeated with other values for the majority and minimum support parameters, if required. In addition to the algorithmic considerations and other run-time optimizations, we implemented query

sequence segmentation and program parallelization, which allow large-scale data analysis with a throughput of several Gbs per day on a standard multiprocessor system.

The program's scope is also not limited to the taxonomic assignment of metagenomes: It can be applied to any DNA or RNA sequence. For instance, another successful in-house application is the detection of contaminations in isolate sequencing data. Furthermore, the program taxator within taxator-tk provides taxonomic information for individual query segments (Supplementary Fig. 2, 25), which could be used to identify assembly errors or regions acquired by lateral gene transfer.

3.6 Acknowledgments

Computational support and infrastructure was provided by the "Centre for Information and Media Technology" (ZIM) at the University of Düsseldorf (Germany).

3.7 Funding

The authors gratefully acknowledge funding by the Max-Planck society and Heinrich Heine University Düsseldorf.

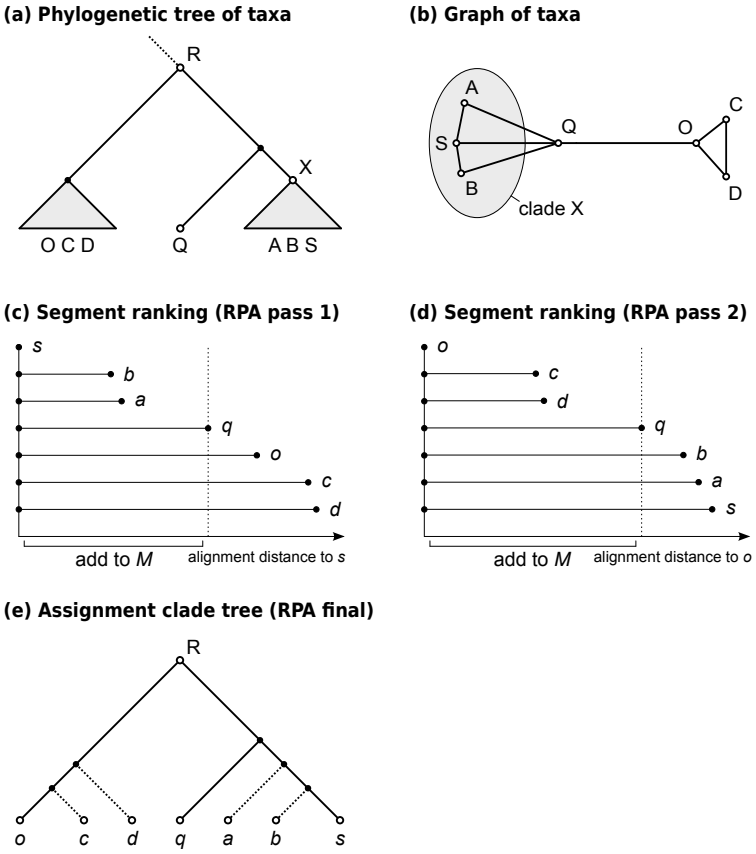


Figure 3.2: Algorithm for taxonomic labeling of query segments (realignment placement algorithm/RPA). The RPA assigns a taxon ID to a query segment *q*. (a) Species reference tree with query taxon Q and reference taxa A, B, C, D, O and S. This will be approximated by the segment phylogenetic tree for the query segment and homologous segments of reference taxa. (b) Approximate graph representing pairwise distances between the taxa. The subgraph for clade X is highlighted. (c,d) Show the two alignment passes which add segment taxa to an (empty) set *M*. Segment *s* is the segment with the smallest local alignment score (distance) to *q* in the initial similarity search. (c) First, all segments are aligned to segment *s*. The resulting distances are ordered and the taxa with equal or smaller distances than *distance(s,q)* are added to *M*. The outgroup segment, here *o*, is the next most similar segment to *s* after *q*, with *distance(o,s)* > *distance(s,q)*. (d) All segments are aligned to *o*. From the ranked distances, taxa with distances smaller than *distance(o,q)* are also added to *M*. Thus, *M* includes all the nearest evolutionary neighbors for the query segment *q* (the taxa corresponding to segments *a*, *b*, *c*, *d*, *o*, *s*). The taxon ID then assigned to *q* is the lowest common ancestor in the reference species tree (reference taxonomy) of these taxa in *M*. (e) Partially resolved segment subtree at node R that is implied by distances obtained in (c) and (d), where the exact position of some segments (*a*, *b*, *c*, *d*; dashed branches) is left unresolved by the RPA.

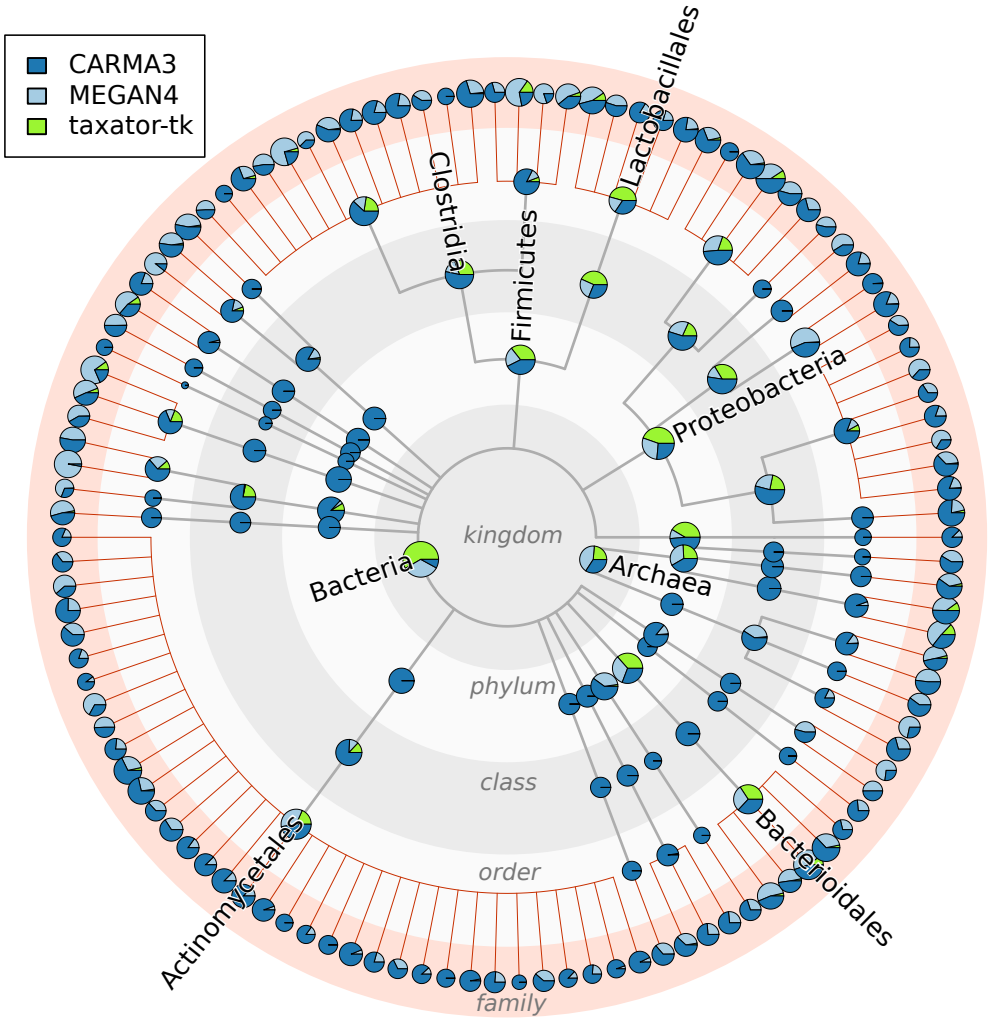


Figure 3.3: Comparison of three classifiers for a novel-family simulation using a simulated metagenome sample (simArt49e) with 49 species. CARMA3, MEGAN4 and taxator-tk: The outer ring with red background shading shows family-level assignments for all orders included in the simulated data set. These are all false in the chosen evaluation scenario, as no data from the families of the query sequences were included in the reference collection in the leave-one-taxon-out cross-validation experiments. Clearly, taxator-tk had the fewest assignments at family level, demonstrating its high precision in assignments. Assignments at inner rings, grey background shading, can be correct in principle, demonstrating at which taxonomic ranks the different methods tend to make their assignments, with taxator-tk tending towards producing higher-ranking assignments, as a trade-off for the high precision.

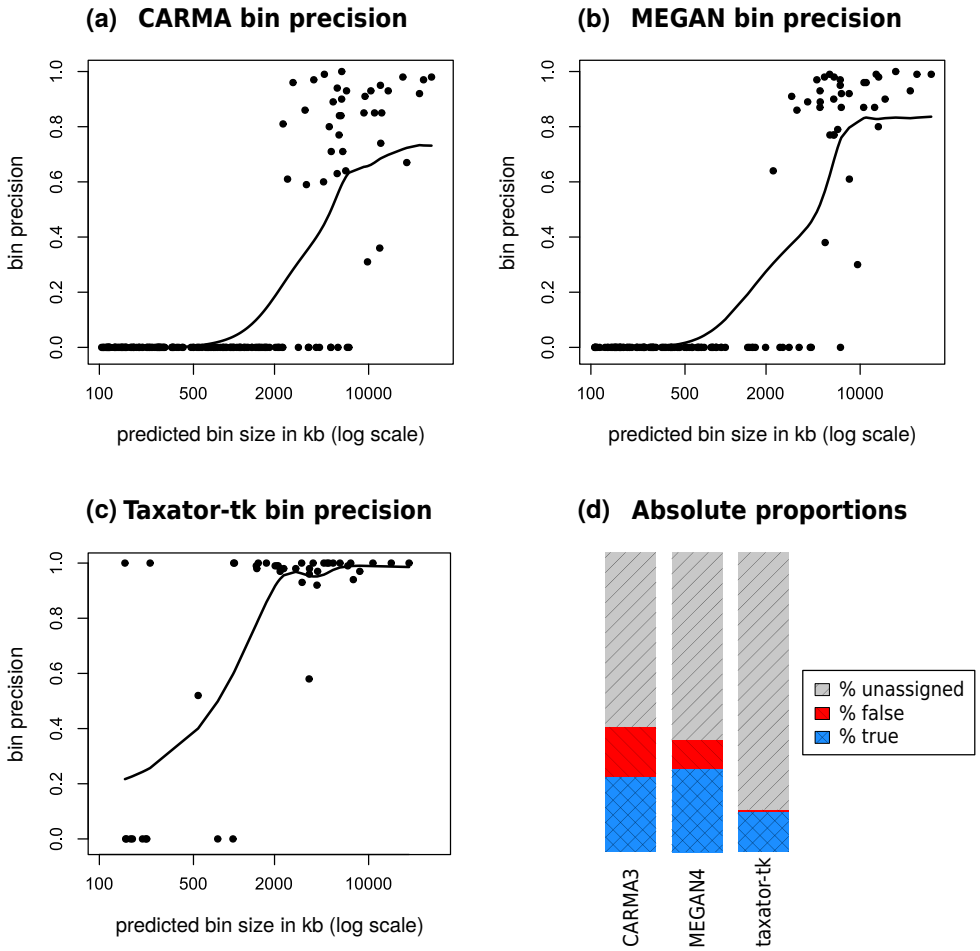


Figure 3.4: Family-level bin precision for the simulated metagenome sample with 49 species (simArt49e). (a-c) Each family bin's assignment precision related to logarithmic bin size for seven cross-validation experiments with simArt49e. The results of the single experiments were added to assess the taxonomic assignment performance across a range of evolutionary distances between the query and the reference sequences, excluding the least abundant bins (1% of total bp). We calculated the precision values for (a) CARMA3, (b) MEGAN4 and (c) taxator-tk, counting assignments to lower-ranking taxa at the family level, and added a smoothed k-nearest-neighbor estimate of the mean precision in R using wapply (width=0.3) followed by smooth.spline (df=10). CARMA3 and MEGAN4 incorrectly identified many small taxonomic bins, substantially more than taxator-tk. (d) gives the amount of correct, false and undetermined family-level assignments for the different classifiers with simArt49e.

Chapter 4

A Probabilistic Model to Recover Genomes in Shotgun Metagenomics

Johannes Dröge[1], Alexander Schönhuth[2], Alice C. McHardy[1*]

[1]Helmholtz Centre for Infection Research, Braunschweig, Germany

[2]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

---

This is an author-produced version of an article under revision in *PeerJ Computer Science*. This article version has been adapted to the thesis layout. The original open-access article is accessible by DOI 10.7287/peerj.preprints.2626.

4.1 Abstract

Shotgun metagenomics of microbial communities reveals information about strains of relevance for applications in medicine, biotechnology and ecology.

Recovering their genomes is a crucial, but very challenging step, due to the complexity of the underlying biological system and technical factors. Microbial communities are heterogeneous, with oftentimes hundreds of present genomes deriving from different species or strains, all at varying abundances and with different degrees of similarity to each other and reference data. We present a versatile probabilistic model for genome recovery and analysis, which aggregates three types of information that are commonly used for genome recovery from metagenomes. As potential applications we showcase metagenome contig classification, genome sample enrichment and genome bin comparisons. The open source implementation MGLEX is available via the Python Package Index and on GitHub and can be embedded into metagenome analysis workflows and programs.

4.2 Introduction

Shotgun sequencing of DNA extracted from a microbial community recovers genomic data from different community members while bypassing the need to obtain pure isolate cultures. It thus enables novel insights into ecosystems, especially for those genomes which are inaccessible by cultivation techniques and isolate sequencing. However, current metagenome assemblies are oftentimes highly fragmented, including unassembled reads, and require further processing to separate data according to the underlying genomes. Assembled sequences, called contigs, that originate from the same genome are placed together in this process, which is known as metagenome binning (Tyson et al., 2004; Dröge & McHardy, 2012) and for which many programs have been developed. Some are trained on reference sequences, using contig k-mer frequencies or sequence similarities as sources of information (McHardy et al., 2007; Dröge, Gregor & McHardy, 2014; Wood & Salzberg, 2014; Gregor et al., 2016), which can be adapted to specific ecosystems. Others cluster the contigs into genome bins, using contig k-mer frequencies and read coverage (Chatterji et al., 2008; Kislyuk et al., 2009; Wu et al., 2014; Nielsen et al., 2014; Imelfort et al., 2014; Alneberg et al., 2014; Kang et al., 2015; Lu et al., 2016).

Recently, oftentimes multiple biological or technical samples of the same environment are sequenced to produce distinct genome copy numbers across samples, sometimes using different sequencing protocols and technologies, such as Illumina and PacBio sequencing (Hagen et al., 2016). Genome copies are reflected by corresponding read coverage variation in the assemblies which allows to resolve samples with many genomes. The combination of experimental techniques helps to overcome platform-specific shortcomings such as short reads or high error rates in the data analysis. However, reconstructing high-quality bins of individual strains remains difficult without very high numbers of replicates. Often, genome reconstruction may improve by manual intervention and iterative analysis (Figure 4.1) or additional sequencing experiments.

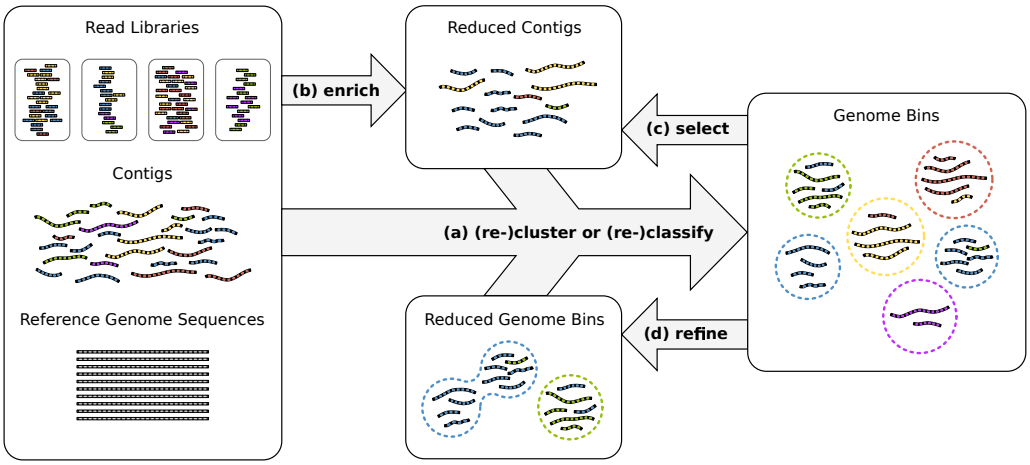


Figure 4.1: Genome reconstruction workflow. To recover genomes from environmental sequencing data, the illustrated processes can be iterated. Different programs can be run for each process and iteration. MGLEX can be applied in all steps: (a) to classify contigs or to cluster by embedding the probabilistic model into an iterative procedure; (b) to enrich a metagenome for a target genome to reduce its size and to filter out irrelevant sequence data; (c) to select contigs of existing bins based on likelihoods and p-values and to repeat the binning process with a reduced dataset; (d) to refine existing bins, for instance to merge bins as suggested by bin analysis.

Genome bins can be constructed by consideration of genome-wide sequence properties. Currently, oftentimes the following types of information are considered:

- Read contig coverage: sequencing read coverage of assembled contigs, which reflects the genome copy number (organismal abundance) in the community. Abundances can vary across biological or technical replicates, and co-vary for contigs from the same genome, supplying more information to resolve individual genomes (Baran & Halperin, 2012; Albertsen et al., 2013).
- Nucleotide sequence composition: the frequencies of short nucleotide subsequences of length k called k-mers. The genomes of different species have a characteristic k-mer spectrum (Karlin, Mrazek & Campbell, 1997; McHardy et al., 2007).
- Sequence similarity to reference sequences: a proxy for the phylogenetic relationship to species which have already been sequenced. The similarity is usually inferred by alignment to a reference collection and can be expressed using taxonomy (McHardy et al., 2007).

Probabilities represent a convenient and efficient way to represent and combine information that is uncertain by nature. Here, we

- propose a probabilistic aggregate model for binning based on three commonly used information sources, which can easily be extended to include new features.
- outline the features and submodels for each information type. As the feature types listed above derive from distinct processes, we define for each of them independently a suitable probabilistic submodel.
- showcase several applications related to the binning problem

A model with data-specific structure poses an advantage for genome recovery in metagenomes because it uses data more efficiently for fragmented assemblies with short contigs or a low number of samples for differential coverage binning. Being probabilistic, it generates probabilities instead of hard labels so that a contig can be assigned to several, related genome bins and the uncertainty can easily be assessed. The models can be applied in different ways, not just classification, which we show in our application examples. Most importantly, there is a rich

repertoire of higher-level procedures based on probabilistic models, including Expectation Maximization (EM) and Markov Chain Monte Carlo (MCMC) methods for clustering without or with few prior knowledge of the modeled genomes.

We focus on defining explicit probabilistic models for each feature type and their combination into an aggregate model. In contrast, binning methods often concatenate and transform features (Chatterji et al., 2008; Imelfort et al., 2014; Alneberg et al., 2014) before clustering. Specific models for the individual data types can be better tailored to the data generation process and will therefore generally enable a better use of information and a more robust fit of the aggregate model while requiring fewer data. We propose a flexible model with regard to both the included features and the feature extraction methods. There already exist parametric likelihood models in the context of clustering, for a limited set of features. For instance, Kislyuk et al. (2009) use a model for nucleotide composition and Wu et al. (2014) integrated distance-based probabilities for 4-mers and absolute contig coverage using a Poisson model. We extend and generalize this work so that the model can be used in different contexts such as classification, clustering, genome enrichment and binning analysis. Importantly, we are not providing an automatic solution to binning but present a flexible framework to target problems associated with binning. This functionality can be used in custom workflows or programs for the steps illustrated in Figure 4.1. As input, the model incorporates genome abundance, nucleotide composition and additionally sequence similarity (via taxonomic annotation). The latter is common as taxonomic binning output (Dröge, Gregor & McHardy, 2014; Wood & Salzberg, 2014; Gregor et al., 2016) and for quality assessment but has rarely been systematically used as features in binning (Chatterji et al., 2008; Lu et al., 2016). We show that taxonomic annotation is valuable information that can improve binning considerably.

4.3 Methods

4.3.1 Classification models

Classification is a common concept in machine learning. Usually, such algorithms use training data for different classes to construct a model which then contains the condensed information about the important properties that distinguish the data of the classes. In probabilistic modeling, we describe these properties as parameters of likelihood functions, often written as θ. After θ has been determined by training, the model can be applied to assign novel data to the modeled classes. In our application, classes are genomes, or bins, and the data are nucleotide sequences like contigs. Thus, contigs can be assigned to genomes bins but we need to provide training sequences for the genomes. Such data can be selected by different means, depending on the experimental and algorithmic context. One can screen metagenomes for genes which are unique to clades, or which can be annotated by phylogenetic approaches, and use the corresponding sequence data for training (Gregor et al., 2016). Independent assemblies or reference genomes can also serve as training data for genome bins (Brady & Salzberg, 2009; Patil et al., 2011; Gregor et al., 2016). Another direct application is to learn from existing genome bins, which were derived by any means, and then to (re)assign contigs to these bins. This is useful for short contigs which are often excluded from binning and analysis due to their high variability. Finally, probabilistic models can be embedded into iterative clustering algorithms with random initialization.

4.3.2 Aggregate model

Let $1 \leq i \leq D$ be an index referring to D contigs resulting from a shotgun metagenomic experiment. In the following we will present a generative probabilistic aggregate model that consists of components, indexed by $1 \leq k \leq M$, which are generative probabilistic models in their own right, yielding probabilities $P_k(\text{contig}_i \mid \text{genome})$ that contig_i belongs to a particular genome. Each of the components k reflects a particular feature such as

- a weight w_i (contig length)
- sample abundance feature vectors $\boldsymbol{a_i}$ and $\boldsymbol{r_i}$, one entry per sample
- a compositional feature vector $\boldsymbol{c_i}$, one entry per compositional feature (e.g. a k-mer)
- a taxonomic feature vector $\boldsymbol{t_i}$, one entry per taxon

We define the individual feature vectors in the corresponding sections. As mentioned before, each of the M features gives rise to a probability $P_k(\text{contig}_i \mid \text{genome})$ that contig_i belongs to a specific genome by means of its component model. Those probabilities are then collected into an aggregate model that transforms those feature specific probabilities $P_k(i \mid \text{genome})$ into an overall probability $P(i \mid \text{genome})$ that contig i is associated with the genome. In the following, we describe how we construct this model with respect to the individual submodels $P_k(i \mid \text{genome})$, the feature representation of the contigs and how we determine the optimal set of parameters from training sequences.

For the i^{th} contig, we define a joint likelihood for genome bin g (Equation 4.1, the probabilities written as a function of the genome parameters), which is a weighted product over M independent component likelihood functions, or submodels, for the different feature types. For the k^{th} submodel, $\boldsymbol{\Theta_k}$ is the corresponding parameter vector, $\boldsymbol{F_{i,k}}$ the feature vector of the i^{th} contig and α_k defines the contribution of the respective submodel or feature type. β is a free scaling parameter to adjust the smoothness of the aggregate likelihood distribution over the genome bins (bin posterior).

$$\mathcal{L}(\boldsymbol{\Theta_g} \mid \mathbf{F_i}) = \left(\prod_{k=1}^{M} \mathcal{L}(\boldsymbol{\Theta_{gk}} \mid \boldsymbol{F_{ik}})^{\alpha_k} \right)^{\beta} \tag{4.1}$$

We assume statistical independence of the feature subtypes and multiply likelihood values from the corresponding submodels. This is a simplified but reasonable assumption: e.g., the species abundance in a community can be altered by external factors without impacting the nucleotide composition of the genome or its taxonomic position. Also, there is no direct relation between a genome's k-mer distribution and taxonomic annotation via reference sequences.

All model parameters, $\mathbf{\Theta_g}$, $\boldsymbol{\alpha}$ and β, are learned from training sequences. We will explain later, how the weight parameters $\boldsymbol{\alpha}$ and β are chosen and begin with a description of the four component likelihood functions, one for each feature type.

In the following, we denote the j^{th} position in a vector $\boldsymbol{x_i}$ with $x_{i,j}$. To simplify notation, we also define the sum or fraction of two vectors of the same dimension as the positional sum or fraction and write the length of vector $\boldsymbol{x}$ as $len(\boldsymbol{x})$.

4.3.3 Absolute abundance

We derive the average number of reads covering each contig position from assembler output or by mapping the reads back onto contigs. This mean coverage is a proxy for the genome abundance in the sample because it is roughly proportional to the genome copy number. A careful library preparation causes the copy numbers of genomes to vary differently over samples, so that each genome has a distinct relative read distribution. Depending on the amount of reads in each sample being associated with every genome, we obtain for every contig a coverage vector $\boldsymbol{a_i}$ where $len(\boldsymbol{a_i})$ is the number of samples. Therefore, if more sample replicates are provided, contigs from different genomes are generally better separable since every additional replicate adds an entry to the feature vectors.

Random sequencing followed by perfect read assembly theoretically produces positional read counts which are Poisson distributed, as described in Lander & Waterman (1988). In Equation 4.2, we derived a similar likelihood using mean coverage values (see Supplementary Methods for details). The likelihood function is a normalized product over the independent Poisson functions $P_{\theta_j}(a_{i,j})$ for each sample. The expectation parameter θ_j represents the genome copy number in the j^{th} sample.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{a_i}) = \sqrt[len(\boldsymbol{a_i})]{\prod_{j=1}^{len(\boldsymbol{a_i})} P_{\theta_j}(a_{i,j})} = \sqrt[len(\boldsymbol{a_i})]{\prod_{j=1}^{len(\boldsymbol{a_i})} \frac{\theta_j^{a_{i,j}}}{a_{i,j}!} e^{-\theta_j}} \tag{4.2}$$

The Poisson explicitly accounts for low and zero counts, unlike a Gaussian model. Low counts are often observed for undersequenced and rare taxa. Note that $a_{i,j}$

is independent of $\boldsymbol{\theta}$. We derived the model likelihood function from the joint Poisson over all contig positions by approximating the first data-term with mean coverage values (Supplementary Methods).

The maximum likelihood estimate (MLE) for $\boldsymbol{\theta}$ on training data is the weighted average of mean coverage values for each sample in the training data (Supplementary Methods).

$$\hat{\boldsymbol{\theta}} = \frac{\sum_{i=1}^{N} w_i \, \boldsymbol{a_i}}{\sum_{i=1}^{N} w_i} \tag{4.3}$$

4.3.4 Relative abundance

In particular for shorter contigs, the absolute read coverage is often overestimated. Basically, the Lander-Waterman assumptions (Lander & Waterman, 1988) are violated if reads do not map to their original locations due to sequencing errors or if they "stack" on certain genome regions because they are ambiguous (i.e. for repeats or conserved genes), rendering the Poisson model less appropriate. The Poisson, when constrained on the total sum of coverages in all samples, leads to a binomial distribution as shown by (Przyborowski & Wilenski, 1940). Therefore, we model differential abundance over different samples using a binomial in which the parameters represent a relative distribution of genome reads over the samples. For instance, if a particular genome had the same copy number in a total of two samples, the genome's parameter vector $\boldsymbol{\theta}$ would simply be $[0.5, 0.5]$. As for absolute abundance, the model becomes more powerful with a higher number of samples. Using relative frequencies as model parameters instead of absolute coverages, however, has the advantage that any constant coverage factor cancels in the division term. For example, if a genome has two similar gene copies which are collapsed during assembly, twice as many reads will map onto the assembled gene in every sample but the relative read frequencies over samples will stay unaffected. This makes the binomial less sensitive to read mapping artifacts but requires two or more samples because one degree of freedom (DF) is lost by the

division.

The contig features $\boldsymbol{r_i}$ are the mean coverages in each sample, which is identical to $\boldsymbol{a_i}$ in the absolute abundance model, and the model's parameter vector $\boldsymbol{\theta}$ holds the relative read frequencies in the samples, as explained before. In Equation 4.4 we ask: how likely is the observed mean contig coverage $r_{i,j}$ in sample j given the genome's relative read frequency θ_j of the sample and the contig's total coverage R_i for all samples. The corresponding likelihood is calculated as a normalized product over the binomials $B_{R_i,\theta_j}(r_{i,j})$ for every sample.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{r_i}) = \sqrt[len(\boldsymbol{r_i})]{\prod_{j=1}^{len(\boldsymbol{r_i})} B_{R_i,\theta_j}(r_{i,j})} = \sqrt[len(\boldsymbol{r_i})]{\prod_{j=1}^{len(\boldsymbol{r_i})} \binom{R_i}{r_{i,j}} \theta_j^{r_{i,j}} \left(1-\theta_j\right)^{(R_i - r_{i,j})}} \quad (4.4)$$

R_i is the sum of the abundance vector $\boldsymbol{r_i}$. Because both R_i and r_i can contain real numbers, we need to generalize the binomial coefficient to positive real numbers via the gamma function Γ.

$$\binom{n}{k} = \frac{\Gamma(n+1)\Gamma(k+1)}{\Gamma(n-k+1)} \quad (4.5)$$

Because the binomial coefficient is a constant factor and independent of $\boldsymbol{\theta}$, it can be omitted in ML classification (when comparing between different genomes) or be retained upon parameter updates. As for the Poisson, the model accounts for low and zero counts (by the binomial coefficient). We derived the likelihood function from the joint distribution over all contig positions by approximating the binomial data-term with mean coverage values (see Supplementary Methods).

The MLE $\hat{\boldsymbol{\theta}}$ for the model parameters on training sequence data corresponds to the amount of read data (base pairs) in each sample divided by the total number of base pairs in all samples. We express this as a weighted sum of contig mean coverage values (see Supplementary Methods).

$$\hat{\boldsymbol{\theta}} = \frac{\sum_{i=1}^{N} w_i \, \boldsymbol{r_i}}{\sum_{i=1}^{N} w_i \, R_i} \tag{4.6}$$

It is obvious that absolute and relative abundance models are not independent when the identical input vectors (here $\boldsymbol{a_i} = \boldsymbol{r_i}$) are used. However, we can instead apply the Poisson model to the total coverage R_i (summed over all samples) because this sum also follows a Poisson distribution. To illustrate the total abundance, this compares to mixing the samples before sequencing so that the resolution of individual samples is lost. The binomial, in contrast, only captures the relative distribution of reads over the samples (one DF is lost in the ratio transform). This way, we can combine both absolute and relative abundance submodels in the aggregate model.

4.3.5 Nucleotide composition

Microbial genomes have a distinct "genomic fingerprint" (Karlin, Mrazek & Campbell, 1997) which is typically determined by means of k-mers. Each contig has a relative frequency vector $\boldsymbol{c_i}$ for all possible k-mers of size k. The nature of shotgun sequencing demands that each k-mer is counted equally to its reverse complement because the orientation of the sequenced strand is typically unknown. With increasing k, the feature space grows exponentially and becomes sparse. Thus, it is common to select k from 4 to 6 (Teeling et al., 2004; McHardy et al., 2007; Kislyuk et al., 2009). Here, we simply use 5-mers ($len(\boldsymbol{c_i}) = \frac{4^5}{2} = 512$) but other choices can be made.

For its simplicity and effectiveness, we chose a likelihood model assuming statistical independence of features so that the likelihood function in Equation 4.7 becomes a simple product over observation probabilities (or a linear model when transforming into a log-likelihood). Though k-mers are not independent due to their overlaps and reverse complementarity (Kislyuk et al., 2009), the model has been successfully applied to k-mers (Wang et al., 2007), and we can replace k-mers in our model with better-suited compositional features, i.e. using locality-

sensitive hashing (Luo et al., 2016). A genome's background distribution $\boldsymbol{\theta}$ is a vector which holds the probabilities to observe each k-mer and the vector $\boldsymbol{c_i}$ does the same for the i^{th} contig. The composition likelihood for a contig is a weighted and normalized product over the background frequencies.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{c_i}) = \prod_{i=1}^{len(\boldsymbol{c_i})} \theta_i^{c_i} \tag{4.7}$$

The genome parameter vector $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood on training sequence data can be estimated by a weighted average of feature counts (Supplementary Methods).

$$\hat{\boldsymbol{\theta}} = \frac{\sum_{i=1}^{N} w_i \, \boldsymbol{c_i}}{\sum_{i=1}^{N} w_i} \tag{4.8}$$

4.3.6 Similarity to reference

We can compare contigs to reference sequences, for instance by local alignment. Two contigs that align to closely related taxa are more likely to derive from the same genome than sequences which align to distant clades. We convert this indirect relationship to explicit taxonomic features which we can compare without direct consideration of reference sequences. A taxon is a hierarchy of nested classes which can be written as a tree path, for example, the species *E. coli* could be written as [Bacteria, Gammaproteobacteria, Enterobacteriaceae, *E. coli*].

We assume that distinct regions of a contig, such as genes, can be annotated with different taxa. Each taxon has a corresponding weight which in our examples is a positive alignment score. The weighted taxa define a spectrum over the taxonomy for every contig and genome. It is not necessary that the alignment reference be complete or include the respective species genome but all spectra must be equally biased. Since each contig is represented by a hierarchy of L numeric weights, we incorporated these features into our multi-layer model. First, each contig's taxon weights are transformed to a set of sparse feature vectors $\boldsymbol{t_i} = \{\boldsymbol{t_{i,l}} \mid 1 \leq l \leq L\}$,

one for each taxonomic level, by inheriting and accumulating scores for higher-level taxa (see Table 4.1 and Figure 4.2).

Table 4.1: Calculating the contig features $\boldsymbol{t_i}$ for a simplified taxonomy. There are five original integer alignment scores for nodes (c), (e), (f), (g) and (h) which are summed up at higher levels to calculate the feature vectors $\boldsymbol{t_{i,l}}$. The corresponding tree structure is shown in Figure 4.2.

Node	Taxon	Level l	Index j	Score	$t_{i,l,j}$
a	Bacteria	1	1	0	7
b	Gammaproteobacteria	2	1	0	6
c	Betaproteobacteria	2	2	1	1
d	Enterobacteriaceae	3	1	0	5
e	Yersiniaceae	3	2	1	1
f	*E. vulneris*	4	1	1	1
g	*E. coli*	4	2	3	3
h	*Yersinia sp.*	4	3	1	1

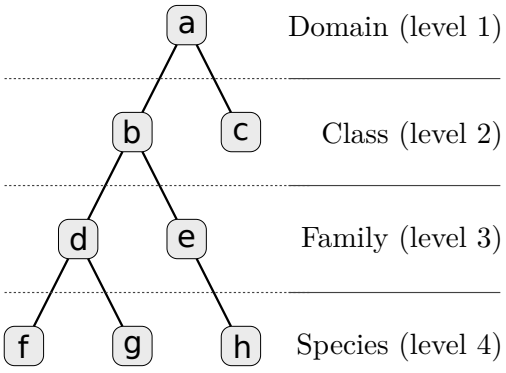


Figure 4.2: Taxonomy for which is simplified to four levels and eight nodes. A full taxonomy may consist of thousands of nodes.

Each vector $\boldsymbol{t_{i,l}}$ contains the scores for all T_l possible taxa at level l. A genome is represented by a similar set of vectors $\boldsymbol{\theta} = \{\boldsymbol{\theta_l} \mid 1 \leq l \leq L\}$ with identical dimensions, but here, entries represent relative frequencies on the particular level

l, for instance a distribution over all family taxa. The corresponding likelihood model corresponds to a set of simple frequency models, one for each layer. The full likelihood is a product of the level likelihoods.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{t_i}) = \prod_{l=1}^{L} \prod_{j=1}^{T_l} \theta_{l,j}^{t_{i,l,j}} \tag{4.9}$$

For simplicity, we assume that layer likelihoods are independent which is not quite true but effective. The MLE for each $\boldsymbol{\theta_l}$ is then derived from training sequences similar to the simple frequency model (Supplementary Methods).

$$\hat{\theta}_l = \frac{\sum_{i=1}^{N} t_{i,l}}{\sum_{j=1}^{T_l} \sum_{i=1}^{N} t_{i,l}} \tag{4.10}$$

4.3.7 Inference of weight parameters

The aggregate likelihood for a contig in Equation 4.1 is a weighted product of submodel likelihoods. The weights in vector $\boldsymbol{\alpha}$ balance the contributions, assuming that they must not be equal. When we write the likelihood in logarithmic form (Equation 4.11), we see that each weight α_k sets the variance or width of the contigs' submodel log-likelihood distribution. We want to estimate α_k in a way which is not affected by the original submodel variance because the corresponding normalization exponent is somewhat arbitrary. For example, we normalized the nucleotide composition likelihood as a single feature and the abundance likelihoods as a single sample to limit the range of the likelihood values, because we simply cannot say how much each feature type counts.

$$l(\boldsymbol{\Theta} \mid \mathbf{F_i}) = \beta \sum_{k=1}^{M} \alpha_k \, l(\boldsymbol{\Theta_k} \mid \boldsymbol{F_{i,k}}) \tag{4.11}$$

For any modeled genome, each of the M submodels produces a distinct log-likelihood distribution of contig data. Based on the origin of the contigs, which

is known for model training, the distribution can be split into two parts, the actual genome (positive class) and all other genomes (negative class), as illustrated in Figure 4.3. The positive distribution is roughly unimodal and close to zero whereas the negative distribution, which represents many genomes at once, is diverse and yields strongly negative values. Intuitively, we want to select $\boldsymbol{\alpha}$ such that the positive class is well separated from the negative class in the aggregate log-likelihood function in Equation 4.11.

Because $\boldsymbol{\alpha}$ cannot be determined by likelihood maximization, the contributions are balanced in a robust way by setting $\boldsymbol{\alpha}$ to the inverse standard deviation of the genome (positive class) log-likelihood distributions. More precisely, we calculate the average standard deviation over all genomes weighted by the amount of contig data (bp) for each genome and calculate α_k as the inverse of this value. This scales down submodels with a high average variance. When we normalize the standard deviation of genome log-likelihood distributions in all submodels before summation, we assume that a high variance means uncertainty. This form of weight estimation requires that for at least some of the genomes, a sufficient number of sequences must be available to estimate the standard deviation. In some instances, it might be necessary to split long contigs into smaller sequences to generate a sufficient number of data points for estimation.

Parameter β in Equation 4.11 is only relevant for soft classification but not in the context of ML classification or p-values. It can best be viewed as a sharpening or smoothing parameter of the bin posterior distribution (the probability of a genome or bin given the contig). β is estimated by minimization of the training or test error, as in our simulation.

4.3.8 Data simulation

We simulated reads of a complex microbial community from 400 publicly available genomes (Supplementary Methods and Supplementary Table 1). These comprised 295 unique and 44 species with each two or three strain genomes to mimic strain heterogeneity. Our aim was to create a difficult benchmark dataset under controlled settings, minimizing potential biases introduced by specific software.

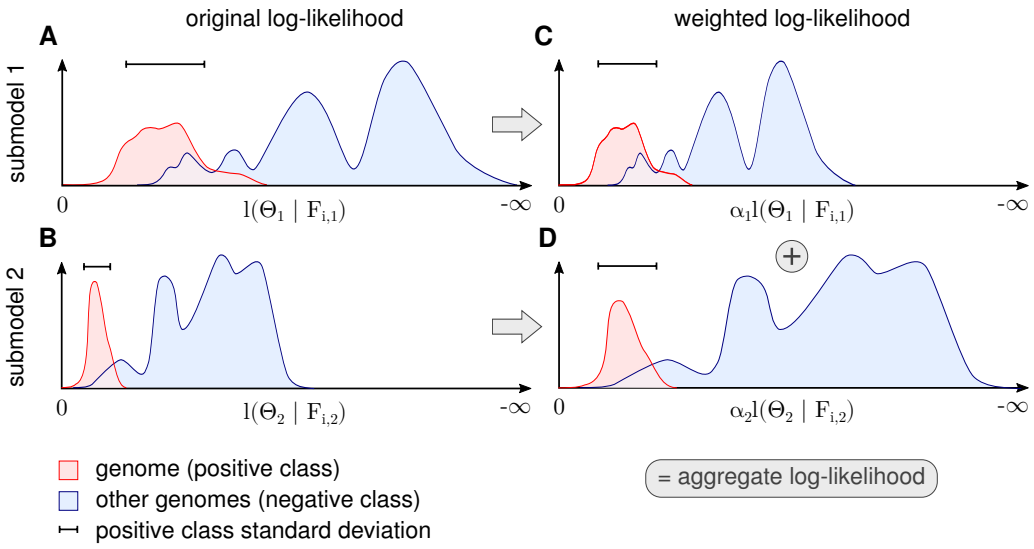


Figure 4.3: Procedure for determination of for each submodel. The figure shows a schematic for a single genome and two submodels. The genome's contig log-likelihood distribution (A and B) is scaled to a standard deviation of one (C and D) before adding the term in the aggregate model in .

We sampled abundances from a lognormal distribution because it has been described as a realistic model (Schloss & Handelsman, 2006). We then simulated a primary community which was then subject to environmental changes resulting in exponential growth of 25% of the community members at growth rates which where chosen uniformly at random between one and ten whereas the other genome abundances remained unchanged. We applied this procedure three times to the primary community which resulted in one primary and three secondary artificial community abundances profiles. With these, we generated 150 bp long Illumina HiSeq reads using the ART simulator (Huang et al., 2012) and chose a yield of 15 Gb per sample. The exact amount of read data for all four samples after simulation was 59.47 Gb. To avoid any bias caused by specific metagenome assembly software and to assure a constant contig length, we divided the original genome sequences into non-overlapping artificial contigs of 1 kb length and selected a random 500 kb of each genome to which we mapped the simulated reads using Bowtie2 (Langmead & Salzberg, 2012). By the exclusion of some genome

reference, we imitated incomplete genome assemblies when mapping reads, which affects the coverage values. Finally, we subsampled 300 kb contigs per genome with non-zero read coverage in at least one of the samples to form the demonstration dataset (120 Mb), which has 400 genomes (including related strains), four samples and contigs of size 1 kb. Due to the short contigs and few samples, this is a challenging dataset for complete genome recovery (Nielsen et al., 2014) but suitable to demonstrate the functioning of our model with limited data. For each contig we derived 5-mer frequencies, taxonomic annotation (removing species-level genomes from the reference sequence data) and average read coverage per sample, as described in the Supplementary Methods.

4.4 Results

4.4.1 Maximum likelihood classification

We evaluated the performance of the model when classifying contigs to the genome with the highest likelihood, a procedure called Maximum Likelihood (ML) classification. We applied a form of three-fold cross-validation, dividing the simulated data set into three equally-sized parts with 100 kb from every genome. We used only 100 kb (training data) of every genome to infer the model parameters and the other 200 kb (test data) to measure the classification error. 100 kb was used for training because it is often difficult to identify sufficient training data in metagenome analysis. For each combination of submodels, we calculated the mean squared error (MSE) and mean pairwise coclustering (MPC) probability for the predicted (ML) probability matrices (Suppl. Methods), averaged over the three test data partitions. We included the MPC as it can easily be interpreted: for instance, a value of 0.5 indicates that on average 50% of all contig pairs of a genome end up in the same bin after classification. Table 4.2 shows that the model integrates information from each data source such that the inclusion of additional submodels resulted in a better MPC and also MSE, with a single exception when combining absolute and relative abdundance models which resulted in a marginal increase of the MSE. We also found that taxonomic annotation rep-

resents the most powerful information type in our simulation. For comparson, we added scores for NBC (Rosen, Reichenberger & Rosenfeld, 2011), a classifier based on nucleotide composition with in-sample training using 5-mers and 15-mers, and Centrifuge (Kim et al., 2016), a similarity-based classifier both with in-sample and reference data. These programs were given the same information as the corresponding submodels and they rank close to these. In a further step, we investigated how the presence of very similar genomes impacted the performance of the model. We first collapsed strains from the same species by merging the corresponding columns in the classification likelihood matrix, retaining the entry with the highest likelihood, and then computed the resulting coclustering performance increase $\Delta\text{MPC}_{\text{ML}}$. Considering assignment on species instead of strain level showed a larger $\Delta\text{MPC}_{\text{ML}}$ for nucleotide composition and taxonomic annotation than for absolute and relative abundance. This is expected, because both do not distinguish among strains, whereas genome abundance does in some, but not all cases.

Table 4.2: Cross-validation performance of ML classification for all possible combinations of submodels. We calculated the mean pairwise coclustering (MPC), the strain to species MPC improvement ($\Delta\text{MPC}_{\text{ML}}$) and the mean squared error (MSE). AbAb = absolute total abundance; ReAb = relative abundance; NuCo = nucleotide composition; TaAn = taxonomic annotation. NBC (v1.1) and Centrifuge (v.1.0.3b) are external classifiers added for comparison. Best values are in bold and worst in italic.

Submodels	MPC_{ML}	$\Delta\text{MPC}_{\text{ML}}$	MSE_{ML}
Centrifuge (in-sample)	*0.01*	+0.01	0.51
NBC (15-mers)	0.02	*+0.00*	*0.66*
AbAb	0.03	*+0.00*	0.58
ReAb	0.08	+0.02	0.61
Centrifuge (reference)	0.13	+0.03	0.45
AbAb + ReAb	0.21	+0.04	0.59
NuCo	0.30	+0.06	0.52
NBC (5-mers)	0.34	+0.06	0.48
ReAb + NuCo	0.41	+0.07	0.48
AbAb + NuCo	0.43	+0.08	0.50

Submodels	MPC_{ML}	ΔMPC_{ML}	MSE_{ML}
TaAn	0.46	+0.09	0.41
AbAb + ReAb + NuCo	0.52	+0.09	0.44
NuCo + TaAn	0.52	+0.09	0.40
AbAb + TaAn	0.54	+0.09	0.39
AbAb + NuCo + TaAn	0.60	+0.10	0.37
ReAb + TaAn	0.60	+0.10	0.36
ReAb + NuCo + TaAn	0.64	**+0.11**	0.34
AbAb + ReAb + TaAn	0.65	+0.10	0.35
AbAb + ReAb + NuCo + TaAn	**0.68**	**+0.11**	**0.33**

4.4.2 Soft assignment

The contig length of 1 kb in our simulation is considerably shorter, and therefore harder to classify, than sequences which can be produced by current assembly methods or by some cutting-edge sequencing platforms (Goodwin, McPherson & McCombie, 2016). In practice, longer contigs can be classified with higher accuracy than short ones, as more information is provided as a basis for assignment. For instance, a more robust coverage mean, a k-mer spectrum derived from more counts or more local alignments to reference genomes can be inferred from longer sequences. However, as short contigs remain frequent in current metagenome assemblies, 1 kb is sometimes considered a minimum useful contig length (Alneberg et al., 2014). To account for the natural uncertainty when assigning short contigs, one can calculate the posterior probabilities over the genomes (see Suppl. Methods), which results in partial assignments of each contig to the genomes. This can reflect situations in which a particular contig is associated with multiple genomes, for instance in case of misassemblies or the presence of homologous regions across genomes.

The free model parameter β in Equation 4.1, which is identical in all genome models, smoothens or sharpens the posterior distribution: $\beta = 0$ produces a uniform posterior and with very high β, the posterior approaches the sharp ML solution.

We determined β by optimizing the MSE on both training and test data, shown in Figure 4.4. As expected, the classification training error was smaller than the test error because the submodel parameters were optimized with respect to the training data. Because the minima are close to each other, the full aggregate model seems robust to overfitting of β on training data. The comparison of soft vs. hard assignment shows that the former has a smaller average test classification MSE of ~ 0.28 (the illustrated minimum in Figure 4.4) compared to the latter (ML) assignment MSE of ~ 0.33 in Table 4.2. Thus, soft assignment seems more suitable to classify 1 kb contigs, which tend to produce similar likelihoods under more than one genome model.

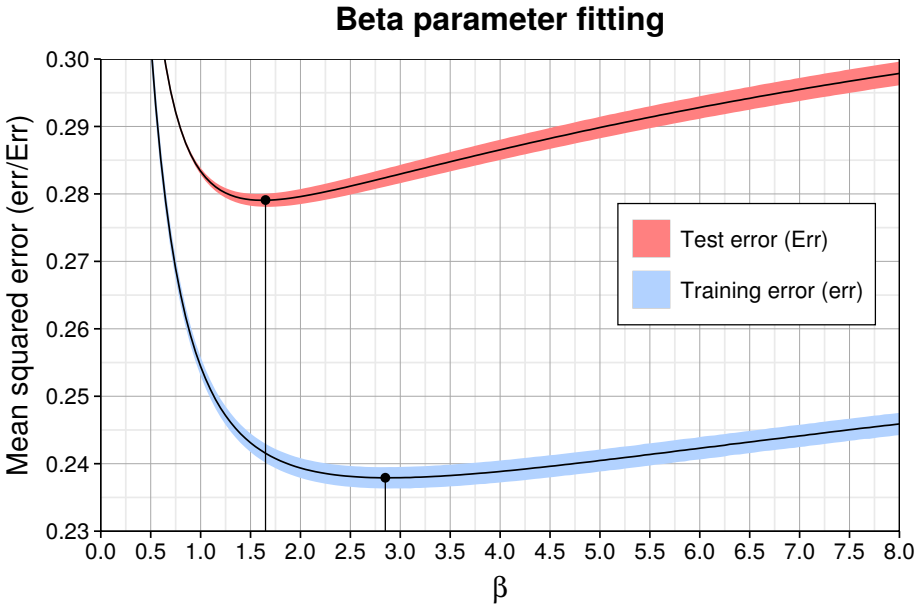


Figure 4.4: Model training (err) and test error (Err) as a function of β for the complete aggregate model including all submodels and feature types. The solid curve shows the average and the colored shading the standard deviation of the three partitions in cross-validation. The corresponding optimal values for β are marked by black dots and vertical lines. The minimum average training error is 0.238 ($\beta = 2.85$) and test error is 0.279 at $\beta = 1.65$.

4.4.3 Genome enrichment

Enrichment is commonly known as an experimental technique to increase the concentration of a target substance relative to others in a probe. Thus, an enriched metagenome still contains a mixture of different genomes, but the target genome will be present at much higher frequency than before. This allows a more focused analysis of the contigs or an application of methods which seem prohibitive for the full data by runtime or memory considerations. In the following, we demonstrate how to filter metagenome contigs by p-value to enrich *in-silico* for specific genomes. Often, classifiers model an exhaustive list of alternative genomes but in practice it is difficult to recognize all species or strains in a metagenome with appropriate training data. When we only look at individual likelihoods, for instance the maximum among the genomes, this can be misleading if the contig comes from a missing genome. For better judgment, a p-value tells us how frequent or extreme the actual likelihood is for each genome. Many if not all binning methods lack explicit significance calculations. We can take advantage of the fact that the classification model compresses all features into a genome likelihood and generate a null (log-)likelihood distribution on training data for each genome. Therefore, we can associate empirical p-values with each newly classified contig and can, for sufficiently small p-values, reject the null hypothesis that the contig belongs to the respective genome. Since this is a form of binary classification, there is the risk to reject a good contig which we measure as sensitivity.

We enriched a metagenome by first training a genome model and then calculating the p-values of remaining contigs using this model. Contigs with higher p-values than the chosen critical value were discarded. The higher this cutoff is, the smaller the enriched sample becomes, but also the target genome will be less complete. We calculated the reduced sample size as a function of the p-value cutoff for our simulation (Figure 4.5). Selecting a p-value threshold of 2.5% shrinks the test data on average down to 5% of the original size. Instead of an empirical p-value, we could also use a parametrized distribution or select a critical log-likelihood value by manual inspection of the log-likelihood distribution (see Figure 4.3 for an example of such a distribution). This example shows that generally a large part of a metagenome dataset can be discarded while retaining most of the target

genome sequence data.

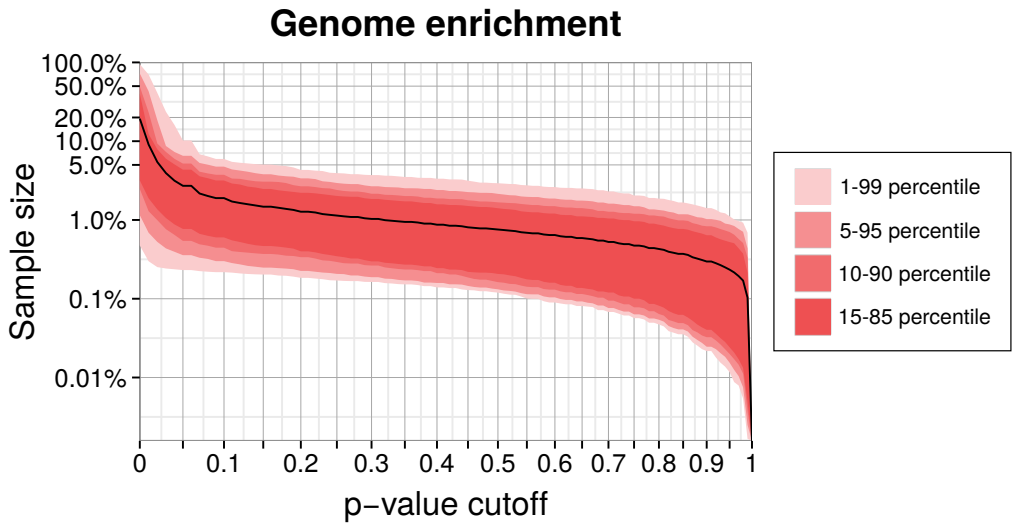


Figure 4.5: Genome enrichment for 400 genomes with three-fold cross-validation. For each genome, we measured the test sample size relative to the full dataset after filtering by a p-value cutoff and summing over the three data partitions. The solid line shows the resulting average sample size over all 400 genomes. The variability between genomes is shown as quantiles in red. Both axes are logarithmic to show the relevant details for lower p-values cutoffs. The corresponding sensitivity, shown in Suppl. Figure 1, is approximately a linear function of the p-value.

4.4.4 Bin analysis

The model can be used to analyze bins of metagenome contigs, regardless of the method that was used to infer these bins. Specifically, one can measure the similarity of two bins in terms of the contig likelihood instead of, for instance, an average euklidean distance based on the contig or genome k-mer and abundance vectors. We compare bins to investigate the relation between the given data, represented by the features in the model, and their grouping into genome bins. For instance, one could ask whether the creation of two genome bins is sufficiently backed up by the contig data or whether they should be merged into a single bin.

For readability, we write the likelihood of a contig in bin A to:

$$L(\theta_A \mid \text{contig } i) = L_i(\theta_A) = L(\theta_A) = L_A$$

To compare two specific bins, we select the corresponding pair of columns in the classification likelihood matrix and calculate two mixture likelihoods for each contig (rows), $\hat{L}$, using the MLE of the parameters for both bins and L_{swap} under the hypothesis that we swap the model parameters of both bins. The partial assignment weights $\hat{\pi}_A$ and $\hat{\pi}_B$, called responsibilities, are estimated by normalization of the two bin likelihoods.

$$\hat{L} = \hat{\pi}_A\, L_A + \hat{\pi}_B\, L_B = \left(\tfrac{L_A}{L_A+L_B}\right) L_A + \left(\tfrac{L_B}{L_A+L_B}\right) L_B = \frac{L_A^2 + L_B^2}{L_A + L_B} \tag{4.12}$$

$$L_{swap} = \hat{\pi}_A\, L_B + \hat{\pi}_B\, L_A = \left(\tfrac{L_A}{L_A+L_B}\right) L_B + \left(\tfrac{L_B}{L_A+L_B}\right) L_A = \frac{2L_AL_B}{L_A + L_B} \tag{4.13}$$

For example, if $\hat{\pi}_A$ and $\hat{\pi}_B$ assign one third of a contig to the first, less likely bin and two thirds to the second, more likely bin using the optimal parameters, then L_{swap} would simply exchange the contributions in the mixture likelihood so that one third are assigned to the more likely and two thirds to the less likely bin. The ratio $L_{swap}/\hat{L}$ ranges from zero to one and can be seen as a percentage similarity. We form a joint relative likelihood for all N contigs, weighting each contig by its optimal mixture likelihood $\hat{L}$ and normalizing over these likelihood values.

$$\mathrm{S}(A,B) = \sqrt[Z]{\prod_{i=1}^{N} \left(\frac{2\, L_i(\theta_A)\, L_i(\theta_B)}{L_i^2(\theta_A) + L_i^2(\theta_B)} \right)^{\frac{L_i^2(\theta_A)+L_i^2(\theta_B)}{L_i(\theta_A)+L_i(\theta_B)}}} \tag{4.14}$$

normalized by the total joint mixture likelihood

$$Z = \sum_{i=1}^{N} \frac{L_i^2(\theta_A) + L_i^2(\theta_B)}{L_i(\theta_A) + L_i(\theta_B)} \tag{4.15}$$

The quantity in Equation 4.14 ranges from zero to one, reaching one when the two bin models produce identical likelihood values. We can therefore interpret the ratio as a percentage similarity between any two bins. A connection to the Kullback-Leibler divergence can be constructed (Supplementary Methods).

To demonstrate the application, we trained the model on our simulated genomes, assuming they were bins, and created trees (Figure 4.6) for a randomly drawn subset of 50 of the 400 genomes using the probabilistic bin distances $-log(S)$ (Equation 4.14). We computed the distances twice, first with only nucleotide composition and taxonomic annotation submodels and second with the full feature set to compare the bin resolution. The submodel parameters were inferred using the full dataset and β using three-fold crossvalidation. We then applied average linkage clustering to build balanced and rooted trees with equal distance from leave to root for visual inspection. The first tree loosely reflects phylogenetic structure corresponding to the input features. However, many similarities over 50% (outermost ring) show that model and data lack the support for separating these bins. In contrast, the fully informed tree, which additionally includes information about contig coverages, separates the genomes bins, such that only closely related strains remain ambiguous. This analysis shows again that the use of additional features improves the resolution of individual genomes and, specifically, that abundance separates similar genomes. Most importantly, we show that our model provides a measure of support for a genome binning. We know the taxa of the genome bins in this example but for real metagenomes, such an analysis can reveal binning problems and help to refine the bins as in Figure 4.1d.

4.4.5 Genome bin refinement

We applied the model to show one of its current use cases on more realistic data. We downloaded the medium complexity dataset from www.cami-challenge.org. This dataset is quite complex (232 genomes, two sample replicates). We also retrieved the results of two highest-performing automatic binning programs, MaxBin and Metawatt, in the CAMI challenge evaluation (Sczyrba et al., 2017). We took the simplest possible approach: we trained MLGEX on the genome bins

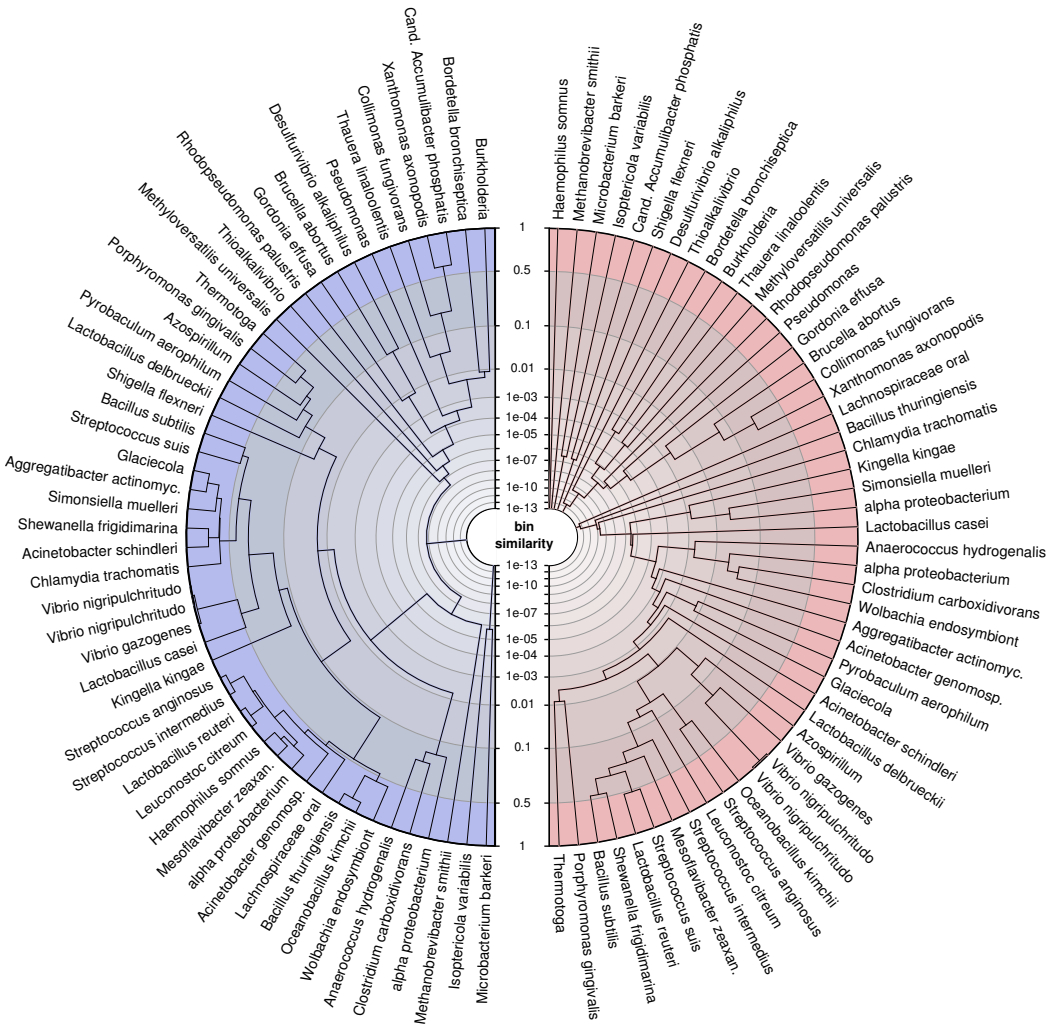


Figure 4.6: Average linkage clustering of a random subset of 50 out of 400 genomes using probabilistic distances $-log(S)$ to demonstrate the ability of the model to measure bin resolution. This example compares the left (blue) tree, which was constructed only with nucleotide composition and taxonomic annotations, with the right (red) tree, which uses all available features. The tip labels were shortened to fit into the figure. The similarity axis is scaled as *log(1-log(S))* to focus on values near one. Bins which are more than 50% similar branch in the outermost ring whereas highly dissimilar bins branch close to the center. We created the trees by applying the R function *hclust(method="average")* to MGLEX output.

derived by these methods and classified the contigs to the bins with the highest likelihood, thus ignoring all details of contig splitting, β or p-value calculation

and changes in the number of genome bins. When contigs were assigned to multiple bins with equal probability, we attributed them to the first bin in the list because the evaluation framework does not allow sharing contigs between bins. We only used information provided to the contestants by the time of the challenge in the process. We report the results for two settings for each method using the recall, the fraction of overall assigned contigs (bp), and the Adjusted Rand index (ARI) as defined in the CAMI evaluation paper. In the first, we swapped contigs which were originially assigned between bins. In the second, all available contigs were assigned to the bins, thus maximizing the recall. Table 4.3 shows that MGLEX bin refinement improved the genome bins in terms of the ARI for both sets of genome bins and increased the recall for Metawatt but not MaxBin. This is likely due to the fact that MaxBin has fewer but relatively complete bins to which the other contigs cannot correctly be recruited. Further improvement would involve disection and merging of bins within and among methods, for which MGLEX likelihoods can be considered.

Table 4.3: Genome bin refinement for CAMI medium complexity dataset with 232 genomes and two samples. The recall is the fraction of overall assigned contigs (bp). The Adjusted Rand index (ARI) is a measure of binning precision. The unmodified genome bins are the submissions to the CAMI challenge using the corresponding unsupervised binning methods Metawatt and MaxBin. MGLEX swapped contigs: contigs in original genome bins reassigned to the bin with highest MGLEX likelihood. MGLEX all contigs: all contigs (with originally uncontained) assigned to the bin with highest MGLEX likelihood. The lowest scores are written in italic and highest in bold.

Binner	Variant	Bin count	Recall (bp)	ARI
Metawatt	unmodified	285	*0.94*	*0.75*
Metawatt	MGLEX swapped contigs	285	*0.94*	**0.82**
Metawatt	MGLEX all contigs	285	**1.00**	0.77
MaxBin	unmodified	125	*0.82*	0.90
MaxBin	MGLEX swapped contigs	125	*0.82*	**0.92**
MaxBin	MGLEX all contigs	125	**1.00**	*0.76*

4.4.6 Implementation

We provide a Python package called MGLEX, which includes the described model. Simple text input facilitates the integration of external programs for feature extraction like k-mer counting or read mapping, which are not included. MGLEX can process millions of sequences with vectorized arithmetics using NumPy (Walt, Colbert & Varoquaux, 2011) and includes a command line interface to the main functionality, such as model training, classification, p-value and error calculations. It is open source (GPLv3) and freely available via the Python Package Index[1] and on GitHub[2].

4.5 Discussion

We describe an aggregate likelihood model for the reconstruction of genome bins from metagenome data sets and show its value for several applications. The model can learn from and classify nucleotide sequences from metagenomes. It provides likelihoods and posterior bin probabilities for existing genome bins, as well as p-values, which can be used to enrich a metagenome dataset with a target genome. The model can also be used to quantify bin similarity. It builds on four different submodels that make use of different information sources in metagenomics, namely contig coverage, nucleotide composition and previous taxonomic assignments. By its modular design, the model can easily be extended to include additional information sources. This modularity also helps in interpretation and computations. The former, because different features can be analyzed separately and the latter, because submodels can be trained independently and in parallel.

In comparison to previously described parametric binning methods, our model incorporates two new types of features. The first is relative differential coverage, for which, to our knowledge, this is the first attempt to use binomials to account for systematic bias in the read mapping for different genome regions. As such, the binomial submodel represents the parametric equivalent of covariance distance

[1] https://pypi.python.org/pypi/mglex/
[2] https://www.github.com/hzi-bifo/mglex/

clustering. The second new type is taxonomic annotation, which substantially improved the classification results in our simulation. Taxonomic annotations, as used in the model and in our simulation, were not correct up to the species level and need not be, as seen in the classification results. We only require the same annotation method be applied to all sequences. In comparison to previous methods, our aggregate model has weight parameters to combine the different feature types and allows tuning the bin posterior distribution by selection of an optimal smoothing parameter β.

We showed that probabilistic models represent a good choice to handle metagenomes with short contigs or few sample replicates, because they make soft, not hard decisions, and because they can be applied in numerous ways. When the individual submodels are trained, genome bin properties are compressed into fewer model parameters, such as mean values, which are mostly robust to outliers and therefore tolerate a certain fraction of bin pollution. This property allows to reassign contigs to bins, which we demonstrated in the "Genome bin refinement" section. Measuring the performance of the individual submodels and their corresponding features on short simulated contigs (Table 4.2), we find that they discriminate genomes or species pan-genomes by varying degrees. Genome abundance represents, in our simulation with four samples, the weakest single feature type, which will likely become more powerful with increasing sample numbers. Notably, genomes of individual strains are more difficult to distinguish than species level pangenomes using any of the features. In practice, if not using idealized assemblies as in our current evaluation, strain resolution poses a problem to metagenome assembly, which is currently not resolved in a satisfactory manner (Sczyrba et al., 2017).

The current MGLEX model is somewhat crude because it makes many simplifying assumptions in the submodel definitions. For instance, the multi-layer model for taxonomic annotation assumes that the probabilities in different layers are independent, the series of binomials for relative abundance should be replaced by a multinomial to accout for the parameter dependencies or the absolute abdundance Poisson model should incorporate overdispersion to model the data more appropriately. Exploiting this room for improvement can lead to further im-

provement in the performance while the overall framework and usage of MGLEX stays unchanged. When we devised our model, we had an embedding into more complex routines in mind. In the future, the model can be used in inference procedures such as EM or MCMC to infer or improve an existing genome binning. Thus, MGLEX provides a software package for use in other programs. However, it also represents a powerful stand-alone tool for the adept user in its current form.

Currently, MGLEX does not yet have support for multiple processors and only provides the basic functionality presented here. However, training and classification can easily be implemented in parallel because they are expressed as matrix multiplications. The model requires sufficient training data to robustly estimate the submodel weights $\boldsymbol{\alpha}$ using the standard deviation of the empirical log-likelihood distributions and requires linked sequences to estimate β using error minimization. In situations with a limited number of contigs per genome bin, we therefore advise to generate linked training sequences of a certain length, as in our simulation, for instance by splitting assembled contigs. The optimal length for splitting may depend on the overall fragmentation of the metagenome.

Our open-source Python package MGLEX provides a flexible framework for metagenome analysis and binning which we intent to develop further together with the metagenomics research community. It can be used as a library to write new binning applications or to implement custom workflows, for example to supplement existing binning strategies. It can build upon a present metagenome binning by taking assignments to bins as input and deriving likelihoods and p-values that allow for critical inspection of the contig assignments. Based on the likelihood, MGLEX can calculate bin similarities to provide insight into the structure of data and community. Finally, genome enrichment of metagenomes can improve the recovery of particular genomes in large datasets.

4.6 Acknowledgments

We thank S. Reimering, A. Weimann and A. Bremges for proofreading and constructive feedback.

Chapter 5

References

Aguiar-Pulido V., Huang W., Suarez-Ulloa V., Cickovski T., Mathee K., Narasimhan G. 2016. Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis. *Evolutionary Bioinformatics Online* 12:5–16. DOI: 10.4137/EBO.S36436.

Albertsen M., Hugenholtz P., Skarshewski A., Nielsen K are L., Tyson GW., Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* 31:533–8. DOI: 10.1038/nbt.2579.

Alneberg J., Bjarnason BS., de Bruijn I., Schirmer M., Quick J., Ijaz UZ., Lahti L., Loman NJ., Andersson AF., Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 11:1144–1146. DOI: 10.1038/nmeth.3103.

Baran Y., Halperin E. 2012. Joint analysis of multiple metagenomic samples. *PLoS computational biology* 8:e1002373. DOI: 10.1371/journal.pcbi.1002373.

Berger SA., Krompass D., Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology* 60:291–302. DOI: 10.1093/sysbio/syr010.

Berry D., Widder S. 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* 5.

DOI: 10.3389/fmicb.2014.00219.

Brady A., Salzberg SL. 2009. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature methods* 6:673–6. DOI: 10.1038/nmeth.1358.

Brady A., Salzberg S. 2011. PhymmBL expanded: Confidence scores, custom databases, parallelization and more. *Nature methods* 8:367. DOI: 10.1038/nmeth0511-367.

Bremges A., Singer E., Woyke T., Sczyrba A. 2016. MeCorS: Metagenome-enabled error correction of single cell sequencing reads. *Bioinformatics* 32:2199–2201. DOI: 10.1093/bioinformatics/btw144.

Buchfink B., Xie C., Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12:59–60. DOI: 10.1038/nmeth.3176.

Bulgarelli D., Garrido-Oter R., Münch PC., Weiman A., Dröge J., Pan Y., McHardy AC., Schulze-Lefert P. 2015. Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley. *Cell Host & Microbe* 17:392–403. DOI: 10.1016/j.chom.2015.01.011.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: Architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

Carr R., Shen-Orr SS., Borenstein E. 2013. Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS computational biology* 9:e1003292. DOI: 10.1371/journal.pcbi.1003292.

Chatterji S., Yamazaki I., Bai Z., Eisen JA. 2008. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In: *Annual International Conference on Research in Computational Molecular Biology*. Springer, 17–28.

Cuvelier ML., Allen AE., Monier A., McCrow JP., Messié M., Tringe SG., Woyke T., Welsh RM., Ishoey T., Lee J-H., Binder BJ., DuPont CL., Latasa M., Guigand C., Buck KR., Hilton J., Thiagarajan M., Caler E., Read B., Lasken RS., Chavez FP., Worden AZ. 2010. Tar-

geted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences* 107:14679–14684. DOI: 10.1073/pnas.1001665107.

Darling AE., Jospin G., Lowe E., Matsen F a., Bik HM., Eisen J a. 2014. PhyloSift: Phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. DOI: 10.7717/peerj.243.

Dohm JC., Lottaz C., Borodina T., Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* 36:e105. DOI: 10.1093/nar/gkn425.

Dong X., Dröge J., von Toerne C., Marozava S., McHardy AC., Meckenstock RU. 2017. Reconstructing metabolic pathways of a member of the genus Pelotomaculum suggesting its potential to oxidize benzene to carbon dioxide with direct reduction of sulfate. *FEMS Microbiology Ecology* 93. DOI: 10.1093/femsec/fiw254.

Dröge J., McHardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics* 13:646–655. DOI: 10.1093/bib/bbs031.

Dröge J., Gregor I., McHardy AC. 2014. Taxator-tk: Precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics (Oxford, England)*:1–8. DOI: 10.1093/bioinformatics/btu745.

Dröge J., Schönhuth A., McHardy AC. 2017. A probabilistic model to recover individual genomes from metagenomes. *PeerJ Computer Science* 3:e117. DOI: 10.7717/peerj-cs.117.

Eloe-Fadrosh EA., Ivanova NN., Woyke T., Kyrpides NC. 2016. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology* 1:15032. DOI: 10.1038/nmicrobiol.2015.32.

Eren AM., Esen ÖC., Quince C., Vineis JH., Morrison HG., Sogin ML., Delmont TO. 2015. Anvi'o: An advanced analysis and visualization

platform for 'omics data. *PeerJ* 3:e1319. DOI: 10.7717/peerj.1319.

Frith MC., Hamada M., Horton P. 2010. Parameters for accurate genome alignment. *BMC bioinformatics* 11:80. DOI: 10.1186/1471-2105-11-80.

Fuhrman JA., Cram JA., Needham DM. 2015. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology* 13:133–146. DOI: 10.1038/nrmicro3417.

Garrett RA., Prangishvili D., Shah SA., Reuter M., Stetter KO., Peng X. 2010. Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. *Environmental Microbiology* 12:2918–2930. DOI: 10.1111/j.1462-2920.2010.02266.x.

Gawad C., Koh W., Quake SR. 2016. Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics* 17:175–188. DOI: 10.1038/nrg.2015.16.

Gerlach W., Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids research*:1–11. DOI: 10.1093/nar/gkr225.

Ghurye JS., Cepeda-Espinoza V., Pop M. 2016. Metagenomic Assembly: Overview, Challenges and Applications. *The Yale Journal of Biology and Medicine* 89:353–362.

Gillespie DE., Brady SF., Bettermann AD., Cianciotto NP., Liles MR., Rondon MR., Clardy J., Goodman RM., Handelsman J. 2002. Isolation of Antibiotics Turbomycin A and B from a Metagenomic Library of Soil Microbial DNA. *Applied and Environmental Microbiology* 68:4301–4306. DOI: 10.1128/AEM.68.9.4301-4306.2002.

Goodwin S., McPherson JD., McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17:333–351. DOI: 10.1038/nrg.2016.49.

Gregor I., Dröge J., Schirmer M., Quince C., McHardy AC. 2014. PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking

taxonomic bins from metagenomes. *arxiv.org*:1–67.

Gregor I., Dröge J., Schirmer M., Quince C., McHardy AC. 2016. *PhyloPythiaS+*: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* 4:e1603. DOI: 10.7717/peerj.1603.

Hagen LH., Frank JA., Zamanzadeh M., Eijsink VGH., Pope PB., Horn SJ., Arntzen MØ. 2016. Quantitative metaproteomics highlight the metabolic contributions of uncultured phylotypes in a thermophilic anaerobic digester. *Applied and Environmental Microbiology*:AEM.01955–16. DOI: 10.1128/AEM.01955-16.

Hamady M., Knight R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research* 19:1141–52. DOI: 10.1101/gr.085464.108.

Handelsman J. 2004. Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews : MMBR* 68:669–85. DOI: 10.1128/MMBR.68.4.669-685.2004.

Hastie T., Tibshirani R., Friedman J. 2001. *The Elements of Statistical Learning*. Springer New York Inc.

Hauswedell H., Singer J., Reinert K. 2014. Lambda: The local aligner for massive biological data. *Bioinformatics* 30:i349–i355. DOI: 10.1093/bioinformatics/btu439.

Hess M., Sczyrba A., Egan R., Kim T-W., Chokhawala H., Schroth G., Luo S., Clark DS., Chen F., Zhang T., Mackie RI., Pennacchio L a., Tringe SG., Visel A., Woyke T., Wang Z., Rubin EM. 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N.Y.)* 331:463–7. DOI: 10.1126/science.1200387.

Hu X., Yuan J., Shi Y., Lu J., Liu B., Li Z., Chen Y., Mu D., Zhang H., Li N., Yue Z., Bai F., Li H., Fan W. 2012. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics (Oxford, England)* 28:1533–5. DOI:

10.1093/bioinformatics/bts187.

Huang X., Madan A. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Research* 9:868–877. DOI: 10.1101/gr.9.9.868.

Huang W., Li L., Myers JR., Marth GT. 2012. ART: A next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* 28:593–4. DOI: 10.1093/bioinformatics/btr708.

Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome biology* 3:REVIEWS0003.

Hugenholtz P., Tyson GW. 2008. Microbiology: Metagenomics. *Nature* 455:481–483. DOI: 10.1038/455481a.

Huson DH., Xie C. 2014. A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA. *Bioinformatics (Oxford, England)* 30:38–9. DOI: 10.1093/bioinformatics/btt254.

Huson DH., Mitra S., Ruscheweyh H-J., Weber N., Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome research* 21:1552–60. DOI: 10.1101/gr.120618.111.

Imelfort M., Parks D., Woodcroft BJ., Dennis P., Hugenholtz P., Tyson GW. 2014. GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2:e603. DOI: 10.7717/peerj.603.

Iverson V., Morris RM., Frazar CD., Berthiaume CT., Morales RL., Armbrust EV. 2012. Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science* 335:587–590. DOI: 10.1126/science.1212665.

Kang DD., Froula J., Egan R., Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. DOI: 10.7717/peerj.1165.

Karlin S., Mrazek J., Campbell AM. 1997. Compositional biases of bacterial genomes and evolutionary implications. *Journal of bacteriology* 179:3899–

3913.

Kim D., Song L., Breitwieser FP., Salzberg SL. 2016. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research* 26:1721–1729. DOI: 10.1101/gr.210641.116.

Kislyuk A., Bhatnagar S., Dushoff J., Weitz JS. 2009. Unsupervised statistical clustering of environmental shotgun sequences. *BMC bioinformatics* 10:316. DOI: 10.1186/1471-2105-10-316.

Klumpp J., Fouts DE., Sozhamannan S. 2012. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage* 2:190–199. DOI: 10.4161/bact.22111.

Koslicki D., Foucart S., Rosen G. 2013. Quikr: A method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics (Oxford, England)* 29:2096–102. DOI: 10.1093/bioinformatics/btt336.

Kunin V., Engelbrektson A., Ochman H., Hugenholtz P. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12:118–123. DOI: 10.1111/j.1462-2920.2009.02051.x.

Lander ES., Waterman MS. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* 2:231–239. DOI: 10.1016/0888-7543(88)90007-9.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359. DOI: 10.1038/nmeth.1923.

Lasken RS., McLean JS. 2014. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics* 15:577–584. DOI: 10.1038/nrg3785.

Lin H-H., Liao Y-C. 2016. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientific Reports* 6. DOI: 10.1038/srep24175.

Lindner MS., Renard BY. 2013. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic acids research* 41:e10. DOI:

10.1093/nar/gks803.

Lu YY., Chen T., Fuhrman JA., Sun F. 2016. COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment, and paired-end read LinkAge. *Bioinformatics*:btw290.

Luo R., Liu B., Xie Y., Li Z., Huang W., Yuan J., He G., Chen Y., Pan Q., Liu Y., Tang J., Wu G., Zhang H., Shi Y., Liu Y., Yu C., Wang B., Lu Y., Han C., Cheung DW., Yiu S-M., Peng S., Xiaoqian Z., Liu G., Liao X., Li Y., Yang H., Wang J., Lam T-W., Wang J. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18. DOI: 10.1186/2047-217X-1-18.

Luo Y., Zeng J., Berger B., Peng J. 2016. Low-density locality-sensitive hashing boosts metagenomic binning.

Matsen FA., Kodner RB., Armbrust EV. 2010. Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* 11:538. DOI: 10.1186/1471-2105-11-538.

Mavromatis K., Ivanova N., Barry K., Shapiro H., Goltsman E., McHardy AC., Rigoutsos I., Salamov A., Korzeniewski F., Land M., Lapidus A., Grigoriev I., Richardson P., Hugenholtz P., Kyrpides NC. 2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods* 4:495–500. DOI: 10.1038/nmeth1043.

McHardy AC., Martín HG., Tsirigos A., Hugenholtz P., Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods* 4:63–72. DOI: 10.1038/nmeth976.

Melsted P., Pritchard JK. 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* 12:333. DOI: 10.1186/1471-2105-12-333.

Mende DR., Aylward FO., Eppley JM., Nielsen TN., DeLong EF. 2016. Improved Environmental Genomes via Integration of Metagenomic and Single-

Cell Assemblies. *Frontiers in Microbiology* 7. DOI: 10.3389/fmicb.2016.00143.

Miller JR., Koren S., Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–27. DOI: 10.1016/j.ygeno.2010.03.001.

Monzoorul Haque M., Ghosh TS., Komanduri D., Mande SS. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics (Oxford, England)* 25:1722–30. DOI: 10.1093/bioinformatics/btp317.

Nielsen HB., Almeida M., Juncker AS., Rasmussen S., Li J., Sunagawa S., Plichta DR., Gautier L., Pedersen AG., Le Chatelier E., Pelletier E., Bonde I., Nielsen T., Manichanh C., Arumugam M., Batto J-M., Quintanilha dos Santos MB., Blom N., Borruel N., Burgdorf KS., Boumezbeur F., Casellas F., Doré J., Dworzynski P., Guarner F., Hansen T., Hildebrand F., Kaas RS., Kennedy S., Kristiansen K., Kultima JR., Léonard P., Levenez F., Lund O., Moumen B., Le Paslier D., Pons N., Pedersen O., Prifti E., Qin J., Raes J., Sørensen S., Tap J., Tims S., Ussery DW., Yamada T., MetaHIT Consortium., Renault P., Sicheritz-Ponten T., Bork P., Wang J., Brunak S., Ehrlich SD. 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* 32:822–828. DOI: 10.1038/nbt.2939.

Patil KR., Haider P., Pope PB., Turnbaugh PJ., Morrison M., Scheffer T., McHardy AC. 2011. Taxonomic metagenome sequence assignment with structured output models. *Nature Methods* 8:191–192. DOI: 10.1038/nmeth0311-191.

Pell J., Hintze A., Canino-Koning R., Howe A., Tiedje JM., Brown CT. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences* 109:13272–13277.

Ponomarova O., Patil KR. 2015. Metabolic interactions in microbial communities: Untangling the Gordian knot. *Current Opinion in Microbiology*

27:37–44. DOI: 10.1016/j.mib.2015.06.014.

Pope PB., Smith W., Denman SE., Tringe SG., Barry K., Hugenholtz P., McSweeney CS., McHardy a C., Morrison M. 2011. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science (New York, N.Y.)* 333:646–8. DOI: 10.1126/science.1205760.

Price ND., Reed JL., Palsson BØ. 2004. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nature Reviews Microbiology* 2:886–897. DOI: 10.1038/nrmicro1023.

Przyborowski J., Wilenski H. 1940. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* 31:313. DOI: 10.2307/2332612.

Qin J., Li R., Raes J., Arumugam M., Burgdorf KS., Manichanh C., Nielsen T., Pons N., Levenez F., Yamada T., Mende DR., Li J., Xu J., Li S., Li D., Cao J., Wang B., Liang H., Zheng H., Xie Y., Tap J., Lepage P., Bertalan M., Batto J-M., Hansen T., Le Paslier D., Linneberg A., Nielsen HB., Pelletier E., Renault P., Sicheritz-Ponten T., Turner K., Zhu H., Yu C., Li S., Jian M., Zhou Y., Li Y., Zhang X., Li S., Qin N., Yang H., Wang J., Brunak S., Doré J., Guarner F., Kristiansen K., Pedersen O., Parkhill J., Weissenbach J., Bork P., Ehrlich SD., Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65. DOI: 10.1038/nature08821.

Quince C., Curtis TP., Sloan WT. 2008. The rational exploration of microbial diversity. *The ISME journal* 2:997–1006. DOI: 10.1038/ismej.2008.69.

Quince C., Lanzén A., Curtis TP., Davenport RJ., Hall N., Head IM., Read LF., Sloan WT. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods* 6:639–41. DOI: 10.1038/nmeth.1361.

Riesenfeld CS., Goodman RM., Handelsman J. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental*

Microbiology 6:981–989. DOI: 10.1111/j.1462-2920.2004.00664.x.

Riesenfeld CS., Schloss PD., Handelsman J. 2004. Metagenomics: Genomic analysis of microbial communities. *Annual review of genetics* 38:525–52. DOI: 10.1146/annurev.genet.38.072902.091216.

Rondon MR., August PR., Bettermann AD., Brady SF., Grossman TH., Liles MR., Loiacono KA., Lynch BA., MacNeil IA., Minor C., others. 2000. Cloning the soil metagenome: A strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology* 66:2541–2547.

Rosen GL., Reichenberger ER., Rosenfeld AM. 2011. NBC: The Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics (Oxford, England)* 27:127–9. DOI: 10.1093/bioinformatics/btq619.

Sayers EW., Barrett T., Benson D., Bryant SH., Canese K., Chetvernin V., Church DM., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer LY., Helmberg W., Kapustin Y., Landsman D., Lipman DJ., Madden TL., Maglott DR., Miller V., Mizrachi I., Ostell J., Pruitt KD., Schuler GD., Sequeira E., Sherry ST., Shumway M., Sirotkin K., Souvorov A., Starchenko G., Tatusova T a., Wagner L., Yaschenko E., Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 37:D5–15. DOI: 10.1093/nar/gkn741.

Schloissnig S., Arumugam M., Sunagawa S., Mitreva M., Tap J., Zhu A., Waller A., Mende DR., Kultima JR., Martin J., Kota K., Sunyaev SR., Weinstock GM., Bork P. 2013. Genomic variation landscape of the human gut microbiome. *Nature* 493:45–50. DOI: 10.1038/nature11711.

Schloss PD., Handelsman J. 2006. Toward a census of bacteria in soil. *PLoS computational biology* 2:e92. DOI: 10.1371/journal.pcbi.0020092.

Sczyrba A., Hofmann P., Belmann P., Koslicki D., Janssen S., Droege J., Gregor I., Majda S., Fiedler J., Dahms E., Bremges A., Fritz A., Garrido-Oter R., Jorgensen TS., Shapiro N., Blood PD., Gure-

vich A., Bai Y., Turaev D., DeMaere MZ., Chikhi R., Nagarajan N., Quince C., Hansen LH., Sorensen SJ., Chia BKH., Denis B., Froula JL., Wang Z., Egan R., Kang DD., Cook JJ., Deltel C., Beckstette M., Lemaitre C., Peterlongo P., Rizk G., Lavenier D., Wu Y-W., Singer SW., Jain C., Strous M., Klingenberg H., Meinicke P., Barton M., Lingner T., Lin H-H., Liao Y-C., Silva GGZ., Cuevas DA., Edwards RA., Saha S., Piro VC., Renard BY., Pop M., Klenk H-P., Goeker M., Kyrpides N., Woyke T., Vorholt JA., Schulze-Lefert P., Rubin EM., Darling AE., Rattei T., McHardy AC. 2017. Critical Assessment of Metagenome Interpretation – a benchmark of computational metagenomics software. *bioRxiv*:099127. DOI: 10.1101/099127.

Sedlar K., Kupkova K., Provaznik I. 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal* 15:48–55. DOI: 10.1016/j.csbj.2016.11.005.

Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*:1–7. DOI: 10.1038/nmeth.2066.

Silva GGZ., Cuevas D a., Dutilh BE., Edwards R a. 2014. FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2:e425. DOI: 10.7717/peerj.425.

Stark M., Berger S., Stamatakis A., von Mering C. 2010. MLTreeMap - accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC genomics* 11:461. DOI: 10.1186/1471-2164-11-461.

Stewart EJ. 2012. Growing Unculturable Bacteria. *Journal of Bacteriology* 194:4151–4160. DOI: 10.1128/JB.00345-12.

Strous M., Kraft B., Bisdorf R., Tegetmeyer H. 2012. The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures. *Frontiers*

in Microbiology 3. DOI: 10.3389/fmicb.2012.00410.

Sunagawa S., Mende DR., Zeller G., Izquierdo-Carrasco F., a Berger S., Kultima JR., Coelho LP., Arumugam M., Tap J., Nielsen HB., Rasmussen S., Brunak S., Pedersen O., Guarner F., de Vos WM., Wang J., Li J., Doré J., Ehrlich SD., Stamatakis A., Bork P. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods* 10:1196–9. DOI: 10.1038/nmeth.2693.

Sutton GG., White O., Adams MD., Kerlavage AR. 1995. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science and Technology* 1:9–19. DOI: 10.1089/gst.1995.1.9.

Teeling H., Waldmann J., Lombardot T., Bauer M., Glöckner FO. 2004. TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC bioinformatics* 5:163. DOI: 10.1186/1471-2105-5-163.

Turnbaugh PJ., Gordon JI. 2008. An Invitation to the Marriage of Metagenomics and Metabolomics. *Cell* 134:708–713. DOI: 10.1016/j.cell.2008.08.025.

Tyson GW., Chapman J., Hugenholtz P., Allen EE., Ram RJ., Richardson PM., Solovyev VV., Rubin EM., Rokhsar DS., Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43. DOI: 10.1038/nature02340.

Ufarté L., Potocki-Veronese G., Laville É. 2015. Discovery of new protein families and functions: New challenges in functional metagenomics for biotechnologies and microbial ecology. *Frontiers in Microbiology* 6. DOI: 10.3389/fmicb.2015.00563.

Ulyantsev VI., Kazakov SV., Dubinkina VB., Tyakht AV., Alexeev DG. 2016. MetaFast: Fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* 32:2760–2767. DOI: 10.1093/bioinformatics/btw312.

Venter JC., Remington K., Heidelberg JF., Halpern AL., Rusch D., Eisen J a., Wu D., Paulsen I., Nelson KE., Nelson W., Fouts DE.,

Levy S., Knap AH., Lomas MW., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y-H., Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)* 304:66–74. DOI: 10.1126/science.1093857.

Vinh LV., Lang TV., Binh LT., Hoai TV. 2015. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithms for Molecular Biology* 10:2. DOI: 10.1186/s13015-014-0030-4.

Walt S van der., Colbert SC., Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* 13:22–30. DOI: 10.1109/MCSE.2011.37.

Wang Q., Garrity GM., Tiedje JM., Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* 73:5261–7. DOI: 10.1128/AEM.00062-07.

Wommack KE., Bhavsar J., Ravel J. 2008. Metagenomics: Read Length Matters. *Applied and Environmental Microbiology* 74:1453–1463. DOI: 10.1128/AEM.02181-07.

Wood DE., Salzberg SL. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15:R46. DOI: 10.1186/gb-2014-15-3-r46.

Woyke T., Tighe D., Mavromatis K., Clum A., Copeland A., Schackwitz W., Lapidus A., Wu D., McCutcheon JP., McDonald BR., Moran N a., Bristow J., Cheng J-F. 2010. One bacterial cell, one complete genome. *PloS one* 5:e10314. DOI: 10.1371/journal.pone.0010314.

Woyke T., Xie G., Copeland A., González JM., Han C., Kiss H., Saw JH., Senin P., Yang C., Chatterji S., Cheng J-F., Eisen J a., Sieracki ME., Stepanauskas R. 2009. Assembling the marine metagenome, one cell at a time. *PloS one* 4:e5299. DOI: 10.1371/journal.pone.0005299.

Wu D., Hugenholtz P., Mavromatis K., Pukall R., Dalin E., Ivanova NN., Kunin V., Goodwin L., Wu M., Tindall BJ., Hooper SD., Pati

A., Lykidis A., Spring S., Anderson IJ., D'haeseleer P., Zemla A., Singer M., Lapidus A., Nolan M., Copeland A., Han C., Chen F., Cheng J-F., Lucas S., Kerfeld C., Lang E., Gronow S., Chain P., Bruce D., Rubin EM., Kyrpides NC., Klenk H-P., Eisen J a. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056–60. DOI: 10.1038/nature08656.

Wu Y-W., Tang Y-H., Tringe SG., Simmons BA., Singer SW. 2014. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26. DOI: 10.1186/2049-2618-2-26.

Yu F., Blainey PC., Schulz F., Woyke T., Horowitz MA., Quake SR. 2017. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *bioRxiv*:114496. DOI: 10.1101/114496.

Zhao Y., Tang H., Ye Y. 2011. RAPSearch2: A fast and memory-efficient protein similarity search tool for next generation sequencing data. *Bioinformatics* 28:125–126. DOI: 10.1093/bioinformatics/btr595.

Appendices

Appendix A

Supplementary Material "*Taxator-tk*: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods"

Supplementary Methods for *Taxator-tk*: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods

I. Taxonomic Assignment of Sequence Segments

Here we describe in detail the individual steps and the run-time properties of the algorithm which is implemented in the program *taxator,* the second stage of the overall binning workflow using *taxator-tk* (Fig. 2b). We propose the realignment placement algorithm (RPA) for the taxonomic assignment of a query segment *q*, which can be any subsequence of the full query sequence (i.e. the query can be a read, contig, scaffold or a complete genome sequence). The algorithm constitutes **two pairwise alignment passes** and in each, *q* is aligned to segments of nucleotide reference sequences. It aims at identifying as many as possible taxa of the prediction clade (node R in Fig. 2a) without explicitly resolving its phylogenetic structure.

1. Among the given set of homologous segments constructed from overlapping alignments before application of the RPA, we define *s* to be the most similar segment to *q*, i.e. the one with the best local alignment score of all reference segments. In the first pass, all segments are aligned against *s* (n alignments). The resulting pairwise scores, our implementation uses the **edit distance** (mismatches + gaps), define an ordering among all segments or their corresponding taxa. The distinction between segments and associated taxa will be neglected in the following for better readability. All taxa which are less distant to *s* than *q*, including *s* itself, are added to an empty set *M* which holds all identified taxa of the prediction clade. The first more distant taxon than *q* is defined to be the outgroup segment *o* (Fig. 2c) and used as the alignment target in the following second and last pass in which similar taxa to *o* are added *M*.

2. We align all segments, including *q*, against *o* and rank the resulting scores. Then we add all taxa to *M* which have a lower score than *q*. With some fine-tuning, we chose to also add taxa with a higher score than *q*, within a small range accounting for erroneous scores, because *o* and *q* can be very distant homologs with noisy alignment. The width of this **error band** is determined on a per-segment basis as a linear score function of the taxonomic disorder in the alignment scores and not a universal or configurable run-time parameter. We interpret a rank disorder (e.g. a known family member of *o* being more similar to *o* than a corresponding

species member segment) as a discordance between gene tree and taxonomy and proportionally scale the effective score of *q* to enlarge *M* by taxa which are slightly more distant to *o* than *q*. This second pass requires $n-1$ new alignments, or less if some segments are identical to either *q* or *s*.

If multiple best references (*s)* or outgroup segments (*o)* were present in these two passes with identical alignment scores, the calculations are repeated for every such segment in order to produce stable output. We reduced the additional computational effort in our implementation by detecting frequent identical segments and uninformative homologs. The final assignment taxon ID of *q* is the lowest common ancestor (LCA) of the taxa in *M*, or none if no outgroup had been found. The theoretical run-time in the segment assignment algorithm measured in units "number of pairwise alignments" is in $O(n)$ and about $2n$, where n denotes the number of homologous segments. The run-time complexity for a single pairwise alignment is $O(l^2)$ and scales quadratically with the segment length l. Therefore the total run-time complexity per segment is $O(nl^2)$ and the total worst-case run-time for the entire query sequence can be bounded above by mL^2 where m denotes the maximum number n of segment homologs among all query segments and L is the total length of the query sequence. Thus, the run-time for the entire sample in the worst case scales linearly with the amount of sequence data (bp) and linearly with the number of homologs but quadratically with the length of the individual segments. Segments with an excessive number of homologs, most often short segments of abundant and uninformative regions, have a negative impact on the program run-time. We currently limit the number of homologs per query to the **top-scoring 50** by default in our pipeline scripts (configurable run-time parameter in program *alignments-filter* or directly in the local alignment search program), before passing them to *taxator*. Other tested **values gave similar results** and the parameter, if changed, should be chosen based on hardware limitations. If this parameter is set lower, then the number of reference segments drops below a critical value such that no outgroup can be determined for some *q* and which therefore remain unassigned (but without impacting the taxon ID of other segments).

II. Consensus Binning Algorithm

Due to sparse segments and taxonomic assignment thereof with *taxator* in stage two of the workflow (Fig. 1b), a final processing step (Fig. 1c) is required to de-

termine a taxon ID for the entire query sequence. Therefore we have implemented a simplistic, weighted consensus assignment scheme in the program *binner,* which optionally permits to apply custom constraints, e.g. the minimum percentage identity (PID) for classification at the species level or the removal of taxa with low counts in the whole sample. However, there are currently only **two mandatory run-time parameters** to control the actual post-processing consensus algorithm. First we define the support of a query segment to be the number of total identical positions to the best reference segment. The first run-time parameter specifies the **minimum combined support** at any rank (50 positions by default) and serves to ignore false predictions caused by short and often noisy segments. The other parameter specifies the **minimum percentage of the summed support** (70% by default) to allow a majority taxon to outvote a contradicting minority. Inconsistent taxa below this support value are resolved by the LCA operation until the threshold is reached. Probably due to the conservative nature of the RPA, we found those two parameters to have minimal impact on the binning results in practice. The output of taxator additionally includes the taxa in the evolutionary neighborhood, a score reflecting the agreement between the segment tree and the taxonomy, as well as a score for interpolation of the query-branch location between the R and X nodes of Fig. 2. We provide Python language bindings for processing with other applications.

III. Taxonomy and Phylogeny

Taxator-tk assumes that the NCBI taxonomy used for the assignment correctly captures the evolutionary process of speciation, although we know that the categorization of some taxa might be inconsistent with their evolution. If the phylogenetic information inferred from similarity scores disagrees with the taxonomic structure, assignments are made to a consistent higher rank. For instance horizontal gene transfer and upstream sequence misassembly can cause multiple similar copies of a sequence to be distributed across unrelated taxa. In case a query sequence cannot be traced by the algorithm to have evolved with either copy, it is usually assigned to the LCA of these clades. However, if the donor clade is unknown, the query may also be assigned to the recipient clade and the horizontal transfer or misassembly can go undetected. Thus assignment errors caused by the evolution of genes, upstream technical errors or taxonomy cannot always be eliminated in this framework. It remains to be assessed whether the use of an alternative microbial taxonomy such as the

GreenGenes[1] or the SILVA[2] taxonomy would improve on the taxonomic assignment.

IV. Comparison and Innovations

Taxator-tk shares some ideas with previous programs: Starting with *MEGAN*[3], which uses local alignments scores to define a "neighborhood of related sequences" and then makes a taxonomic estimate which is the LCA of the corresponding taxa. This neighborhood threshold is a percentage of the local alignment score and can be interpreted to reflect the rate of evolution within a taxonomic group. Its value is empirical and lacks stronger justification. The neighborhood definition has been improved in *taxator-tk* and other programs. To our knowledge, *SOrt-ITEMS*[4] was the first algorithm to use the logic of realignment to the best reference (termed reciprocal similarity) for read assignment but is restricted to protein level alignment and is implemented as a wrapper around (the legacy C version of) *BLAST+*[3]. Protein-level alignment in general triples the run-time of the local alignment step (translation into three frame shifts) and cannot make use of faster nucleotide aligners. *SOrt-ITEMS* also uses fixed similarity thresholds in terms of percentage identity to define universal levels of conservation within taxonomic groups assuming the same rate of evolution for different genetic regions and clades. Furthermore *SOrt-ITEMS* was primarily designed for reads and if it performs well for longer sequences, its run-time is expected to increase proportionally with input sequence lengths. Both follow-up programs *taxator-tk* and *CARMA3*[6] adopted the logic of reciprocal alignment, extended it and removed the assumption of universal conservation levels. *CARMA3* accounts for a heterogeneous rate of evolution for different genetic regions. The initial identification of similar sequences in the reference can be based on nucleotide or protein *BLAST* search or profile Hidden Markov Models with *HMMER*[7]. In *BLAST* mode, *CARMA3*, like *SOrt-ITEMS*, uses a single reciprocal alignment search and then extra or interpolates alignment scores to select a taxonomic rank for prediction. It therefore assumes a parameterized model for the conservation level at a taxonomic rank: a linear function which is fitted to the observed local alignment scores.

With *taxator-tk*, we use a non-parametric score ranking algorithm, instead. Also, to our knowledge, we provide the first algorithm to determine a proper outgroup and to sparsify the input data being able to assign distinct regions on the query sequence to possibly different taxonomic groups. Also, we at most assume segment-wise constant rates of evolution (equally long branches from a common ancestor).

This makes the major algorithmic component parameter-less and robust in itself, independent of the individual segment sizes. Through the sparsification procedure it incorporates structural rearrangements among distant relatives and scales better with the length of the input sequences. The individual segment assignments allow for a robust consensus voting scheme for the assignment of entire sequence fragments. The segment-specific classifications could also be used to detect the inconsistent taxonomic composition of an input sequence which can be caused by horizontal gene transfers events (HGTs) and assembly errors. Different from most previous approaches, *taxator-tk* was developed for and tested using fast nucleotide sequence local alignments instead of protein sequence alignments, although for the local alignments in stage 1 of the workflow both can be used. Our comparisons, however, suggest that the additional computations which are required for protein-level homology search do not considerably improve the results with *taxator-tk*. Thus, taxonomic binning of a metagenome sample with *taxator-tk* requires no more than specification of reference sequences, their taxonomic affiliations and an aligner like *BLAST* or *LAST*[8]. On the implementation side, all workflow steps for taxonomic assignment with *taxator-tk* are designed in a modular way making it easy to save, compress, reuse or recompute results. The computation-intensive classification of segments in *taxator* is run in parallel on many CPU cores while at the same time using the open source C++ algorithm library SeqAn[9] for fast pairwise alignment.

V. Performance Measures

As metagenome datasets can have varying taxonomic composition in terms of which taxa are present and their relative abundances, this needs to be taken into consideration in evaluating taxonomic assignment methods. If an algorithm performs better for some clades than for others at a given rank we call it taxonomically biased. Oftentimes a classifier is biased, if it uses parameters that fit one clade better than another. This can be the case if the parameters were chosen to give good overall assignment accuracy (low total number of false predictions) on training data with biased taxonomic composition. Such a method is optimized to perform well for the abundant taxa of these particular training data and will not generalize well when applied to a sample of different taxonomic structure and abundances. To account for uneven taxonomic composition in evaluation datasets and to obtain comparable performance estimates across datasets of different taxonomic composition, we used as the pri-

mary evaluation measure the bin-averaged **precision** (or **positive predictive value**), also known as **macro-precision**.

$$\text{macro-precision} = \frac{1}{N_p} \sum_{i=1}^{N_p} \text{precision}_i \qquad \text{(Equation V.1)}$$

where N_p is number of all predicted bins and

$$\text{precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \qquad \text{(Equation V.2)}$$

True positives TP_i are the correct assignments to the i^{th} bin and false positives FP_i the incorrect assignments to the same bin.

The macro-precision is the fraction of correct sequence assignments over all assignments to a given taxonomic bin, averaged over all predicted bins for a given rank. For falsely predicted bins which do not occur in the data, the precision is therefore zero. This value reflects how trustworthy the bin assignments are on average from a user's perspective, as it is averaged overall predicted bins.

In addition to the macro-precision, we report the raw numbers of true and false predictions for every cross-validation, as well as a quick overall precision for pooled ranks. This overall precision is most informative for species+genus+family and reports the fraction of true classifications among the predictions for all these ranks in a single pooled bin.

$$\text{overall-precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{(Equation V.3)}$$

We measure the taxonomic bias of a method in terms of the standard deviation over all individual bin precisions.

$$\text{sd}_{prec} = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} \left(\text{precision}_i - \overline{\text{precision}}\right)^2} \qquad \text{(Equation V.4)}$$

where

$$\overline{\text{precision}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \text{precision}_i \qquad \text{(Equation V.5)}$$

The standard deviation is small if all predicted bins have a similar precision. A universally good method should have a high macro-precision with a low taxonomic

bias.

The **recall** (or **sensitivity**) is a measure of completeness of a predicted bin and, analogously, the **macro-recall** is the fraction of correctly assigned sequences of all sequences belonging to a certain bin, averaged over all existing bins in the test data[10].

$$\text{macro-recall} = \frac{1}{N_r} \sum_{i=1}^{N_r} \text{recall}_i \qquad \text{(Equation V.6)}$$

where N_r is the number of all existing bins in the test data and

$$\text{recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \qquad \text{(Equation V.7)}$$

False negatives (FN_i) are the assignments belonging to the i^{th} bin but which where classified to another bin or left unassigned.

The macro-recall reflects how well the classifier works more from a developer's perspective than from the user's perspective, as it is usually not known which predicted bins correspond to existing ones and which do not.

VI. Low-abundance Filtering

The number of predicted bins at each rank can be quite large, at most the number of known taxa in the taxonomy and reference sequence data. When noise is considered to occur evenly distributed across this large output space, bins with few assigned sequences are more likely to be falsely identified, than larger bins (the chance to independently classify the same bin by chance *n* times is $\left(\frac{1}{m}\right)^n$, where m is the number of possible bins). Since the macro precision is an average over all predicted bins, it is heavily affected by bins with few sequences assigned. As a result, classifiers that predict clades present at low frequencies in the sample score badly under this measure. To correct for this effect, we define a truncated average precision ignoring the least abundant predicted bins and consider only the **largest predicted bins constituting a minimum fraction α of the total assignments** (equal size bins are also included). This modification acts as a noise filter and accounts for different behavior of classifiers without explicitly considering the size of the model space or the number of existing species in the actual sample. We set α to 0.99 for our evaluations.

VII. Cross-validation

Despite the limitations of simulated metagenomes, which incorporate assumptions about sequencing error rates or species abundance distributions, it is very informative to evaluate taxonomic assignment methods on simulated sequence data as real metagenome samples lack taxon IDs for evaluation. Our canonical way of evaluating a method on simulated data is a version of **leave-one-out cross-validation**: Each query sequence is classified by removing all identical or related sequences up to a given rank from the reference collection: For example, to assess the performance in assigning query sequences from a new species, all sequences belonging to this species are removed from the reference sequence collection for the classifier. Performance measures (macro-recall, macro-precision), along with other statistics (true/false/unassigned data, overall precision, bin counts) which are available in the coupled tables, were normally calculated in units of the number of assigned base-pairs or the number of assigned sequences, if these had comparable lengths. These values were calculated for all ranks (species, genus, family, order, class, phylum, domain/superkingdom) for seven simulations: either all reference data was used (per query) or all data from **the query** species, genus, family, order, class or phylum was removed from the reference data prior to classification. The assignments of these seven cross-validation experiments were averaged for a combined performance summary with standard measures.

VIII. Consistency Analysis

In order to evaluate the predictions for real metagenome samples where no underlying correct taxon IDs are known for the sequences, we assigned sequences linked by assembly and calculated an assignment consistency value. We split long contigs into multiple pieces and classified each piece independently. Assuming that the sequence assembly was correct in the first place, contradicting assignments of pieces that originate from the same contig represent false assignments. This unveils part of the errors made by a particular method but some, if not the majority, will go undetected because the actual ID stays unknown and the assignments for a contig can be consistently wrong. Hence these results are generally more difficult to interpret than those from simulated data.

IX. Sequence Homology Search via Local Alignment

In the course of evaluation we created many local alignments as input to the

taxonomic assignment programs *CARMA3*, *MEGAN4/5* and *taxator-tk*. The nucleotide alignments were mostly generated using the alignment program *LAST* (version 320) because it ran faster without noticeable differences in the output alignments than *BLAST+/blastn (version 2.2.28+)*. The protein-level alignments which we used in our evaluations were generated with *BLAST+/tblastx (version 2.2.28+)* because we wanted to compare with identical nucleotide reference sequences. We support and tested with different alignment programs for the fact that *BLAST* is standard and easy to parallelize whereas *LAST* has a faster algorithm but high memory requirements. It ran with comparable speed to the *BLAST+/megablast* algorithm which has a limited sensitivity and in practice resulted in a two to four times reduced amount of query sequences being aligned and classified. For a detailed comparison of alignment programs and how LAST compares to other programs such as *RAPSEARCH2*[11] and *BLAT*[12], consider Niu et al.[13] and Darling et al.[14]. In our evaluations, *LAST* was roughly 50 to 200 times faster than *BLAST+/blastn* and about as fast as *BLAST+/megablast* (which has much reduced sensitivity). *LAST* is also tunable for better sensitivity with protein-coding nucleotide sequences using a special form of seeding. If other alignment programs are found to be better-suited for a particular data type, these can easily be incorporated into the provided workflows. For instance, local protein sequence alignments can be performed in the homology search step, e.g. by using *BLAST+/tblastx*. There are fast aligners such as *RAPSEARCH2*, *PAUDA*[15] and *DIAMOND*[16] that allow searching for homologs in large reference collections of amino acid sequences. To produce compatible input for *taxator-tk*, the amino acid alignment positions must be converted into nucleotide positions.

For our short sequence length evaluation (Supplementary Fig. S6-S8), evaluation of a published SimMC scenario (Supplementary Fig. S21) and evaluation of a simulated metagenome sample with 49 species (Fig. 3, Supplementary Fig. S11-S13), we used a standard *BLAST+/blastn (version 2.2.28+)* and *BLAST+/tblastx* search. We chose the default alignment parameters and scoring schemes with each aligner. The generated alignments were then provided in *BLAST* tabular format to be usable with *CARMA3* and *MEGAN4/MEGAN5*. *Taxator-tk* reads a simplistic tab-separated alignment format that can be generated directly with *BLAST+* or with conversion scripts which we provide for the MAF alignment format of *LAST*. This arrangement ensures that *taxator-tk* can be easily adapted to profit from advancements in the field of local alignment in future. Users can also employ amino acid level align-

ment if the final output is mapped back to positions on the nucleotide reference and query sequences. The easiest way to achieve this is to use *BLAST+/tblastx* although this is computationally more demanding than directly searching a collection of protein sequences for which also nucleotide sequences are available.

X. Program Parameters and Versions

For taxonomic assignment with *MEGAN4* (version 4.70.4) we used minscore=20, toppercent=20, minsupport=5 and mincomplexity=0.44 parameters. With *MEGAN5* (version 5.4.3), we used the default options minsupport=10, minscore=50, max_expected=0.01, minimal_coverage_heuristic=on and top_percent=20, as with *MEGAN4*. In *CARMA3*, we used the standard parameters in the contained configuration file. *Kraken* (version 0.10.4b) was also applied with the standard commands and without shrinking the database (shrink_db.sh). *Taxator-tk* (version 1.1.1-extended) was run with standard settings, being restricted to the 50 best scoring local alignments to avoid long run-times for some of the query sequences. This is purely a convenience filter at the current state of development and is meant to be replaced by an adaptive per-segment heuristic.

XI. 16S Cross-validation

We evaluated the performance of *taxator-tk* in classifying the most widely used taxonomic marker gene in studies of microbial diversity, the 16S rRNA gene, as a proof of concept. For our evaluation, we extracted 7,175 annotated 16S rRNA genes (Suppl. Fig. 5) each with a minimum length of 1 kb from *mRefSeq47* (Suppl. Fig. 9). The sequences were assigned with *taxator-tk* using the entire mRefSeq as reference, not just 16S genes. The cross-validation assesses the performance of 16S gene assignment in a wide range of situations. The performance statistics were calculated based on the number of assigned sequences, as all have comparable length. When using the complete reference sequences, 87% of sequences were assigned to the ranks of species, genus and family with 100% accuracy (Supplementary Fig. S3b), the remaining 13% were correctly assigned at higher ranks. This is an ideal situation showing the baseline on our dataset (in terms of the assigned rank depth). In more realistic simulations, when we tested assignment of genes from novel species or novel higher-level clades, assignments were accordingly made to higher ranks in most cases. For instance, when simulation novels species, 2,678 contigs were assigned to the correct genera, while 491 erroneous species and genus assignments

were made. The macro-precision in the combined cross-validation (Fig. 2) was always above 92%, with standard deviations from 10 to 25%, which demonstrates a good and even performance of *taxator-tk* for all clades in the case of 16S rRNA data.

XII. FAMeS Cross-validation

On the FAMeS contig datasets, taxator-tk produced fewer errors for all taxonomic ranks than *MEGAN4*, which was accompanied by a moderate reduction in macro-recall throughout all individual experiments and in the combined cross-validation experiments: For SimMC, the macro-precision was three to four times as large as *MEGAN4*'s for species to order, with higher macro-recall (Supplementary Fig. S17-S18). The species to family overall precision was ~91% for *taxator-tk* (~59% for *MEGAN4*) and taxator-tk estimated 54 species bins (*MEGAN4* 188) for the 47 actual species in SimMC. Similarly, for SimHC, *taxator-tk* achieved a higher macro-precision for all ranks, which was most pronounced for class and phylum (Supplementary Fig. S19-S20). By contrast, the macro-recall was slightly reduced and both methods underestimated the 96 existing species in SimHC.

XIII. Supplementary Files

The PDF attachment includes informative interactive charts and files which are necessary to reproduce the results which are shown in the article. Larger benchmark data can be downloaded from http://algbio.cs.uni-duesseldorf.de/software/.

Supplementary Methods References

1. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–72 (2006).
2. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–6 (2013).
3. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–86 (2007).
4. Monzoorul Haque, M., Ghosh, T. S., Komanduri, D. & Mande, S. S. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* **25,** 1722–30 (2009).
5. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
6. Gerlach, W. & Stoye, J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.* 1–11 (2011).

7. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39 Suppl 2,** W29–37 (2011).

8. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11,** 80 (2010).

9. SeqAn. at http://www.seqan.de

10. McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* **4,** 63–72 (2007).

11. Zhao, Y., Tang, H. & Ye, Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next generation sequencing data. *Bioinformatics* **28,** 125–126 (2011).

12. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–64 (2002).

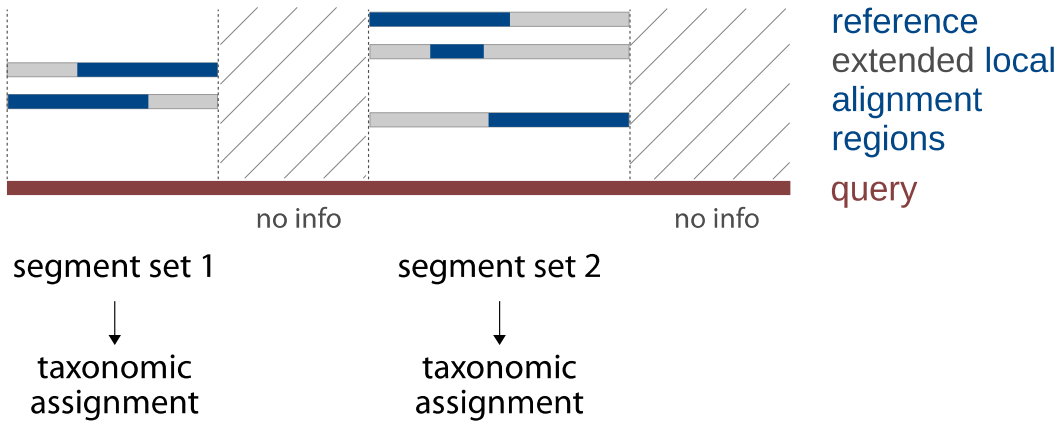
13. Niu, B., Zhu, Z., Fu, L., Wu, S. & Li, W. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27,** 1704–5 (2011).

14. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2,** e243 (2014).

15. Huson, D. H. & Xie, C. A poor man's BLASTX--high-throughput metagenomic protein database search using PAUDA. *Bioinformatics* **30,** 38–9 (2014).

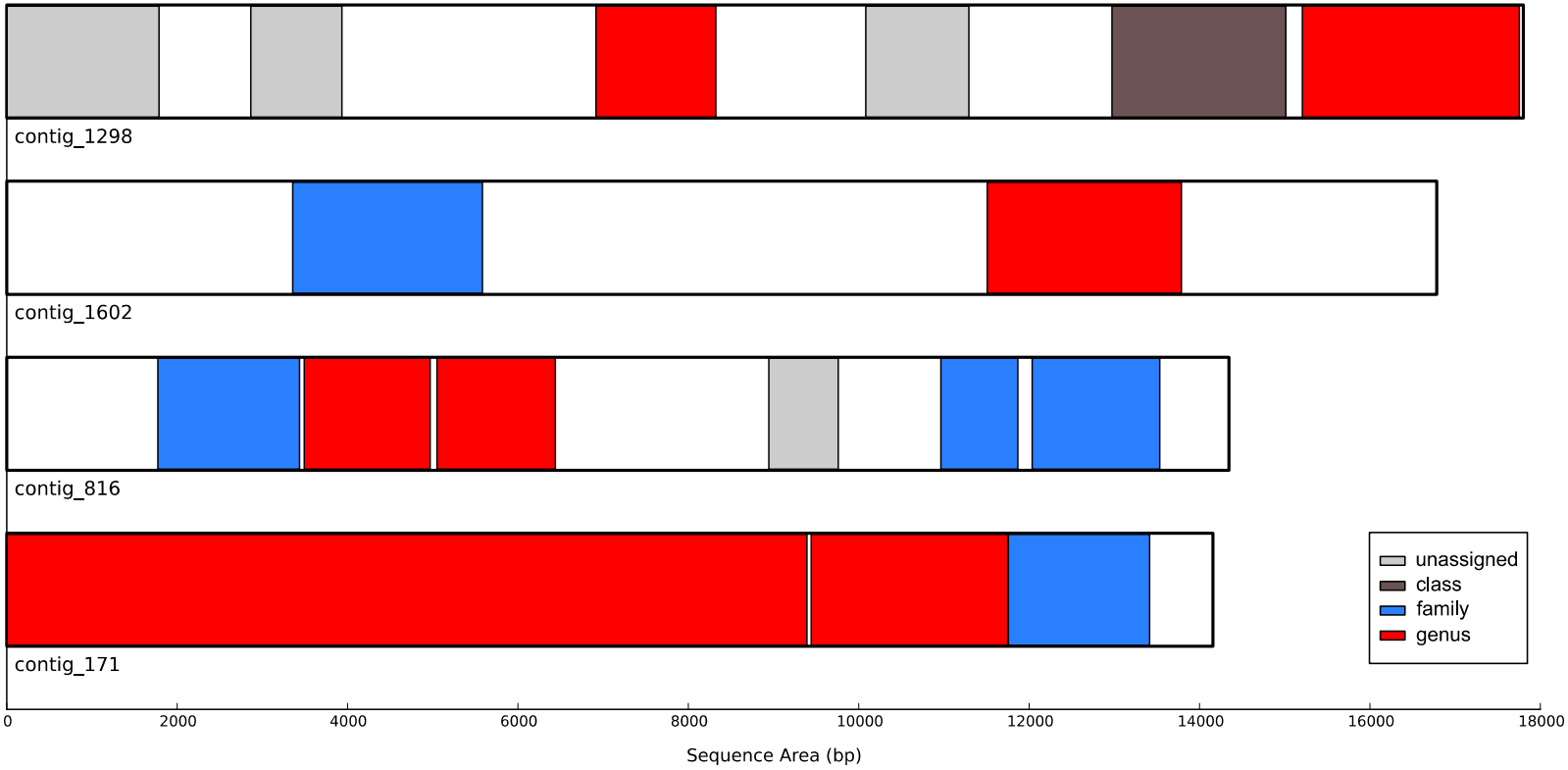
16. Buchfink, B., Xie, C. & Huson, D. H. Fast and Sensitive Protein Alignment using DIAMOND, under review.

Supplementary Figure S1: Query sequence segmentation and segment splicing



Query and corresponding reference segments from local alignment region extension and splicing. Blue bars correspond to original local alignment regions on reference nucleotide sequences which are positionally aligned to the query nucleotide sequence in red. These alignments are generated by a local (nucleotide) sequence aligner such as *BLAST* or *LAST* before running *taxator*. If alignments overlap on the query, they are joined into query segments which are flanked by regions without detected similarity to any known reference sequence. Reference segments are constructed from the original alignment reference regions (blue) by extension (gray bars) with the same number of nucleotides which are missing to match the length of the query segment. The corresponding sets of homologs are the input to the core taxonomic assignment algorithm in *taxator*.

Supplementary Figure S2: Taxonomic assignment of segments

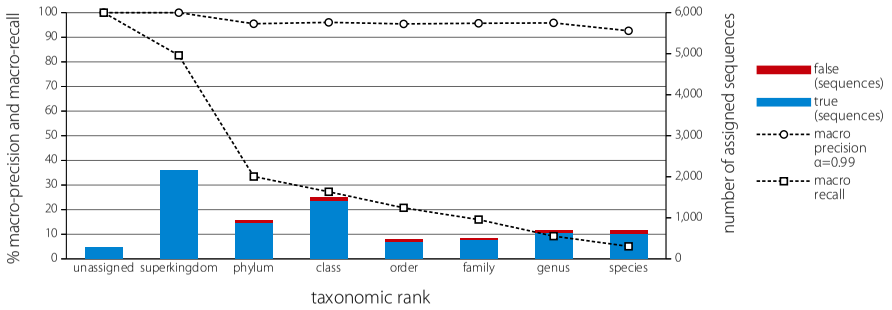


Four long contigs of the SimMC data-set. Colored boxes show segments that where assigned by *taxator*, when all species reference data was removed (new species simulation). White regions in between lack alignments by the local alignment search and have therefore no homologs for assignment. All assigned regions in this example are consistently assigned at the taxonomic ranks genus, family and class. The shown segments are used by the program *binner* to derive consistent whole-sequence taxonomic assignments, as done in our evaluations.

Supplementary Figure S3 - 16S gene assignment with taxator-tk **(a) summary scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	274.4	0.0	0	100.0	0.0	1	100.0	0.0	1	4593.6	0.0	100.0	root+superkingdom
superkingdom	1	2159.6	0.0	0	100.0	0.0	2	82.7	14.2	2				
phylum	2	869.1	66.6	0	95.5	13.8	13	33.5	23.6	32	2707.1	228.6	92.2	phylum+class+order
class	3	1417.9	92.6	0	96.1	10.7	25	27.3	18.4	52				
order	4	420.1	69.4	0	95.4	12.6	62	20.7	14.2	109				
family	5	471.6	26.1	0	95.7	13.5	148	16.0	11.9	235	1732.1	174.1	90.9	family+genus+species
genus	6	636.7	65.0	0	95.8	16.1	342	9.2	8.9	615				
species	7	623.8	83.0	0	92.6	24.5	570	5.1	6.7	1416				
avg/sum	2.6	6598.8	402.7	0	95.9	13.0	166.0	27.8	14.0	351.6			94.2	all but unassigned
avg/sum	2.6	6873.3	402.7	0	96.4	11.4	145.4	36.8	12.2	307.8			94.5	all with unassigned

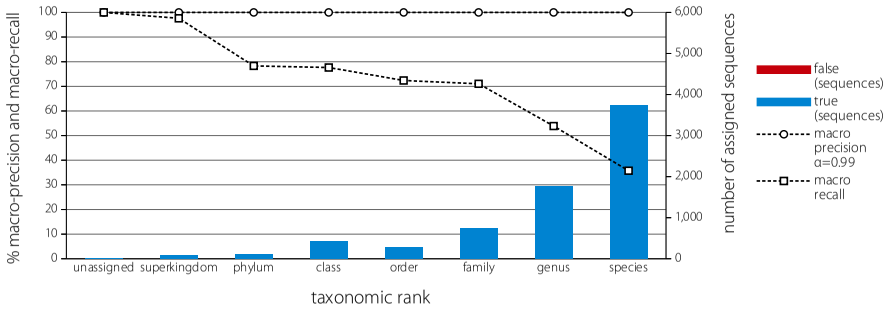
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(b) all reference scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	10	0	0	100.0	0.0	1	100.0	0.0	1	170	0	100.0	root+superkingdom
superkingdom	1	80	0	0	100.0	0.0	2	97.6	2.4	2				
phylum	2	113	0	0	100.0	0.0	16	78.3	38.3	32	813	0	100.0	phylum+class+order
class	3	428	0	0	100.0	0.0	29	77.6	37.2	52				
order	4	272	0	0	100.0	0.0	67	72.4	39.6	109				
family	5	750	0	0	100.0	0.0	158	71.0	40.4	235	6272	0	100.0	family+genus+species
genus	6	1779	0	0	100.0	0.0	337	53.9	48.0	615				
species	7	3743	0	0	100.0	0.0	504	35.8	46.8	1416				
avg/sum	5.0	7165	0	0	100.0	0.0	159.0	69.5	36.1	351.6			100.0	all but unassigned
avg/sum	5.0	7175	0	0	100.0	0.0	139.3	73.3	31.6	307.8			100.0	all with unassigned

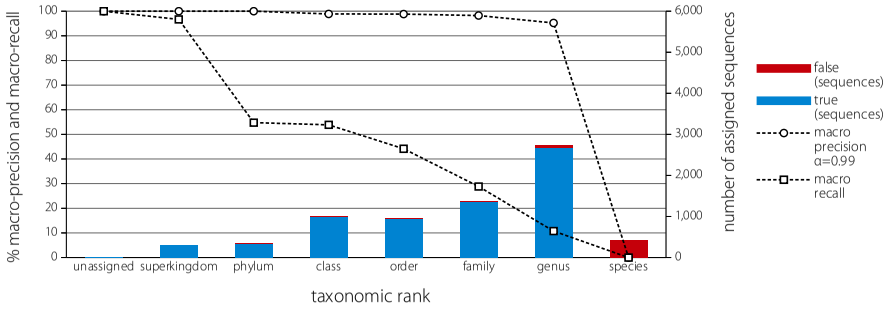
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(c) new species scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	22	0	0	100.0	0.0	1	100.0	0.0	1	648	0	100.0	root+superkingdom
superkingdom	1	313	0	0	100.0	0.0	2	96.7	3.2	2				
phylum	2	347	2	0	100.0	0.0	14	54.8	39.5	32	2284	19	99.2	phylum+class+order
class	3	989	8	0	98.9	3.9	26	53.9	41.6	52				
order	4	948	9	0	98.8	7.9	54	44.2	39.6	109				
family	5	1350	18	0	98.2	7.3	91	28.8	37.5	235	4028	509	88.8	family+genus+species
genus	6	2678	64	0	95.2	18.9	88	10.8	26.3	615				
species	7	0	427	0	0.0	0.0	54	0.0	0.0	1416				
avg/sum	4.6	6625	528	0	84.4	5.4	47.0	41.3	26.8	351.6			92.6	all but unassigned
avg/sum	4.6	6647	528	0	86.4	4.7	41.3	48.6	23.5	307.8			92.6	all with unassigned

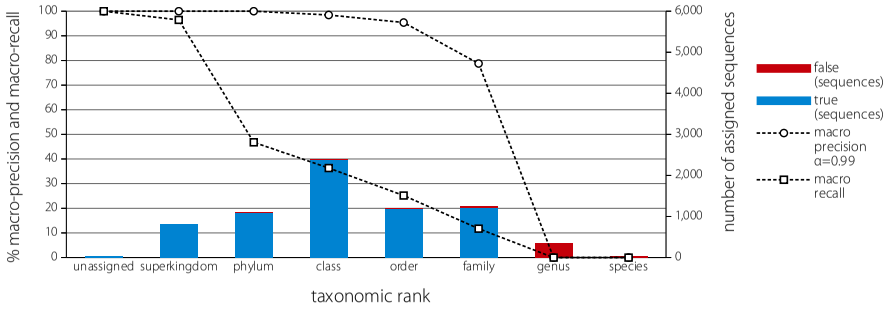
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(d) new genus scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	48	0	0	100.0	0.0	1	100.0	0.0	1	1656	0	100.0	root+superkingdom
superkingdom	1	804	0	0	100.0	0.0	2	96.5	3.0	2				
phylum	2	1098	2	0	100.0	0.0	12	46.8	38.9	32	4680	19	99.6	phylum+class+order
class	3	2392	8	0	98.5	4.7	22	36.3	35.7	52				
order	4	1190	9	0	95.4	17.8	48	25.2	32.9	109				
family	5	1201	36	0	78.8	39.2	59	11.7	26.7	235	1201	423	74.0	family+genus+species
genus	6	0	344	0	0.0	0.0	34	0.0	0.0	615				
species	7	0	43	0	0.0	0.0	8	0.0	0.0	1416				
avg/sum	3.3	6685	442	0	67.5	8.8	26.4	30.9	19.6	351.6			93.8	all but unassigned
avg/sum	3.3	6733	442	0	71.6	7.7	23.3	39.6	17.2	307.8			93.8	all with unassigned

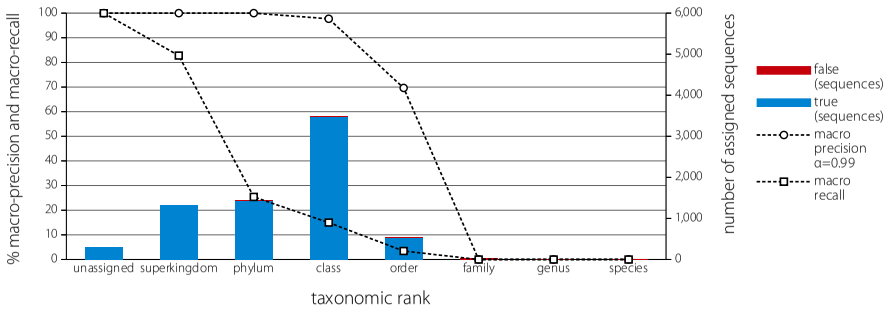
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(e) new family scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	299	0	0	100.0	0.0	1	100.0	0.0	1	2941	0	100.0	root+superkingdom
superkingdom	1	1321	0	0	100.0	0.0	2	82.8	13.7	2				
phylum	2	1442	2	0	100.0	0.0	7	25.4	35.2	32	5458	28	99.5	phylum+class+order
class	3	3485	11	0	97.7	7.2	13	15.0	26.2	52				
order	4	531	15	0	69.6	42.6	28	3.4	12.1	109				
family	5	0	38	0	0.0	0.0	24	0.0	0.0	235	0	69	0.0	family+genus+species
genus	6	0	17	0	0.0	0.0	9	0.0	0.0	615				
species	7	0	14	0	0.0	0.0	3	0.0	0.0	1416				
avg./sum	2.4	6779	97	0	52.5	7.1	12.3	18.1	12.5	351.6			98.6	all but unassigned
avg./sum	2.4	7078	97	0	58.4	6.2	10.9	28.3	10.9	307.8			98.6	all with unassigned

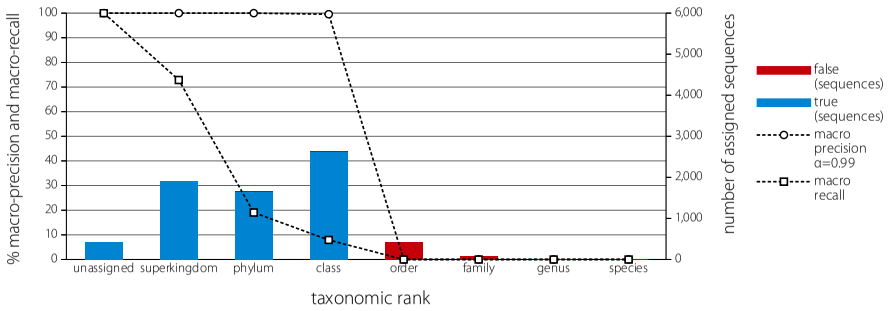
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(f) new order scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	424	0	0	100.0	0.0	1	100.0	0.0	1	4264	0	100.0	root+superkingdom
superkingdom	1	1920	0	0	100.0	0.0	2	72.9	22.3	2				
phylum	2	1665	2	0	100.0	0.1	6	19.1	31.6	32	4296	448	90.6	phylum+class+order
class	3	2631	12	0	99.6	1.1	8	7.9	20.8	52				
order	4	0	434	0	0.0	0.0	17	0.0	0.0	109				
family	5	0	74	0	0.0	0.0	11	0.0	0.0	235	0	87	0.0	family+genus+species
genus	6	0	2	0	0.0	0.0	3	0.0	0.0	615				
species	7	0	11	0	0.0	0.0	1	0.0	0.0	1416				
avg./sum	2.1	6216	535	0	42.8	0.2	6.9	14.3	10.7	351.6			92.1	all but unassigned
avg./sum	2.1	6640	535	0	49.9	0.1	6.1	25.0	9.3	307.8			92.5	all with unassigned

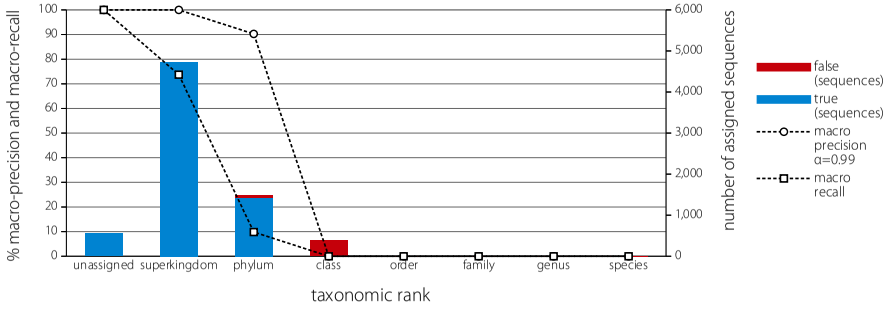
taxator-tk on RefSeq 16S genes



Supplementary Figure S3 - 16S gene assignment with taxator-tk **(g) new class scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	549	0	0	100.0	0.0	1	100.0	0.0	1	10017	0	100.0	root+superkingdom
superkingdom	1	4734	0	0	100.0	0.0	2	73.7	19.6	2				
phylum	2	1419	67	0	90.3	15.4	4	9.8	23.9	32	1419	460	75.5	phylum+class+order
class	3	0	390	0	0.0	0.0	8	0.0	0.0	52				
order	4	0	3	0	0.0	0.0	8	0.0	0.0	109				
family	5	0	9	0	0.0	0.0	6	0.0	0.0	235	0	13	0.0	family+genus+species
genus	6	0	1	0	0.0	0.0	2	0.0	0.0	615				
species	7	0	3	0	0.0	0.0	1	0.0	0.0	1416				
avg./sum	1.2	6153	473	0	27.2	2.2	4.4	11.9	6.2	351.6			92.9	all but unassigned
avg./sum	1.2	6702	473	0	36.3	1.9	4.0	22.9	5.4	307.8			93.4	all with unassigned

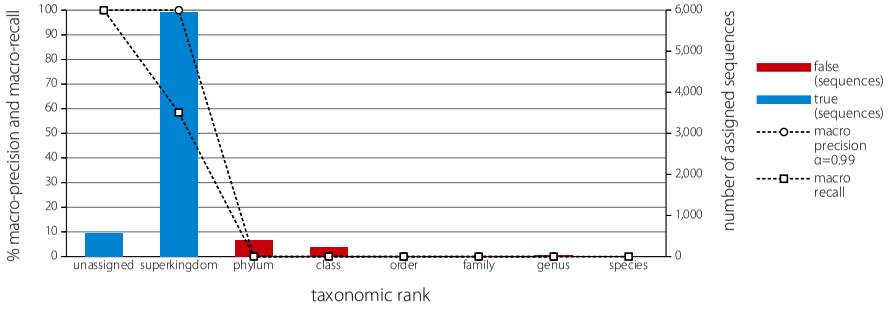
taxator-tk on RefSeq 16S genes



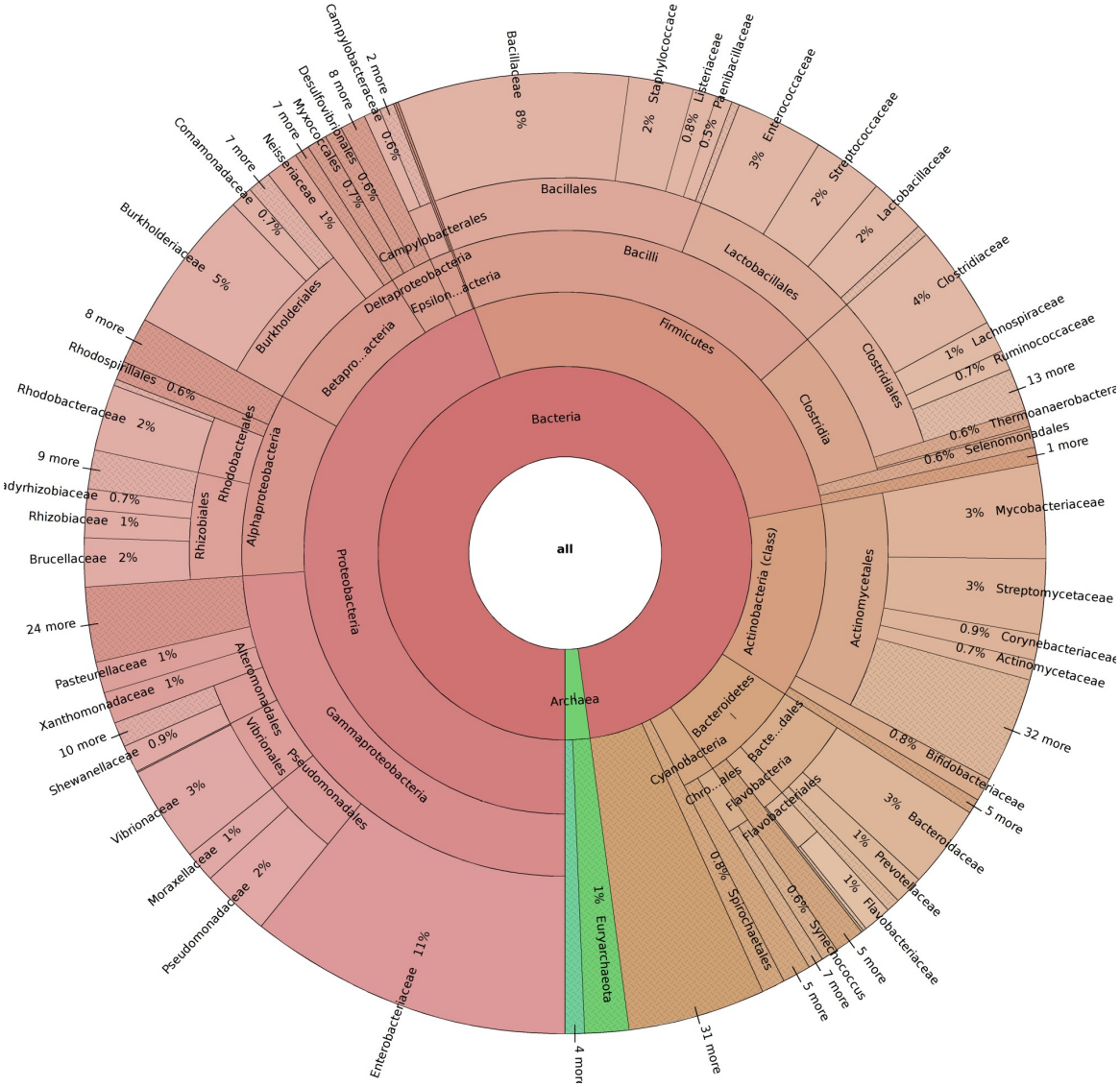
Supplementary Figure S3 - 16S gene assignment with taxator-tk **(h) new phylum scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	569	0	0	100.0	0.0	1	100.0	0.0	1	12459	0	100.0	root+superkingdom
superkingdom	1	5945	0	0	100.0	0.0	1	58.5	35.3	2				
phylum	2	0	391	0	0.0	0.0	5	0.0	0.0	32				
class	3	0	219	0	0.0	0.0	8	0.0	0.0	52	0	626	0.0	phylum+class+order
order	4	0	16	0	0.0	0.0	7	0.0	0.0	109				
family	5	0	8	0	0.0	0.0	5	0.0	0.0	235				
genus	6	0	27	0	0.0	0.0	2	0.0	0.0	615	0	35	0.0	family+genus+species
species	7				nan	nan	0	0.0	0.0	1416				
avg./sum	1.1	5945	661	0	16.7	0.0	4.0	8.4	5.0	351.6			90.0	all but unassigned
avg./sum	1.1	6514	661	0	28.6	0.0	3.6	19.8	4.4	307.8			90.8	all with unassigned

taxator-tk on RefSeq 16S genes

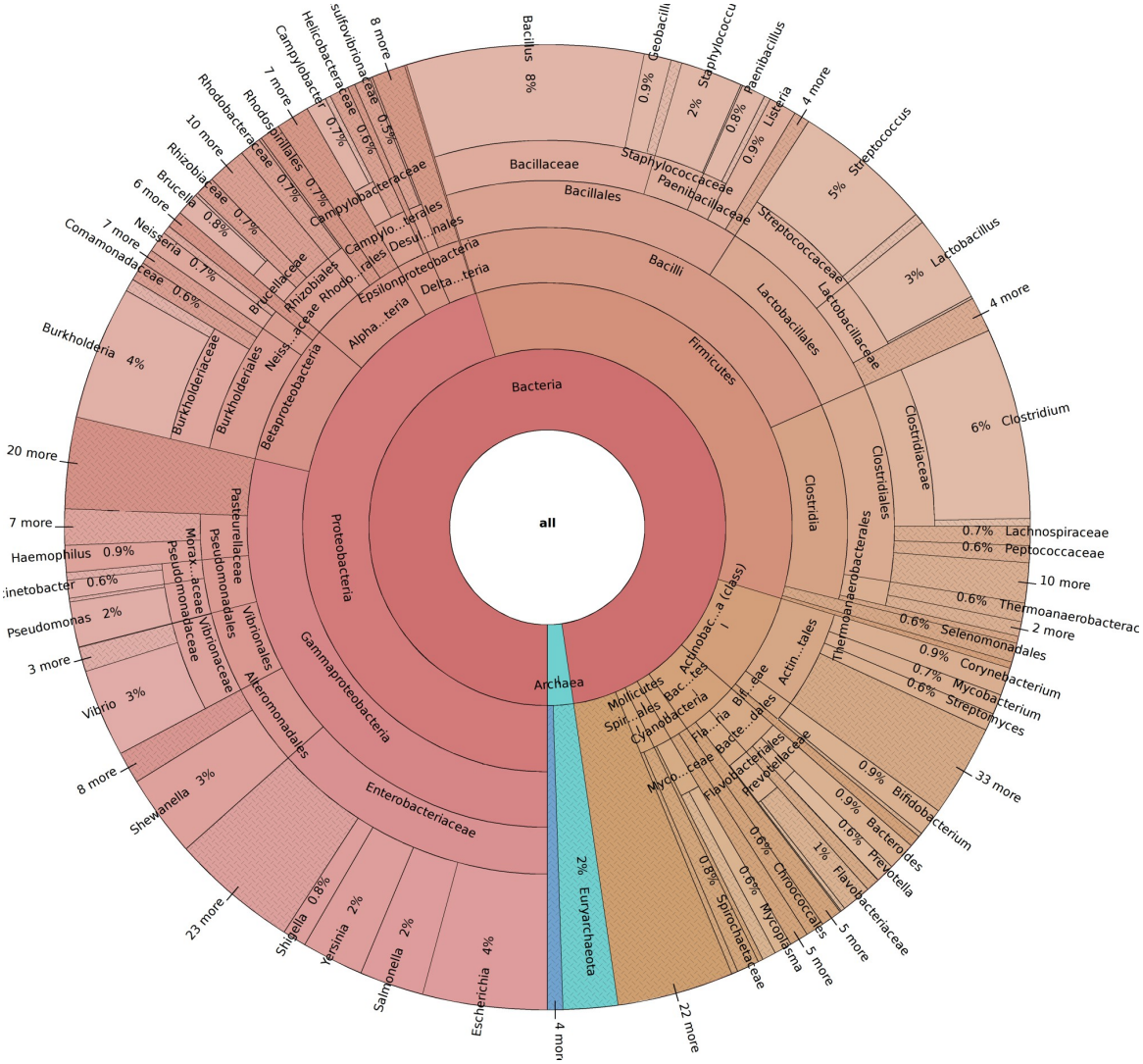


Supplementary Figure S4: Taxonomic composition of microbial RefSeq 47



Taxonomic composition down to family level of the microbial (bacteria, archaea and viruses) portion of the *RefSeq47* sequence data collection using Krona (Ondov et *al*., 2011). An interactive version can be found in the supplementary files (RefSeq47.krona.html). Abundance is measured in terms of accumulated sequence lengths per clade.

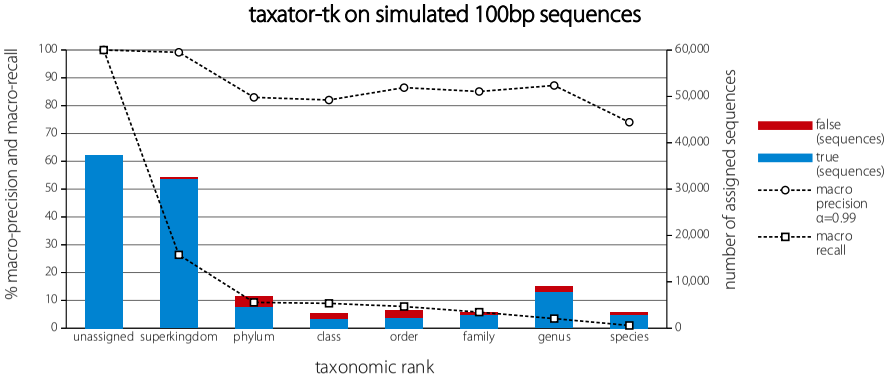
Supplementary Figure S5: Taxonomic composition of 16S genes extracted from *RefSeq47*



Taxonomic composition down to genus level of the 16S benchmark dataset using Krona (Ondov et *al*., 2011). The dataset was simulated by extracting every annotated 16S gene in *RefSeq47* which was at least 1000 bp long. An interactive version can be found in the supplementary files (refseq-16S.krona.html). Abundance is measured as the number of 16S genes.

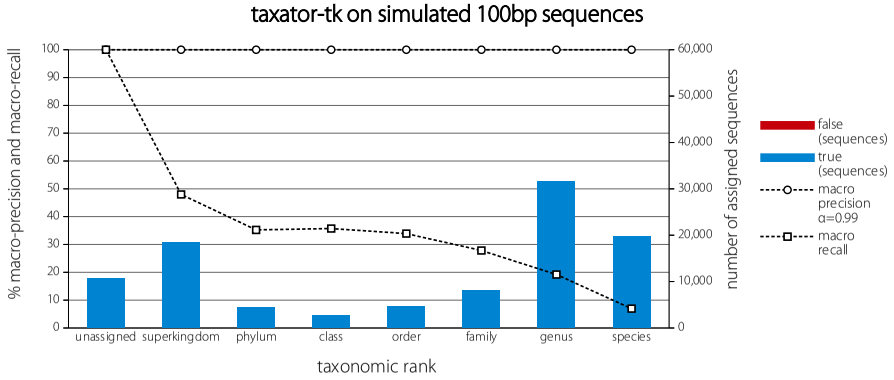
Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(a) summary scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	37391.6	0.0	0	100.0	0.0	1	100.0	0.0	1	101937.3	427.3	99.6	root+superkingdom
superkingdom	1	32272.9	427.3	0	99.2	0.0	1	26.4	26.7	3				
phylum	2	4563.7	2340.3	0	83.0	10.2	11	9.3	8.5	32	8977.7	4995.4	64.2	phylum+class+order
class	3	2164.1	1120.1	0	82.0	13.1	23	8.9	7.6	52				
order	4	2249.9	1535.0	0	86.5	11.1	52	7.8	7.2	110				
family	5	2859.3	591.9	0	85.1	14.7	98	5.8	6.8	240	13520.1	2415.0	84.8	family+genus+species
genus	6	7852.3	1275.7	0	87.3	17.6	202	3.5	5.6	656				
species	7	2808.6	547.4	0	74.0	34.8	431	1.0	2.6	1697				
avg/sum	2.4	54770.7	7837.7	0	85.3	14.5	116.9	8.9	9.3	398.6			87.5	all but unassigned
avg/sum	1.5	92162.3	7837.7	0	87.1	12.7	102.4	20.3	8.1	348.9			92.2	all with unassigned



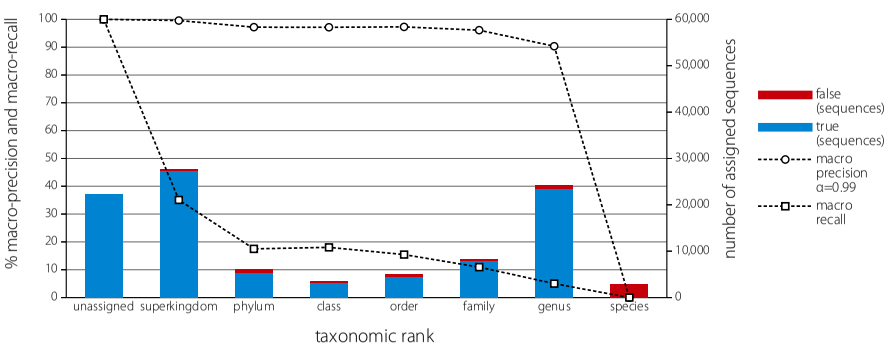
Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(b) all reference scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	10662	0	0	100.0	0.0	1	100.0	0.0	1	47620	0	100.0	root+superkingdom
superkingdom	1	18479	0	0	100.0	0.0	2	48.0	37.0	3				
phylum	2	4362	0	0	100.0	0.0	12	35.2	28.1	32	11598	0	100.0	phylum+class+order
class	3	2607	0	0	100.0	0.0	24	35.7	27.1	52				
order	4	4629	0	0	100.0	0.0	54	33.9	28.2	110				
family	5	8015	0	0	100.0	0.0	104	27.8	29.2	240	59261	0	100.0	family+genus+species
genus	6	31586	0	0	100.0	0.0	211	19.2	28.2	656				
species	7	19660	0	0	100.0	0.0	365	6.9	18.2	1697				
avg/sum	4.1	89338	0	0	100.0	0.0	110.3	29.5	28.0	398.6			100.0	all but unassigned
avg/sum	3.5	100000	0	0	100.0	0.0	96.6	38.4	24.5	348.9			100.0	all with unassigned



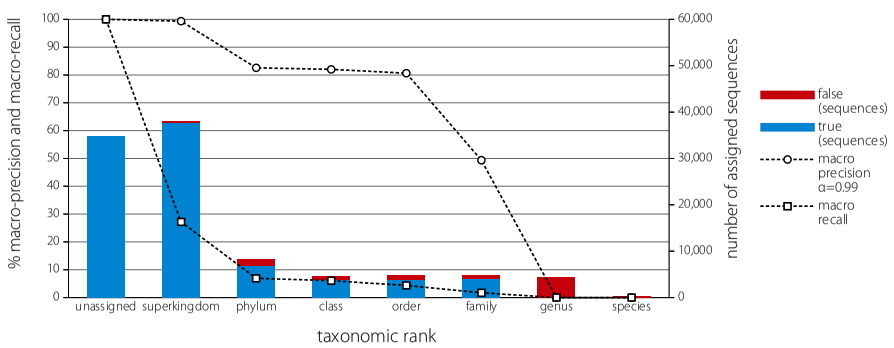
Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(c) new species scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	22319	0	0	100.0	0.0	1	100.0	0.0	1	76901	252	99.7	root+superkingdom
superkingdom	1	27291	252	0	99.6	0.0	1	35.1	32.5	3				
phylum	2	5362	746	0	97.2	1.8	10	17.5	18.3	32	13213	1541	89.6	phylum+class+order
class	3	3240	327	0	97.2	2.9	22	18.1	19.1	52				
order	4	4611	468	0	97.3	3.3	45	15.4	18.5	110				
family	5	7973	255	0	96.1	6.7	75	10.9	18.3	240	31353	4031	88.6	family+genus+species
genus	6	23380	776	0	90.4	21.6	100	5.0	14.2	656				
species	7	0	3000	0	0.0	0.0	217	0.0	0.0	1697				
avg/sum	3.4	71857	5824	0	82.5	5.2	67.1	14.6	17.3	398.6			92.5	all but unassigned
avg/sum	2.6	94176	5824	0	84.7	4.5	58.9	25.3	15.1	348.9			94.2	all with unassigned



Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(d) new genus scenario**

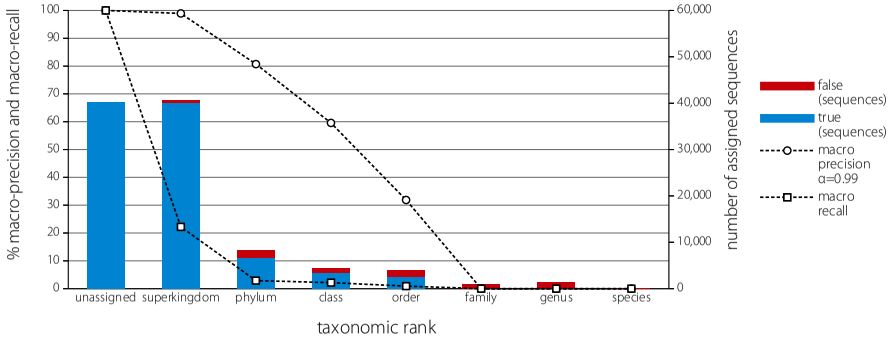
rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	34909	0	0	100.0	0.0	1	100.0	0.0	1	110459	343	99.7	root+superkingdom
superkingdom	1	37775	343	0	99.3	0.0	1	27.2	28.0	3				
phylum	2	6814	1406	0	82.6	22.0	8	6.9	9.4	32	14566	3057	82.7	phylum+class+order
class	3	3906	689	0	82.0	17.8	19	6.1	7.9	52				
order	4	3846	962	0	80.6	17.2	44	4.4	7.5	110				
family	5	4027	657	0	49.4	39.3	77	1.7	5.3	240	4027	5323	43.1	family+genus+species
genus	6	0	4422	0	0.0	0.0	193	0.0	0.0	656				
species	7	0	244	0	0.0	0.0	103	0.0	0.0	1697				
avg/sum	2.1	56368	8723	0	56.3	13.8	63.6	6.6	8.3	398.6			86.6	all but unassigned
avg/sum	1.4	91277	8723	0	61.7	12.0	55.8	18.3	7.3	348.9			91.3	all with unassigned



Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(e) new family scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	40215	0	0	100.0	0.0	1	100.0	0.0	1	120367	525	99.6	root+superkingdom
superkingdom	1	40076	525	0	98.9	0.0	1	22.2	27.0	3				
phylum	2	6632	1627	0	80.7	7.2	6	2.9	5.6	32	12720	3840	76.8	phylum+class+order
class	3	3425	904	0	59.6	27.6	14	2.2	3.9	52				
order	4	2663	1309	0	31.9	33.3	43	0.9	2.5	110				
family	5	0	1045	0	0.0	0.0	120	0.0	0.0	240	0	2624	0.0	family+genus+species
genus	6	0	1413	0	0.0	0.0	133	0.0	0.0	656				
species	7	0	166	0	0.0	0.0	80	0.0	0.0	1697				
avg/sum	1.7	52796	6989	0	38.7	9.7	56.7	4.0	5.6	398.6			88.3	all but unassigned
avg/sum	1.0	93011	6989	0	46.4	8.5	49.8	16.0	4.9	348.9			93.0	all with unassigned

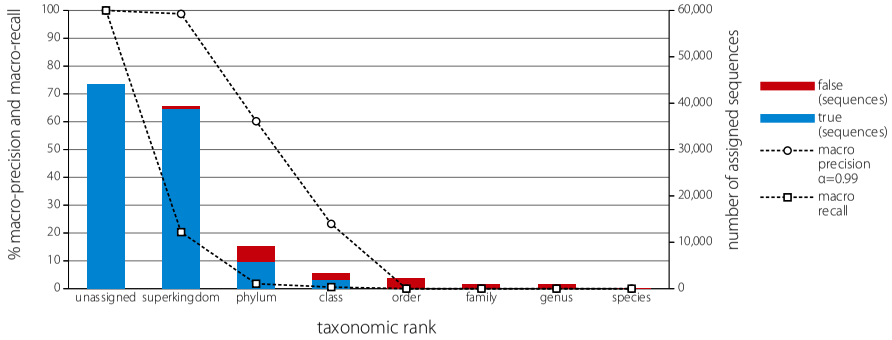
taxator-tk on simulated 100bp sequences



Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(f) new order scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	44037	0	0	100.0	0.0	1	100.0	0.0	1	121439	563	99.5	root+superkingdom
superkingdom	1	38701	563	0	98.8	0.0	1	20.3	25.5	3				
phylum	2	5817	3314	0	60.2	21.5	6	1.8	4.0	32	7788	6917	53.0	phylum+class+order
class	3	1971	1414	0	23.3	23.8	18	0.6	1.5	52				
order	4	0	2189	0	0.0	0.0	49	0.0	0.0	110				
family	5	0	961	0	0.0	0.0	106	0.0	0.0	240	0	1994	0.0	family+genus+species
genus	6	0	873	0	0.0	0.0	118	0.0	0.0	656				
species	7	0	160	0	0.0	0.0	72	0.0	0.0	1697				
avg/sum	1.5	46489	9474	0	26.0	6.5	52.9	3.2	4.4	398.6			83.1	all but unassigned
avg/sum	0.9	90526	9474	0	35.3	5.7	46.4	15.3	3.9	348.9			90.5	all with unassigned

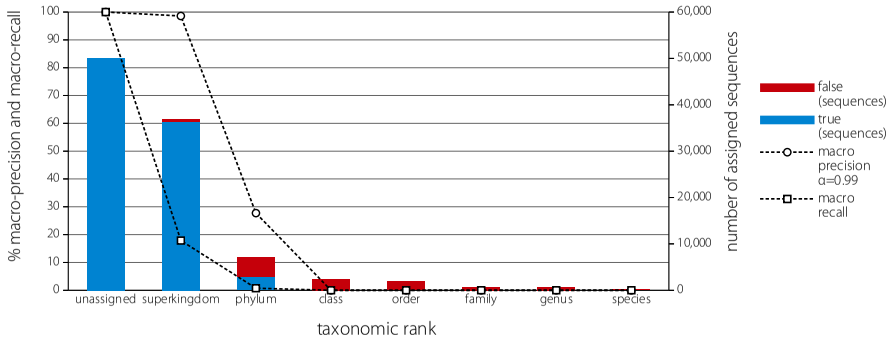
taxator-tk on simulated 100bp sequences



Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(g) new class scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	50150	0	0	100.0	0.0	1	100.0	0.0	1	122726	579	99.5	root+superkingdom
superkingdom	1	36288	579	0	98.6	0.0	1	17.9	22.8	3				
phylum	2	2959	4203	0	27.7	24.4	7	0.7	2.0	32	2959	8462	25.9	phylum+class+order
class	3	0	2365	0	0.0	0.0	20	0.0	0.0	52				
order	4	0	1894	0	0.0	0.0	49	0.0	0.0	110				
family	5	0	692	0	0.0	0.0	100	0.0	0.0	240	0	1562	0.0	family+genus+species
genus	6	0	742	0	0.0	0.0	109	0.0	0.0	656				
species	7	0	128	0	0.0	0.0	63	0.0	0.0	1697				
avg/sum	1.5	39247	10603	0	18.0	3.5	49.9	2.7	3.5	398.6			78.7	all but unassigned
avg/sum	0.7	89397	10603	0	28.3	3.0	43.8	14.8	3.1	348.9			89.4	all with unassigned

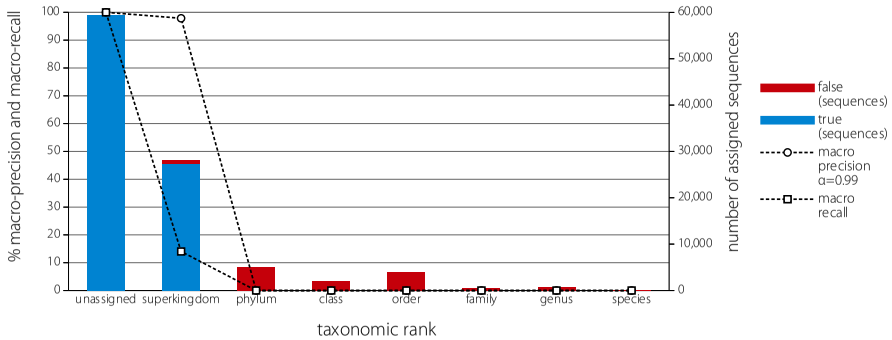
taxator-tk on simulated 100bp sequences



Supplementary Figure S6 - Simulated 100 bp sequence assignment with taxator-tk **(h) new phylum scenario**

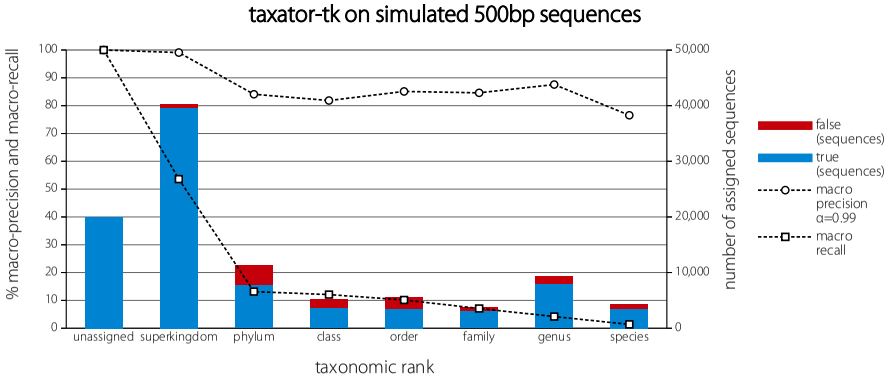
rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	59449	0	0	100.0	0.0	1	100.0	0.0	1	114049	729	99.4	root+superkingdom
superkingdom	1	27300	729	0	97.9	0.0	1	14.0	18.6	3				
phylum	2	0	5086	0	0.0	0.0	11	0.0	0.0	32	0	11151	0.0	phylum+class+order
class	3	0	2142	0	0.0	0.0	20	0.0	0.0	52				
order	4	0	3923	0	0.0	0.0	44	0.0	0.0	110				
family	5	0	533	0	0.0	0.0	98	0.0	0.0	240	0	1371	0.0	family+genus+species
genus	6	0	704	0	0.0	0.0	105	0.0	0.0	656				
species	7	0	134	0	0.0	0.0	60	0.0	0.0	1697				
avg/sum	1.7	27300	13251	0	14.0	0.0	48.4	2.0	2.7	398.6			67.3	all but unassigned
avg/sum	0.7	86749	13251	0	24.7	0.0	42.5	14.3	2.3	348.9			86.7	all with unassigned

taxator-tk on simulated 100bp sequences



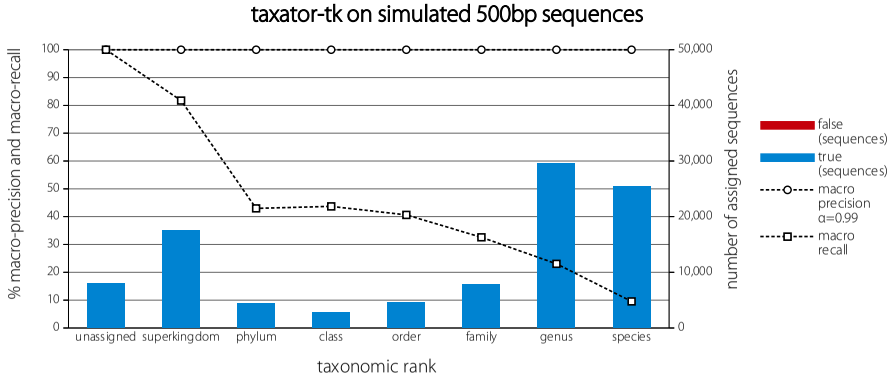
Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(a) summary scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	20001.1	0.0	0	100.0	0.0	1	100.0	0.0	1	99200.0	582.0	99.4	root+superkingdom
superkingdom	1	39599.4	582.0	0	99.1	0.0	1	53.6	26.8	2				
phylum	2	7862.7	3532.7	0	84.1	12.4	12	13.2	11.4	32	15065.4	7226.3	67.6	phylum+class+order
class	3	3756.1	1555.3	0	81.8	14.8	24	12.1	9.4	52				
order	4	3446.6	2138.3	0	85.1	13.1	56	10.2	8.6	110				
family	5	3162.4	702.6	0	84.6	17.2	104	7.1	7.8	240	14666.0	2859.7	83.7	family+genus+species
genus	6	7880.9	1428.4	0	87.6	19.3	212	4.2	6.3	656				
species	7	3622.7	728.7	0	76.5	34.0	480	1.4	3.4	1693				
avg/sum	2.3	69330.9	10668.0	0	85.6	15.8	127.0	14.5	10.5	397.9			86.7	all but unassigned
avg/sum	1.8	89332.0	10668.0	0	87.4	13.8	111.3	25.2	9.2	348.3			89.3	all with unassigned



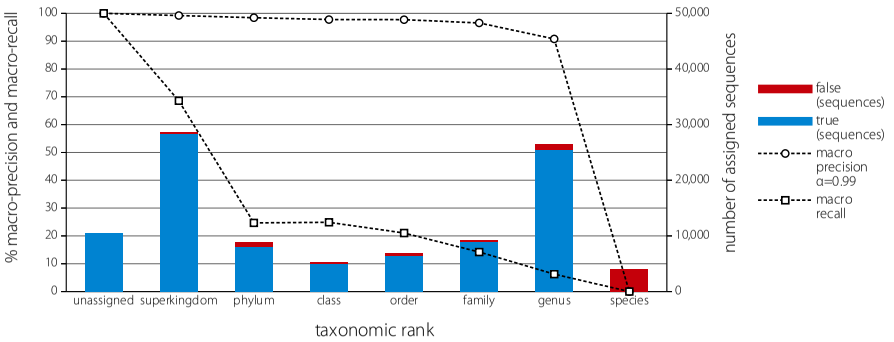
Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(b) all reference scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	7999	0	0	100.0	0.0	1	100.0	0.0	1	42883	0	100.0	root+superkingdom
superkingdom	1	17442	0	0	100.0	0.0	2	81.8	10.7	2				
phylum	2	4415	0	0	100.0	0.0	14	43.0	30.7	32	11750	0	100.0	phylum+class+order
class	3	2699	0	0	100.0	0.0	27	43.7	29.3	52				
order	4	4636	0	0	100.0	0.0	59	40.6	30.9	110				
family	5	7889	0	0	100.0	0.0	109	32.6	31.8	240	62809	0	100.0	family+genus+species
genus	6	29561	0	0	100.0	0.0	221	23.0	32.0	656				
species	7	25359	0	0	100.0	0.0	408	9.5	23.5	1693				
avg/sum	4.0	92001	0	0	100.0	0.0	120.0	39.2	27.0	397.9			100.0	all but unassigned
avg/sum	3.6	100000	0	0	100.0	0.0	105.1	46.8	23.6	348.3			100.0	all with unassigned



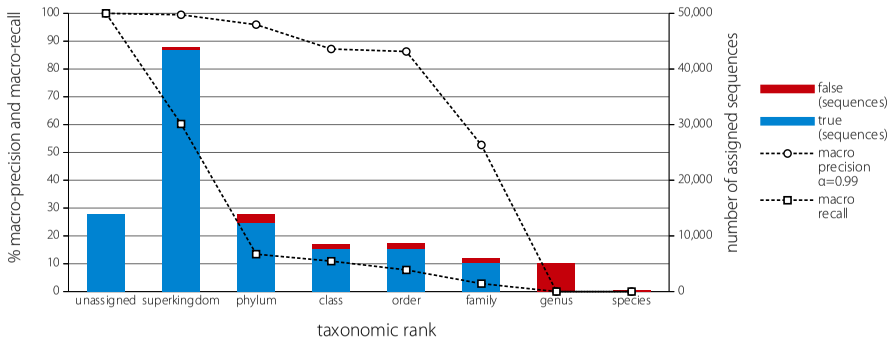
Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(c) new species scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	10520	0	0	100.0	0.0	1	100.0	0.0	1	67276	224	99.7	root+superkingdom
superkingdom	1	28378	224	0	99.2	0.5	2	68.6	21.6	2				
phylum	2	8027	773	0	98.4	1.2	11	24.7	23.6	32	19494	1591	92.5	phylum+class+order
class	3	4991	337	0	97.8	2.6	23	24.9	23.3	52				
order	4	6476	481	0	97.7	2.8	50	21.0	22.3	110				
family	5	9009	253	0	96.6	6.6	79	14.2	21.7	240	34614	5179	87.0	family+genus+species
genus	6	25605	910	0	90.8	21.1	107	6.3	16.4	656				
species	7	0	4016	0	0.0	0.0	237	0.0	0.0	1693				
avg/sum	3.5	82486	6994	0	82.9	5.0	72.7	22.8	18.4	397.9			92.2	all but unassigned
avg/sum	3.1	93006	6994	0	85.1	4.3	63.8	32.5	16.1	348.3			93.0	all with unassigned



Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(d) new genus scenario**

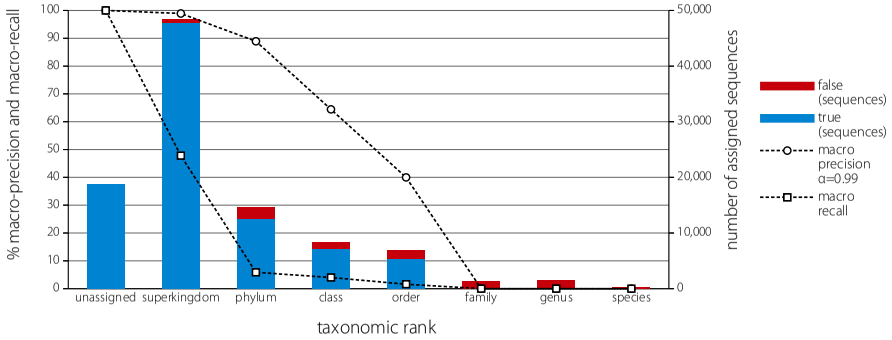
rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	13884	0	0	100.0	0.0	1	100.0	0.0	1	100784	357	99.6	root+superkingdom
superkingdom	1	43450	357	0	99.5	0.0	1	60.2	26.6	2				
phylum	2	12278	1548	0	95.9	2.9	9	13.4	15.2	32	27611	3340	89.2	phylum+class+order
class	3	7723	760	0	87.2	13.6	21	10.9	12.4	52				
order	4	7610	1032	0	86.3	12.4	47	7.8	11.2	110				
family	5	5239	761	0	52.7	40.4	81	2.8	7.9	240	5239	6119	46.1	family+genus+species
genus	6	0	5064	0	0.0	0.0	136	0.0	0.0	656				
species	7	0	294	0	0.0	0.0	105	0.0	0.0	1693				
avg/sum	2.2	76300	9816	0	60.2	9.9	57.1	13.6	10.5	397.9			88.6	all but unassigned
avg/sum	1.9	90184	9816	0	65.2	8.7	50.1	24.4	9.2	348.3			90.2	all with unassigned



Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(e) new family scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	18731	0	0	100.0	0.0	1	100.0	0.0	1	114197	702	99.4	root+superkingdom
superkingdom	1	47733	702	0	98.9	0.0	1	47.8	34.0	2				
phylum	2	12581	2006	0	88.9	4.2	6	5.9	10.5	32	25162	4675	84.3	phylum+class+order
class	3	7177	1099	0	64.4	31.5	16	4.0	6.7	52				
order	4	5404	1570	0	40.0	36.6	42	1.6	4.2	110				
family	5	0	1251	0	0.0	0.0	103	0.0	0.0	240	0	2997	0.0	family+genus+species
genus	6	0	1492	0	0.0	0.0	122	0.0	0.0	656				
species	7	0	254	0	0.0	0.0	82	0.0	0.0	1693				
avg/sum	1.8	72895	8374	0	41.8	10.3	53.1	8.5	7.9	397.9			89.7	all but unassigned
avg/sum	1.5	91626	8374	0	49.0	9.0	46.6	19.9	6.9	348.3			91.6	all with unassigned

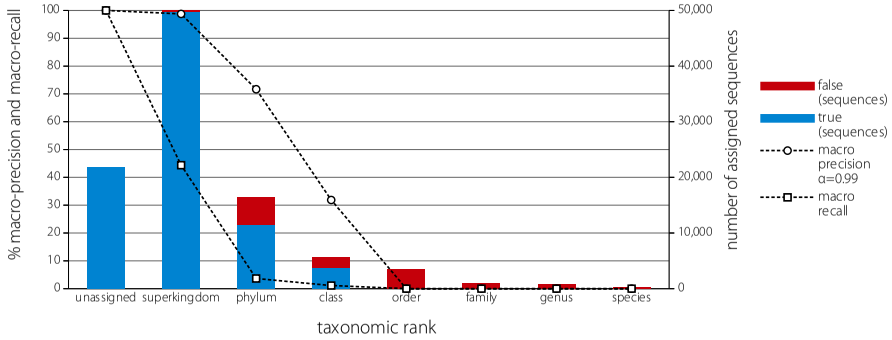
taxator-tk on simulated 500bp sequences



Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(f) new order scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	21881	0	0	100.0	0.0	1	100.0	0.0	1	121329	771	99.4	root+superkingdom
superkingdom	1	49724	771	0	98.8	0.0	1	44.4	34.2	2				
phylum	2	11475	4991	0	71.7	17.0	6	3.6	7.8	32	15178	10362	59.4	phylum+class+order
class	3	3703	1888	0	31.9	29.3	18	1.1	2.5	52				
order	4	0	3483	0	0.0	0.0	56	0.0	0.0	110				
family	5	0	1003	0	0.0	0.0	106	0.0	0.0	240	0	2084	0.0	family+genus+species
genus	6	0	868	0	0.0	0.0	110	0.0	0.0	656				
species	7	0	213	0	0.0	0.0	63	0.0	0.0	1693				
avg/sum	1.6	64902	13217	0	28.9	6.6	51.4	7.0	6.4	397.9			83.1	all but unassigned
avg/sum	1.2	86783	13217	0	37.8	5.8	45.1	18.6	5.6	348.3			86.8	all with unassigned

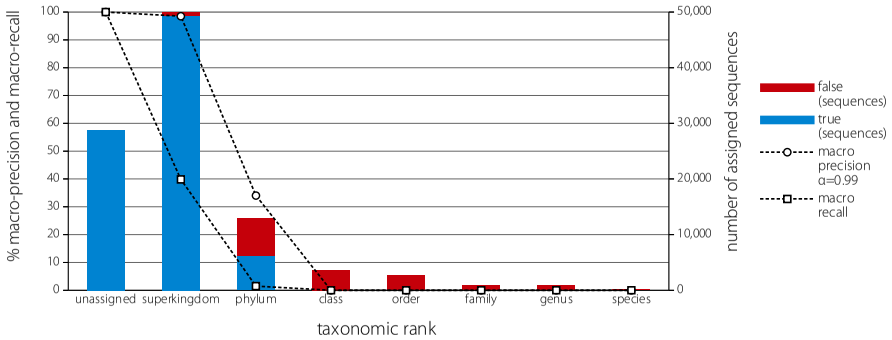
taxator-tk on simulated 500bp sequences



Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(g) new class scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	28770	0	0	100.0	0.0	1	100.0	0.0	1	127418	838	99.3	root+superkingdom
superkingdom	1	49324	838	0	98.5	0.0	1	39.9	31.6	2				
phylum	2	6263	6679	0	34.0	28.9	7	1.5	3.8	32	6263	12967	32.6	phylum+class+order
class	3	0	3676	0	0.0	0.0	21	0.0	0.0	52				
order	4	0	2612	0	0.0	0.0	52	0.0	0.0	110				
family	5	0	852	0	0.0	0.0	108	0.0	0.0	240	0	1838	0.0	family+genus+species
genus	6	0	834	0	0.0	0.0	103	0.0	0.0	656				
species	7	0	152	0	0.0	0.0	56	0.0	0.0	1693				
avg/sum	1.5	55587	15643	0	18.9	4.1	49.7	5.9	5.1	397.9			78.0	all but unassigned
avg/sum	1.1	84357	15643	0	29.1	3.6	43.6	17.7	4.4	348.3			84.4	all with unassigned

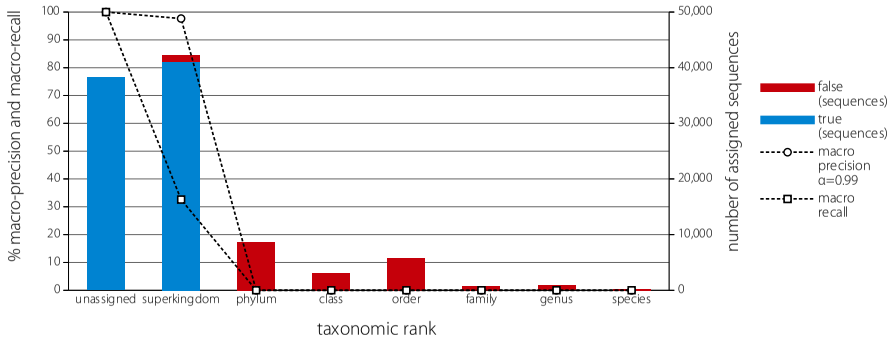
taxator-tk on simulated 500bp sequences



Supplementary Figure S7 - Simulated 500 bp sequence assignment with taxator-tk **(h) new phylum scenario**

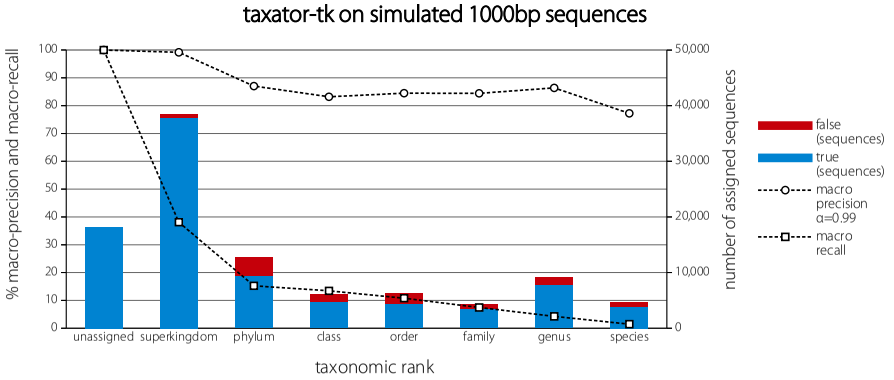
rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	38223	0	0	100.0	0.0	1	100.0	0.0	1	120513	1182	99.0	root+superkingdom
superkingdom	1	41145	1182	0	97.7	0.0	1	32.6	28.9	2				
phylum	2	0	8732	0	0.0	0.0	11	0.0	0.0	32	0	17649	0.0	phylum+class+order
class	3	0	3127	0	0.0	0.0	22	0.0	0.0	52				
order	4	0	5790	0	0.0	0.0	44	0.0	0.0	110				
family	5	0	798	0	0.0	0.0	106	0.0	0.0	240	0	1801	0.0	family+genus+species
genus	6	0	831	0	0.0	0.0	93	0.0	0.0	656				
species	7	0	172	0	0.0	0.0	45	0.0	0.0	1693				
avg/sum	1.6	41145	20632	0	14.0	0.0	46.0	4.7	4.1	397.9			66.6	all but unassigned
avg/sum	1.0	79368	20632	0	24.7	0.0	40.4	16.6	3.6	348.3			79.4	all with unassigned

taxator-tk on simulated 500bp sequences



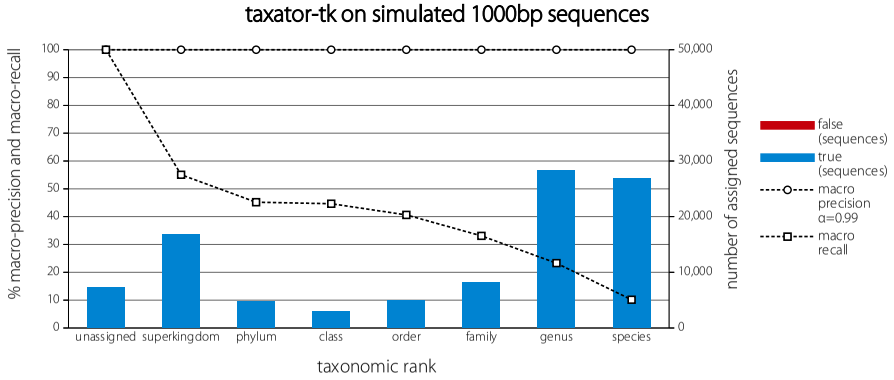
Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(a) summary scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	18217.3	0.0	0	100.0	0.0	1	100.0	0.0	1	93809.9	550.7	99.4	root+superkingdom
superkingdom	1	37796.3	550.7	0	99.2	0.0	1	38.1	33.8	3				
phylum	2	9465.1	3300.7	0	87.0	12.2	12	15.2	12.7	32	18677.1	6634.6	73.8	phylum+class+order
class	3	4795.0	1367.1	0	83.2	14.7	25	13.4	10.3	52				
order	4	4417.0	1966.7	0	84.5	15.1	57	10.8	9.2	110				
family	5	3498.1	817.1	0	84.4	17.9	106	7.5	8.1	240	15169.9	2954.1	83.7	family+genus+species
genus	6	7834.3	1397.6	0	86.4	19.6	219	4.3	6.3	653				
species	7	3837.4	739.4	0	77.2	34.2	472	1.5	3.5	1690				
avg/sum	2.4	71643.3	10139.4	0	86.0	16.2	127.4	13.0	12.0	397.1			87.6	all but unassigned
avg/sum	1.9	89860.6	10139.4	0	87.7	14.2	111.6	23.8	10.5	347.6			89.9	all with unassigned



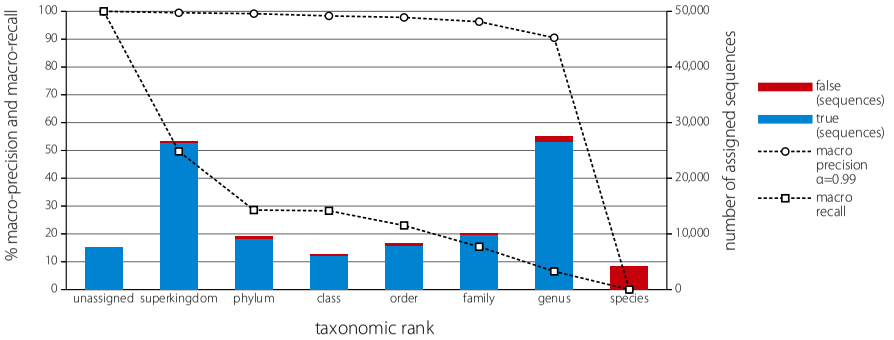
Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(b) all reference scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	7256	0	0	100.0	0.0	1	100.0	0.0	1	40990	0	100.0	root+superkingdom
superkingdom	1	16867	0	0	100.0	0.0	2	55.1	39.9	3				
phylum	2	4739	0	0	100.0	0.0	14	45.2	30.5	32	12595	0	100.0	phylum+class+order
class	3	3024	0	0	100.0	0.0	27	44.7	29.7	52				
order	4	4832	0	0	100.0	0.0	59	40.6	31.1	110				
family	5	8132	0	0	100.0	0.0	112	33.1	32.0	240	63282	0	100.0	family+genus+species
genus	6	28288	0	0	100.0	0.0	221	23.3	32.2	653				
species	7	26862	0	0	100.0	0.0	402	10.1	24.8	1690				
avg/sum	4.0	92744	0	0	100.0	0.0	119.6	36.0	31.5	397.1			100.0	all but unassigned
avg/sum	3.6	100000	0	0	100.0	0.0	104.8	44.0	27.5	347.6			100.0	all with unassigned



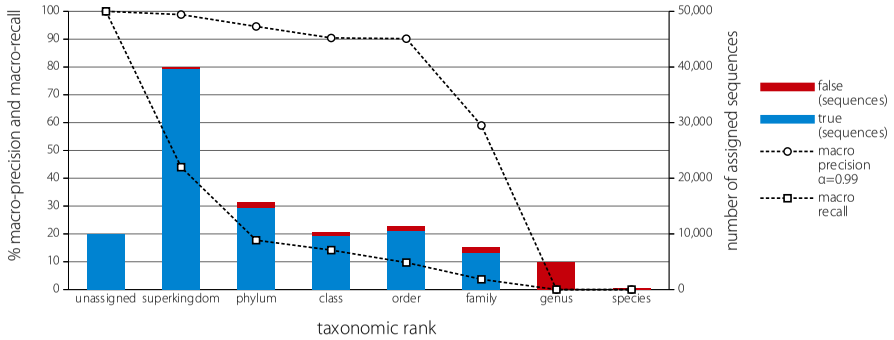
Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(c) new species scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	7557	0	0	100.0	0.0	1	100.0	0.0	1	60223	191	99.7	root+superkingdom
superkingdom	1	26333	191	0	99.5	0.2	2	49.7	38.2	3				
phylum	2	9128	523	0	99.2	0.6	12	28.6	25.3	32	23034	1142	95.3	phylum+class+order
class	3	6028	250	0	98.4	2.4	24	28.3	25.2	52				
order	4	7878	369	0	97.8	4.4	52	23.0	23.6	110				
family	5	9803	325	0	96.3	9.0	83	15.4	22.6	240	36355	5388	87.1	family+genus+species
genus	6	26552	898	0	90.5	22.4	107	6.5	16.7	653				
species	7	0	4165	0	0.0	0.0	230	0.0	0.0	1690				
avg/sum	3.5	85722	6721	0	83.1	5.6	72.9	21.6	21.7	397.1			92.7	all but unassigned
avg/sum	3.3	93279	6721	0	85.2	4.9	63.9	31.4	19.0	347.6			93.3	all with unassigned



Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(d) new genus scenario**

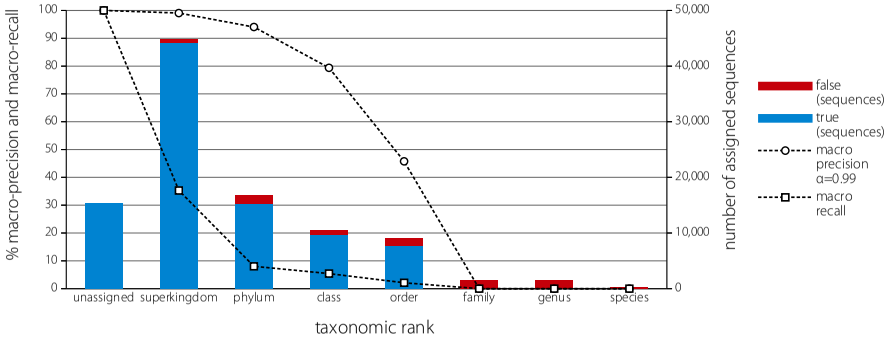
rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	9974	0	0	100.0	0.0	1	100.0	0.0	1	89246	293	99.7	root+superkingdom
superkingdom	1	39636	293	0	98.8	0.8	2	44.0	37.1	3				
phylum	2	14673	1006	0	94.6	9.8	10	17.7	18.6	32	34900	2418	93.5	phylum+class+order
class	3	9782	538	0	90.4	11.6	22	14.2	15.4	52				
order	4	10445	874	0	90.2	10.8	47	9.7	13.5	110				
family	5	6552	955	0	59.0	39.9	82	3.7	9.4	240	6552	6227	51.3	family+genus+species
genus	6	0	4978	0	0.0	0.0	143	0.0	0.0	653				
species	7	0	294	0	0.0	0.0	94	0.0	0.0	1690				
avg/sum	2.4	81088	8938	0	61.9	10.4	57.1	12.8	13.4	397.1			90.1	all but unassigned
avg/sum	2.2	91062	8938	0	66.6	9.1	50.1	23.7	11.7	347.6			91.1	all with unassigned



Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(e) new family scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	15429	0	0	100.0	0.0	1	100.0	0.0	1	103635	644	99.4	root+superkingdom
superkingdom	1	44103	644	0	99.1	0.0	1	35.3	36.3	3				
phylum	2	15330	1458	0	94.1	2.4	6	8.0	13.7	32	32792	3711	89.8	phylum+class+order
class	3	9698	849	0	79.4	24.5	15	5.4	9.1	52				
order	4	7764	1404	0	45.8	38.6	40	2.1	5.6	110				
family	5	0	1548	0	0.0	0.0	104	0.0	0.0	240	0	3321	0.0	family+genus+species
genus	6	0	1507	0	0.0	0.0	127	0.0	0.0	653				
species	7	0	266	0	0.0	0.0	65	0.0	0.0	1690				
avg/sum	1.9	76895	7676	0	45.5	9.4	51.1	7.3	9.3	397.1			90.9	all but unassigned
avg/sum	1.6	92324	7676	0	52.3	8.2	44.9	18.9	8.1	347.6			92.3	all with unassigned

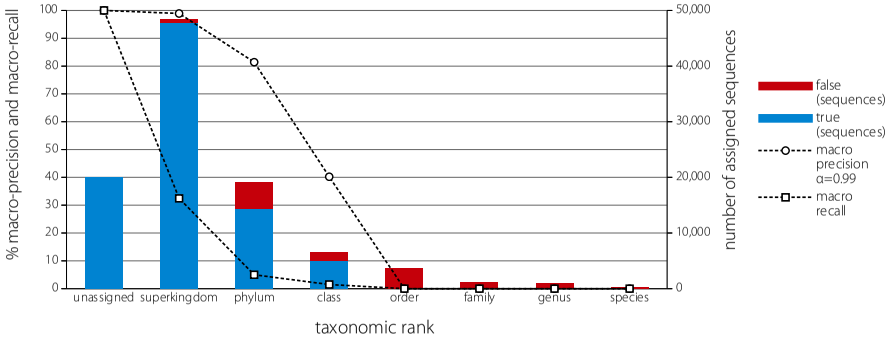
taxator-tk on simulated 1000bp sequences



Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(f) new order scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	19932	0	0	100.0	0.0	1	100.0	0.0	1	115360	725	99.4	root+superkingdom
superkingdom	1	47714	725	0	98.9	0.0	1	32.4	34.7	3				
phylum	2	14366	4764	0	81.4	11.9	6	5.0	10.2	32	19399	9995	66.0	phylum+class+order
class	3	5033	1596	0	40.2	33.9	17	1.5	3.5	52				
order	4	0	3635	0	0.0	0.0	49	0.0	0.0	110				
family	5	0	1133	0	0.0	0.0	80	0.0	0.0	240	0	2235	0.0	family+genus+species
genus	6	0	883	0	0.0	0.0	98	0.0	0.0	653				
species	7	0	219	0	0.0	0.0	50	0.0	0.0	1690				
avg/sum	1.7	67113	12955	0	31.5	6.5	43.0	5.6	6.9	397.1			83.8	all but unassigned
avg/sum	1.3	87045	12955	0	40.1	5.7	37.8	17.4	6.0	347.6			87.0	all with unassigned

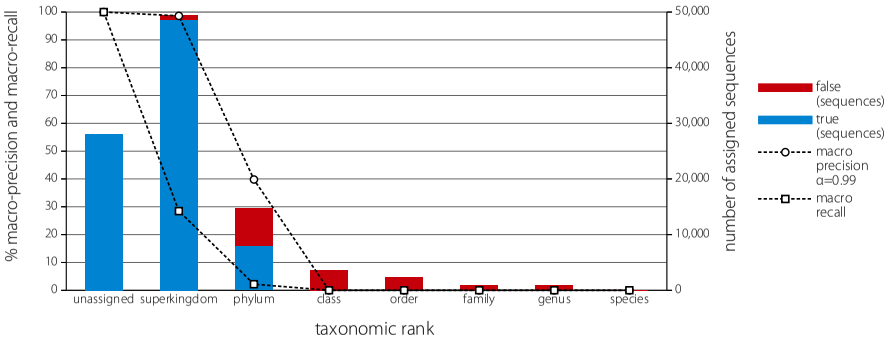
taxator-tk on simulated 1000bp sequences



Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(g) new class scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	28057	0	0	100.0	0.0	1	100.0	0.0	1	125231	824	99.3	root+superkingdom
superkingdom	1	48587	824	0	98.6	0.0	1	28.5	31.4	3				
phylum	2	8020	6696	0	39.8	34.0	7	2.2	5.1	32	8020	12645	38.8	phylum+class+order
class	3	0	3649	0	0.0	0.0	22	0.0	0.0	52				
order	4	0	2300	0	0.0	0.0	48	0.0	0.0	110				
family	5	0	939	0	0.0	0.0	94	0.0	0.0	240	0	1867	0.0	family+genus+species
genus	6	0	835	0	0.0	0.0	94	0.0	0.0	653				
species	7	0	93	0	0.0	0.0	38	0.0	0.0	1690				
avg/sum	1.5	56607	15336	0	19.8	4.9	43.4	4.4	5.2	397.1			78.7	all but unassigned
avg/sum	1.1	84664	15336	0	29.8	4.2	38.1	16.3	4.6	347.6			84.7	all with unassigned

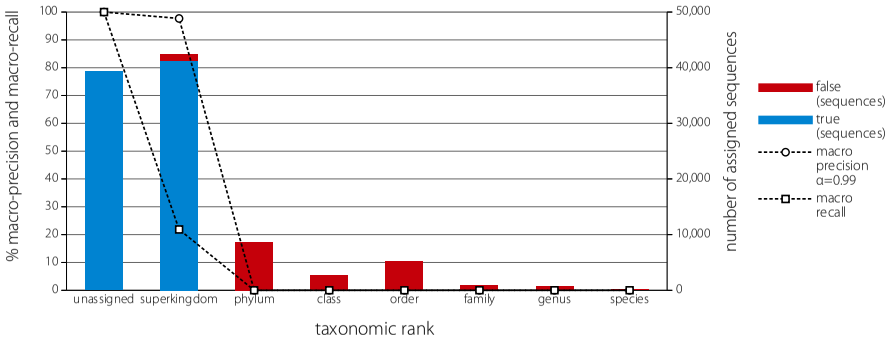
taxator-tk on simulated 1000bp sequences



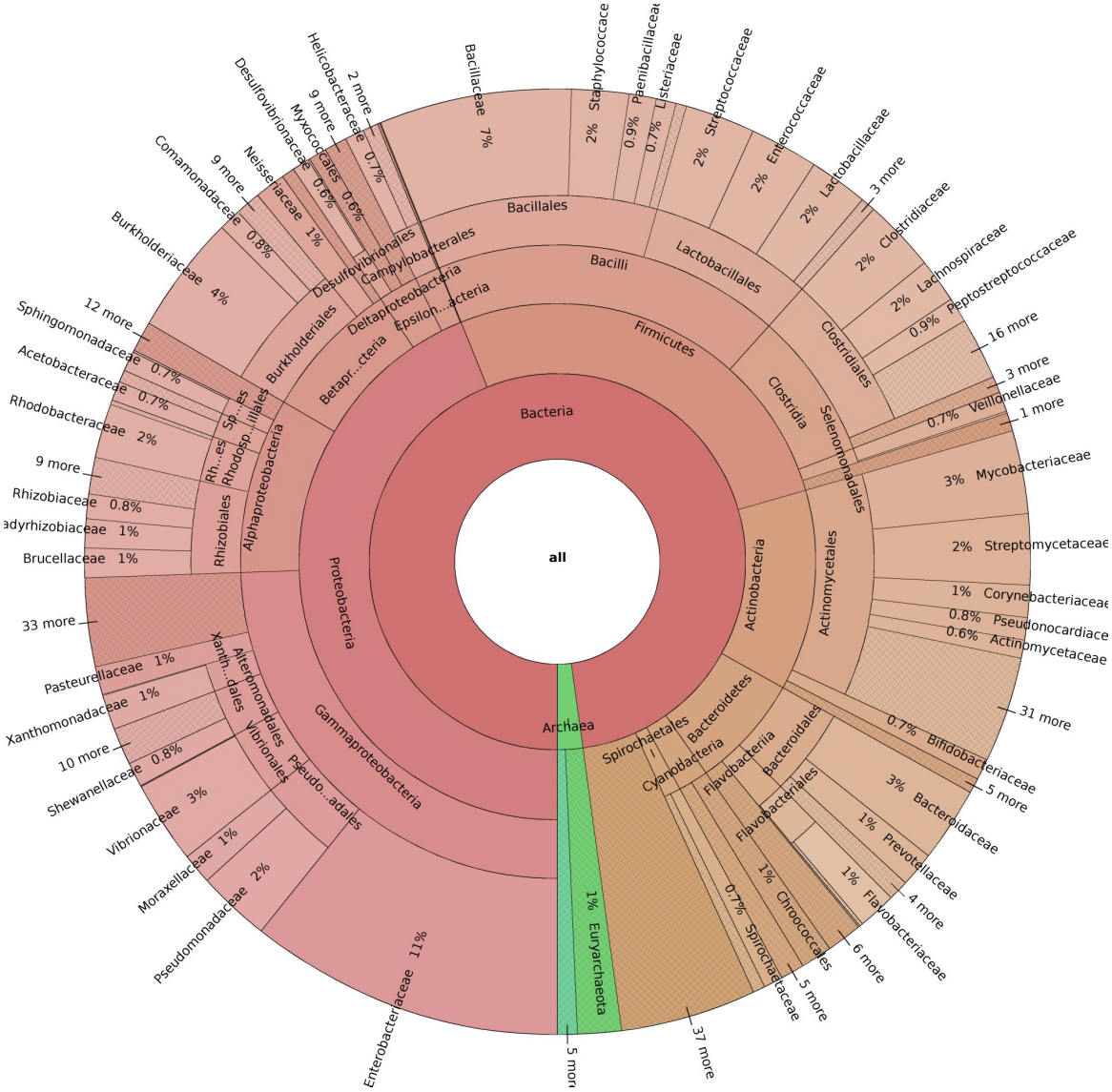
Supplementary Figure S8 - Simulated 1000 bp sequence assignment with taxator-tk **(h) new phylum scenario**

rank	depth	true (sequences)	false (sequences)	unknown (sequences)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (sequences)	sum false (sequences)	overall prec.	description
unassigned	0	39316	0	0	100.0	0.0	1	100.0	0.0	1	121984	1178	99.0	root+superkingdom
superkingdom	1	41334	1178	0	97.7	0.0	1	21.9	27.3	3				
phylum	2	0	8658	0	0.0	0.0	12	0.0	0.0	32	0	16531	0.0	phylum+class+order
class	3	0	2688	0	0.0	0.0	21	0.0	0.0	52				
order	4	0	5185	0	0.0	0.0	44	0.0	0.0	110				
family	5	0	820	0	0.0	0.0	98	0.0	0.0	240	0	1641	0.0	family+genus+species
genus	6	0	682	0	0.0	0.0	99	0.0	0.0	653				
species	7	0	139	0	0.0	0.0	41	0.0	0.0	1690				
avg/sum	1.6	41334	19350	0	14.0	0.0	45.1	3.1	3.9	397.1			68.1	all but unassigned
avg/sum	1.0	80650	19350	0	24.7	0.0	39.6	15.2	3.4	347.6			80.7	all with unassigned

taxator-tk on simulated 1000bp sequences

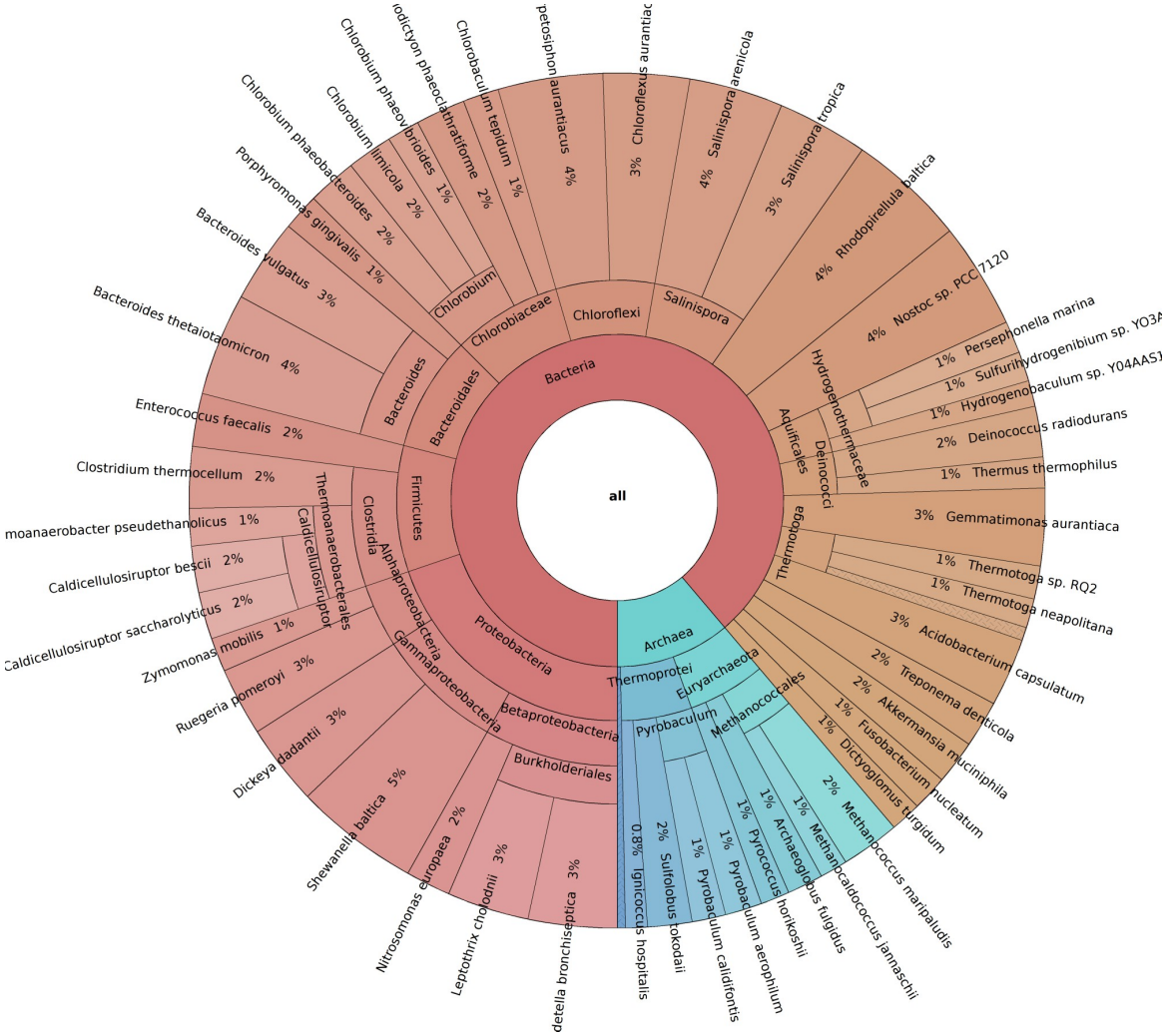


Supplementary Figure S9: Taxonomic composition of microbial *RefSeq54*



Taxonomic composition down to family level of the microbial (bacteria, archaea and viruses) portion of the *RefSeq54* sequence data collection using Krona (Ondov et *al*., 2011). An interactive version can be found in the supplementary files (RefSeq54.krona.html). Abundance is measured in terms of accumulated sequence lengths per clade.

Supplementary Figure S10: Taxonomic composition of simArt49e

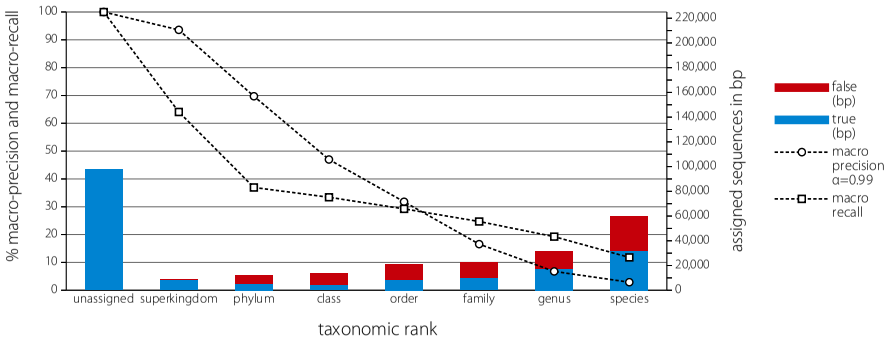


Taxonomic composition of the simulated metagenome sample simArt49e using Krona (Ondov et *al*., 2011). An interactive version can be found in the supplementary files (simArt49e.krona.html). Abundance is measured in terms of accumulated contigs lengths. The reads for this dataset were simulated using equal coverage for every strain, so differences in the data proportions result from a variable genome size and assembly bias.

Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(a) summary scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	97460.6	0.0	0	100.0	0.0	1	100.0	0.0	1	115012.6	500.1	99.6	root+superkingdom
superkingdom	1	8776.0	500.1	0	93.6	3.0	2	64.1	17.1	2				
phylum	2	5011.0	7085.1	0	69.7	22.7	20	36.9	15.2	20	17937.9	28325.0	38.8	phylum+class+order
class	3	4568.4	9153.1	0	47.0	38.5	36	33.4	11.8	23				
order	4	8358.4	12086.7	0	31.8	39.5	78	29.2	9.5	32				
family	5	10303.4	12858.1	0	16.6	33.6	176	24.7	7.1	36				
genus	6	17193.4	13745.6	0	6.7	24.1	553	19.3	5.1	41	59286.4	54892.0	51.9	family+genus+species
species	7	31789.6	28288.3	0	2.9	16.5	1672	11.8	2.6	49				
avg/sum	4.2	86000.3	83717.1	0	38.3	25.4	362.4	31.3	9.8	29.0			50.7	all but unassigned
avg/sum	2.2	183460.9	83717.1	0	46.0	22.2	317.3	39.9	8.5	25.5			68.7	all with unassigned

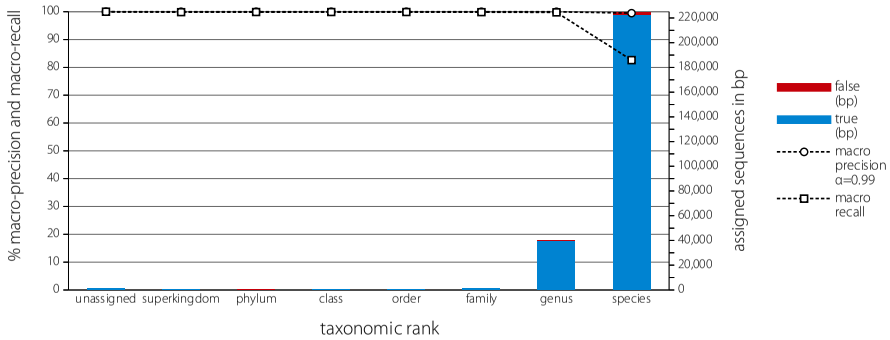
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(b) all reference scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	1071	0	0	100.0	0.0	1	100.0	0.0	1	1403	0	100.0	root+superkingdom
superkingdom	1	166	0	0	100.0	0.0	2	99.9	0.1	2				
phylum	2	130	1	0	100.0	0.0	19	99.9	0.2	20	852	1	99.9	phylum+class+order
class	3	108	0	0	100.0	0.0	23	99.9	0.2	23				
order	4	614	0	0	100.0	0.0	31	99.9	0.2	32				
family	5	1000	0	0	100.0	0.0	35	99.9	0.2	36				
genus	6	39774	30	0	100.0	0.0	40	99.8	0.2	41	263301	1787	99.3	family+genus+species
species	7	222527	1757	0	99.5	2.7	46	82.6	18.3	49				
avg/sum	5.9	264319	1788	0	99.9	0.4	28.0	97.4	2.8	29.0			99.3	all but unassigned
avg/sum	5.8	265390	1788	0	99.9	0.3	24.6	97.7	2.4	25.5			99.3	all with unassigned

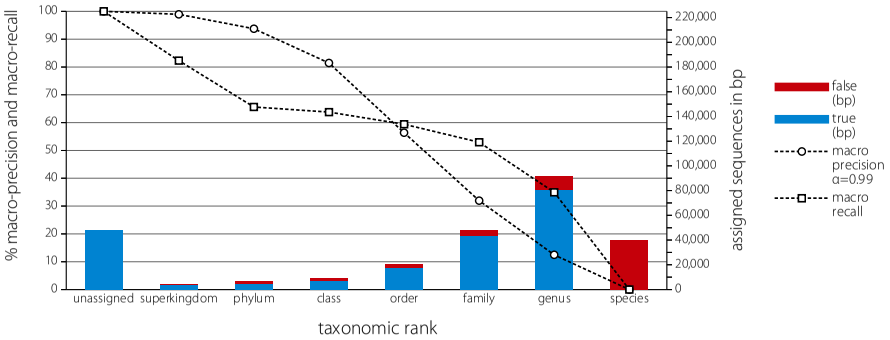
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(c) new species scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	48039	0	0	100.0	0.0	1	100.0	0.0	1	55507	174	99.7	root+superkingdom
superkingdom	1	3734	174	0	98.9	0.2	2	82.3	8.3	2				
phylum	2	5195	1718	0	93.8	5.8	17	65.6	31.8	20	29216	7007	80.7	phylum+class+order
class	3	6657	2267	0	81.4	31.7	24	63.8	28.3	23				
order	4	17364	3022	0	56.4	46.1	48	59.4	32.1	32				
family	5	43855	4055	0	32.0	44.7	96	53.0	33.4	36				
genus	6	80580	10693	0	12.5	32.0	216	35.0	35.5	41	124435	54573	69.5	family+genus+species
species	7	0	39825	0	0.0	0.0	1153	0.0	0.0	49				
avg/sum	5.1	157385	61754	0	53.6	22.9	222.3	51.3	24.2	29.0			71.8	all but unassigned
avg/sum	4.0	205424	61754	0	59.4	20.0	194.6	57.4	21.2	25.5			76.9	all with unassigned

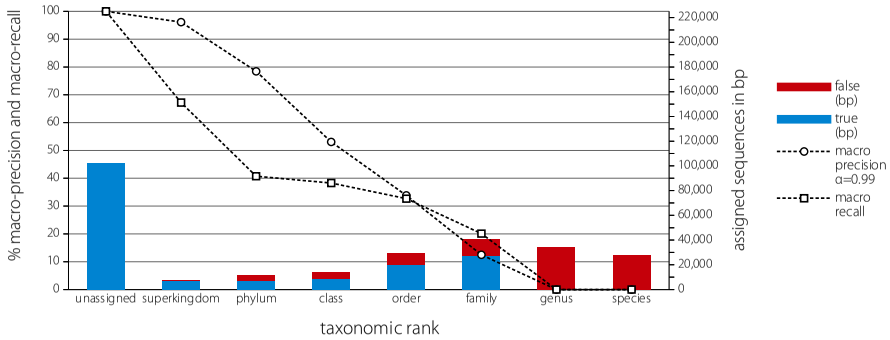
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(d) new genus scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	101939	0	0	100.0	0.0	1	100.0	0.0	1	116715	386	99.7	root+superkingdom
superkingdom	1	7388	386	0	96.1	1.9	2	67.2	14.9	2				
phylum	2	7629	4042	0	78.4	17.6	17	40.7	29.4	20	36454	18595	66.2	phylum+class+order
class	3	8751	5633	0	53.1	39.5	31	38.3	26.6	23				
order	4	20074	8920	0	33.8	39.9	65	32.7	27.5	32				
family	5	27269	13450	0	12.5	29.1	156	20.1	23.1	36				
genus	6	0	33904	0	0.0	0.0	535	0.0	0.0	41	27269	75147	26.6	family+genus+species
species	7	0	27793	0	0.0	0.0	1788	0.0	0.0	49				
avg/sum	4.3	71111	94128	0	39.1	18.3	370.6	28.4	17.4	29.0			43.0	all but unassigned
avg/sum	2.5	173050	94128	0	46.7	16.0	324.4	37.4	15.2	25.5			64.8	all with unassigned

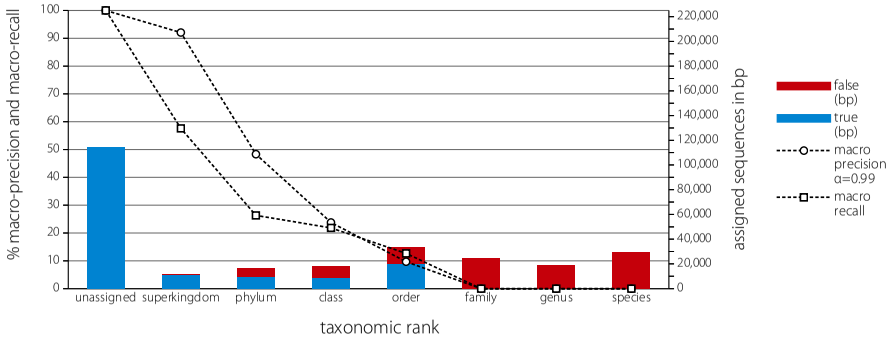
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(e) new family scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	114225	0	0	100.0	0.0	1	100.0	0.0	1	136949	536	99.6	root+superkingdom
superkingdom	1	11362	536	0	92.1	3.8	2	57.6	21.4	2				
phylum	2	9527	6860	0	48.3	32.1	18	26.3	27.3	20	39216	29057	57.4	phylum+class+order
class	3	9232	8904	0	23.8	32.6	36	21.8	25.3	23				
order	4	20457	13293	0	9.7	22.8	81	12.7	19.9	32				
family	5	0	24317	0	0.0	0.0	196	0.0	0.0	36	0	72782	0.0	family+genus+species
genus	6	0	18709	0	0.0	0.0	625	0.0	0.0	41				
species	7	0	29756	0	0.0	0.0	1816	0.0	0.0	49				
avg/sum	3.8	50578	102375	0	24.8	13.0	396.3	16.9	13.4	29.0			33.1	all but unassigned
avg/sum	2.0	164803	102375	0	34.2	11.4	346.9	27.3	11.7	25.5			61.7	all with unassigned

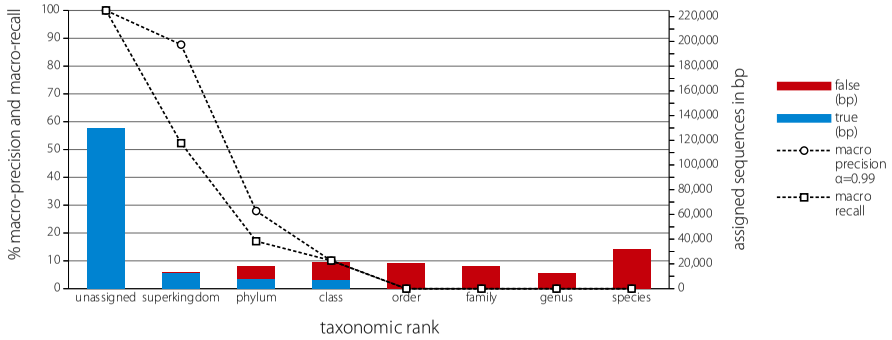
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(f) new order scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	130123	0	0	100.0	0.0	1	100.0	0.0	1	154807	706	99.5	root+superkingdom
superkingdom	1	12342	706	0	87.7	7.2	2	52.3	23.4	2				
phylum	2	8019	10411	0	27.9	30.5	21	17.0	23.2	20	15250	45492	25.1	phylum+class+order
class	3	7231	14435	0	10.0	21.5	39	10.1	16.8	23				
order	4	0	20646	0	0.0	0.0	90	0.0	0.0	32				
family	5	0	18233	0	0.0	0.0	203	0.0	0.0	36	0	63265	0.0	family+genus+species
genus	6	0	12779	0	0.0	0.0	652	0.0	0.0	41				
species	7	0	32253	0	0.0	0.0	1810	0.0	0.0	49				
avg/sum	3.5	27592	109463	0	17.9	8.5	402.4	11.4	9.0	29.0			20.1	all but unassigned
avg/sum	1.6	157715	109463	0	28.2	7.4	352.3	22.4	7.9	25.5			59.0	all with unassigned

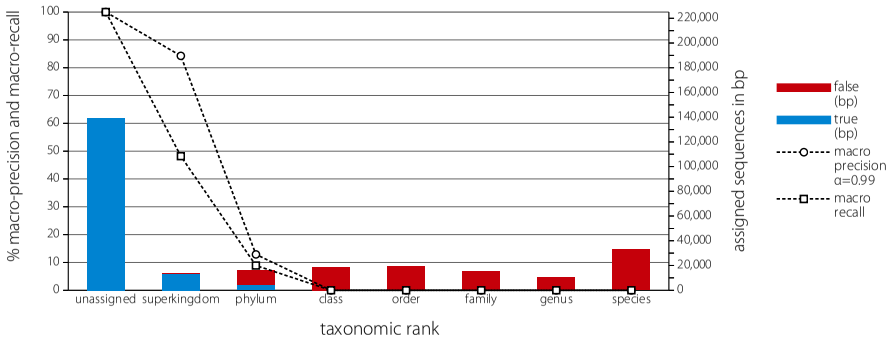
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(g) new class scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	139175	0	0	100.0	0.0	1	100.0	0.0	1	165277	778	99.5	root+superkingdom
superkingdom	1	13051	778	0	84.3	9.8	2	48.2	24.2	2				
phylum	2	4577	12085	0	12.9	20.1	22	8.9	15.4	20	4577	50298	8.3	phylum+class+order
class	3	0	18605	0	0.0	0.0	41	0.0	0.0	23				
order	4	0	19608	0	0.0	0.0	91	0.0	0.0	32				
family	5	0	15771	0	0.0	0.0	206	0.0	0.0	36	0	59299	0.0	family+genus+species
genus	6	0	10539	0	0.0	0.0	657	0.0	0.0	41				
species	7	0	32989	0	0.0	0.0	1814	0.0	0.0	49				
avg/sum	3.4	17628	110375	0	13.9	4.3	404.7	8.2	5.7	29.0			13.8	all but unassigned
avg/sum	1.4	156803	110375	0	24.6	3.7	354.3	19.6	5.0	25.5			58.7	all with unassigned

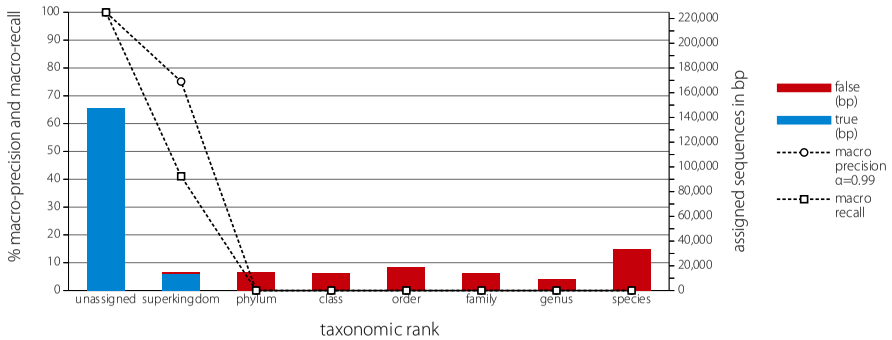
CARMA3 binning on simulated metagenome with 49 species



Supplementary Figure S11 - CARMA binning of simulated metagenome with 49 species (simArt49e) **(h) new phylum scenario**

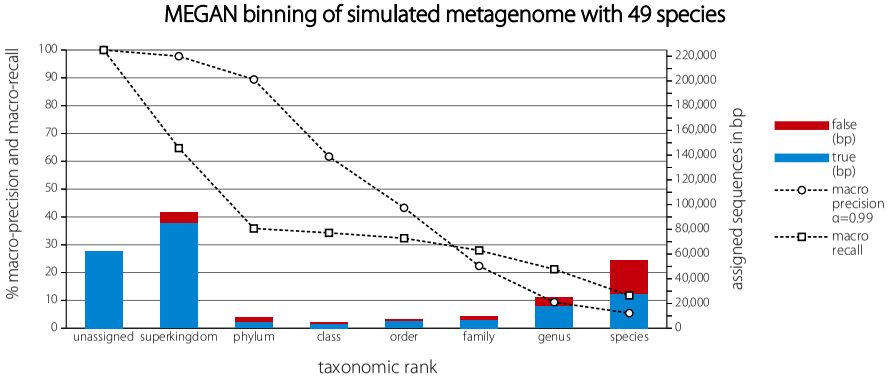
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	147652	0	0	100.0	0.0	1	100.0	0.0	1	174430	921	99.5	root+superkingdom
superkingdom	1	13389	921	0	75.1	17.3	2	41.1	27.2	2				
phylum	2	0	14479	0	0.0	0.0	24	0.0	0.0	20	0	47825	0.0	phylum+class+order
class	3	0	14228	0	0.0	0.0	42	0.0	0.0	23				
order	4	0	19118	0	0.0	0.0	93	0.0	0.0	32				
family	5	0	14181	0	0.0	0.0	214	0.0	0.0	36	0	57391	0.0	family+genus+species
genus	6	0	9565	0	0.0	0.0	664	0.0	0.0	41				
species	7	0	33645	0	0.0	0.0	1820	0.0	0.0	49				
avg/sum	3.4	13389	106137	0	10.7	2.5	408.4	5.9	3.9	29.0			11.2	all but unassigned
avg/sum	1.2	161041	106137	0	21.9	2.2	357.5	17.6	3.4	25.5			60.3	all with unassigned

CARMA3 binning on simulated metagenome with 49 species



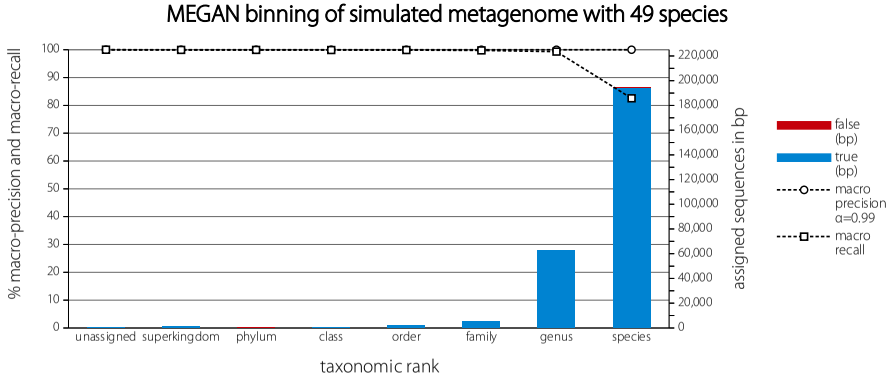
Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(a) summary scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	62255.4	0.0	0	100.0	0.0	1	100.0	0.0	1	232794.3	8388.1	96.5	root+superkingdom
superkingdom	1	85269.4	8388.1	0	97.8	0.1	2	64.8	21.3	2				
phylum	2	5415.0	3937.6	0	89.4	8.3	19	35.9	14.3	20	14756.4	7006.7	67.8	phylum+class+order
class	3	3302.9	1523.4	0	61.7	41.3	33	34.3	9.9	23				
order	4	6038.6	1545.7	0	43.3	45.1	66	32.3	9.3	32				
family	5	6638.4	2415.9	0	22.4	38.7	139	28.0	8.0	36	53023.9	36478.0	59.2	family+genus+species
genus	6	18552.9	6525.4	0	9.3	27.9	400	21.2	5.9	41				
species	7	27832.6	27536.7	0	5.4	21.9	824	11.8	4.5	49				
avg/sum	2.4	153049.7	51872.9	0	47.0	26.2	211.9	32.6	10.4	29.0			74.7	all but unassigned
avg/sum	1.7	215305.1	51872.9	0	53.7	22.9	185.5	41.0	9.1	25.5			80.6	all with unassigned



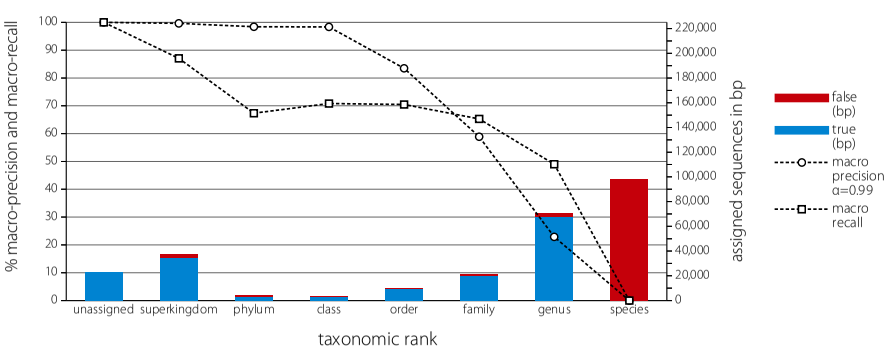
Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(b) all reference scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	595	0	0	100.0	0.0	1	100.0	0.0	1	2301	0	100.0	root+superkingdom
superkingdom	1	853	0	0	100.0	0.0	2	99.9	0.1	2				
phylum	2	515	1	0	100.0	0.0	19	99.9	0.1	20	2948	1	100.0	phylum+class+order
class	3	399	0	0	100.0	0.0	23	99.9	0.1	23				
order	4	2034	0	0	100.0	0.0	31	99.9	0.1	32				
family	5	5388	0	0	100.0	0.0	35	99.8	0.3	36	262771	10	100.0	family+genus+species
genus	6	62555	0	0	100.0	0.0	40	99.3	1.6	41				
species	7	194828	10	0	100.0	0.0	44	82.5	31.3	49				
avg/sum	5.8	266572	11	0	100.0	0.0	27.7	97.3	4.8	29.0			100.0	all but unassigned
avg/sum	5.7	267167	11	0	100.0	0.0	24.4	97.7	4.2	25.5			100.0	all with unassigned



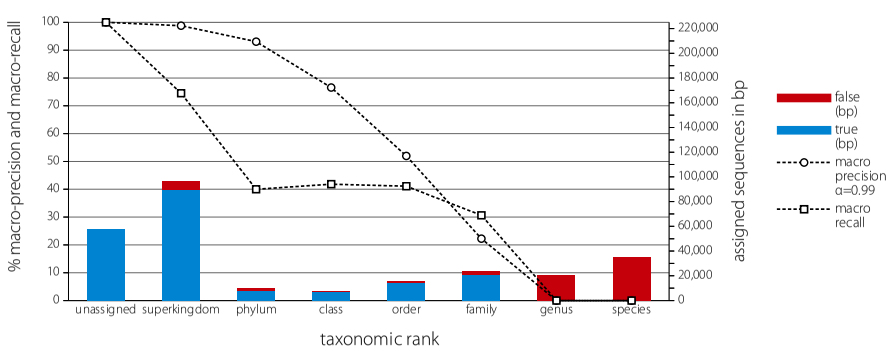
Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(c) new species scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	22828	0	0	100.0	0.0	1	100.0	0.0	1	91304	3128	96.7	root+superkingdom
superkingdom	1	34238	3128	0	99.6	0.1	2	87.1	9.1	2				
phylum	2	3671	867	0	98.4	1.9	17	67.3	33.8	20	15968	1537	91.2	phylum+class+order
class	3	3130	282	0	98.4	1.8	21	70.8	29.0	23				
order	4	9167	388	0	83.5	34.4	35	70.5	31.9	32				
family	5	20053	979	0	58.9	47.3	55	65.3	34.3	36	87368	102111	46.1	family+genus+species
genus	6	67315	3069	0	22.9	41.1	121	49.0	40.8	41				
species	7	0	98063	0	0.0	0.0	218	0.0	0.0	49				
avg/sum	4.3	137574	106776	0	66.0	18.1	67.0	58.6	25.6	29.0			56.3	all but unassigned
avg/sum	3.7	160402	106776	0	70.2	15.8	58.8	63.8	22.4	25.5			60.0	all with unassigned



Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(d) new genus scenario**

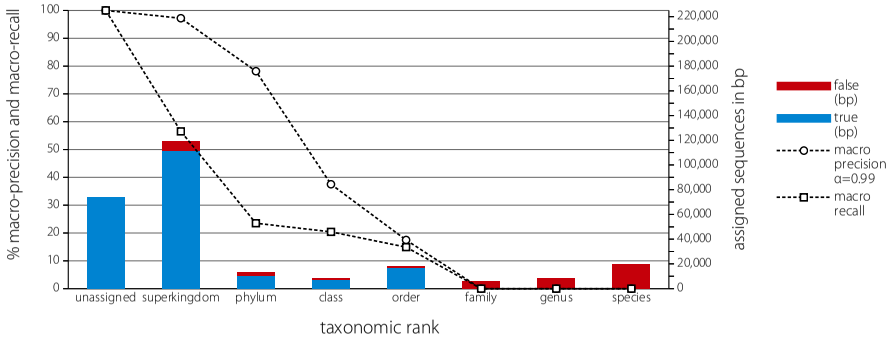
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	58109	0	0	100.0	0.0	1	100.0	0.0	1	238003	6636	97.3	root+superkingdom
superkingdom	1	89947	6636	0	98.8	0.0	2	74.5	15.4	2				
phylum	2	8288	1861	0	93.1	9.2	15	40.0	29.9	20	29489	3515	89.3	phylum+class+order
class	3	6666	657	0	76.6	34.7	25	41.8	26.2	23				
order	4	14535	997	0	52.0	45.1	50	41.1	29.8	32				
family	5	21028	3013	0	22.3	38.5	105	30.6	30.6	36	21028	58454	26.5	family+genus+species
genus	6	0	20343	0	0.0	0.0	274	0.0	0.0	41				
species	7	0	35098	0	0.0	0.0	430	0.0	0.0	49				
avg/sum	2.5	140464	68605	0	49.0	18.2	128.7	32.6	18.8	29.0			67.2	all but unassigned
avg/sum	1.9	198573	68605	0	55.3	15.9	112.8	41.0	16.5	25.5			74.3	all with unassigned



Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(e) new family scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	73809	0	0	100.0	0.0	1	100.0	0.0	1	296855	8005	97.4	root+superkingdom
superkingdom	1	111523	8005	0	97.2	0.4	2	56.5	28.9	2				
phylum	2	10028	2966	0	78.2	27.3	14	23.5	25.8	20				
class	3	7514	1043	0	37.5	42.3	31	20.4	21.9	23	34076	5490	86.1	phylum+class+order
order	4	16534	1481	0	17.5	32.9	69	14.9	21.1	32				
family	5	0	5906	0	0.0	0.0	161	0.0	0.0	36				
genus	6	0	8487	0	0.0	0.0	366	0.0	0.0	41	0	34275	0.0	family+genus+species
species	7	0	19882	0	0.0	0.0	580	0.0	0.0	49				
avg/sum	1.9	145599	47770	0	32.9	14.7	174.7	16.5	14.0	29.0			75.3	all but unassigned
avg/sum	1.3	219408	47770	0	41.3	12.9	153.0	26.9	12.2	25.5			82.1	all with unassigned

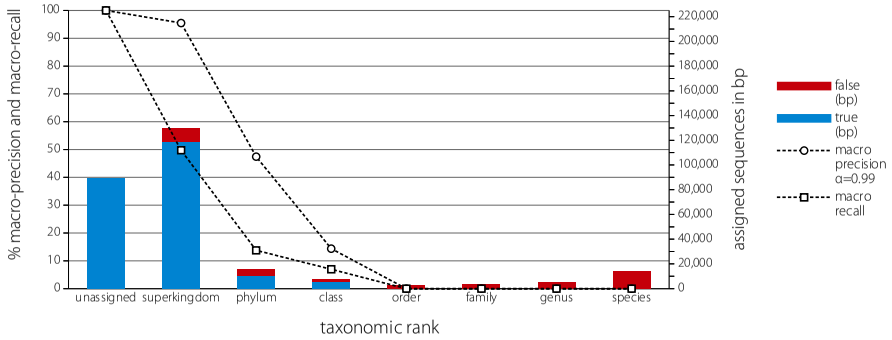
MEGAN binning of simulated metagenome with 49 species



Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(f) new order scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	89682	0	0	100.0	0.0	1	100.0	0.0	1	326876	11005	96.7	root+superkingdom
superkingdom	1	118597	11005	0	95.5	1.3	2	49.8	30.9	2				
phylum	2	10011	5581	0	47.4	37.8	17	13.8	21.7	20				
class	3	5411	1881	0	14.4	28.3	39	7.0	11.7	23	15422	9908	60.9	phylum+class+order
order	4	0	2446	0	0.0	0.0	84	0.0	0.0	32				
family	5	0	3416	0	0.0	0.0	167	0.0	0.0	36				
genus	6	0	5382	0	0.0	0.0	401	0.0	0.0	41	0	22564	0.0	family+genus+species
species	7	0	13766	0	0.0	0.0	565	0.0	0.0	49				
avg/sum	1.5	134019	43477	0	22.5	9.6	182.1	10.1	9.2	29.0			75.5	all but unassigned
avg/sum	1.0	223701	43477	0	32.2	8.4	159.5	21.3	8.0	25.5			83.7	all with unassigned

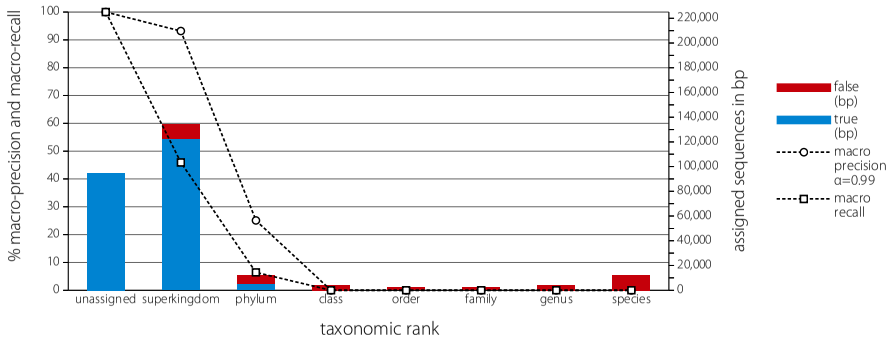
MEGAN binning of simulated metagenome with 49 species



Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(g) new class scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	94817	0	0	100.0	0.0	1	100.0	0.0	1	340147	11467	96.7	root+superkingdom
superkingdom	1	122665	11467	0	93.2	2.8	2	45.9	31.8	2				
phylum	2	5392	7208	0	25.1	30.7	19	6.4	11.6	20				
class	3	0	4356	0	0.0	0.0	43	0.0	0.0	23	5392	13709	28.2	phylum+class+order
order	4	0	2145	0	0.0	0.0	88	0.0	0.0	32				
family	5	0	2203	0	0.0	0.0	172	0.0	0.0	36				
genus	6	0	4437	0	0.0	0.0	446	0.0	0.0	41	0	19128	0.0	family+genus+species
species	7	0	12488	0	0.0	0.0	657	0.0	0.0	49				
avg/sum	1.4	128057	44304	0	16.9	4.8	203.9	7.5	6.2	29.0			74.3	all but unassigned
avg/sum	0.9	222874	44304	0	27.3	4.2	178.5	19.0	5.4	25.5			83.4	all with unassigned

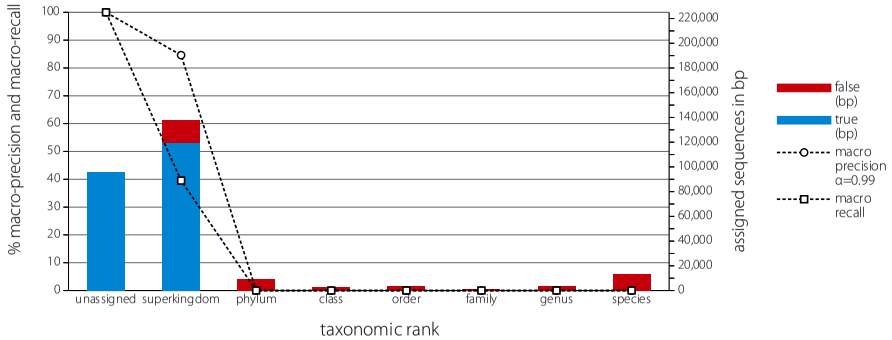
MEGAN binning of simulated metagenome with 49 species



Supplementary Figure S12 - MEGAN binning of simulated metagenome with 49 species (simArt49e) **(h) new phylum scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	95948	0	0	100.0	0.0	1	100.0	0.0	1	334074	18476	94.8	root+superkingdom
superkingdom	1	119063	18476	0	84.6	8.8	2	39.5	33.1	2				
phylum	2	0	9079	0	0.0	0.0	25	0.0	0.0	20				
class	3	0	2445	0	0.0	0.0	45	0.0	0.0	23	0	14887	0.0	phylum+class+order
order	4	0	3363	0	0.0	0.0	96	0.0	0.0	32				
family	5	0	1394	0	0.0	0.0	197	0.0	0.0	36				
genus	6	0	3960	0	0.0	0.0	494	0.0	0.0	41	0	18804	0.0	family+genus+species
species	7	0	13450	0	0.0	0.0	814	0.0	0.0	49				
avg/sum	1.3	119063	52167	0	12.1	1.3	239.0	5.6	4.7	29.0			69.5	all but unassigned
avg/sum	0.8	215011	52167	0	23.1	1.1	209.3	17.4	4.1	25.5			80.5	all with unassigned

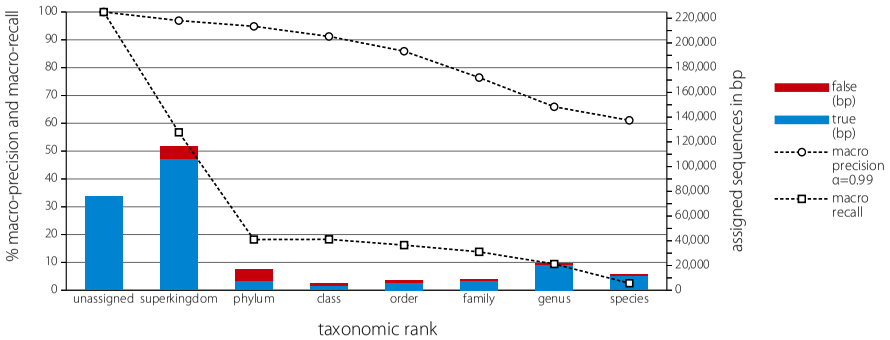
MEGAN binning of simulated metagenome with 49 species



Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(a) summary scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	75644.4	0.0	0	100.0	0.0	1	100.0	0.0	1	288633.9	10293.3	96.6	root+superkingdom
superkingdom	1	106494.7	10293.3	0	96.9	2.5	2	56.8	33.5	2				
phylum	2	7691.6	9493.1	0	94.9	9.2	16	18.2	13.5	20	17180.7	13851.4	55.4	phylum+class+order
class	3	3656.6	2344.0	0	91.2	21.5	21	18.3	11.6	23				
order	4	5832.6	2014.3	0	85.9	31.8	34	16.2	9.5	32				
family	5	7550.6	1079.7	0	76.4	39.8	44	13.8	8.2	36	39774.7	3938.7	91.0	family+genus+species
genus	6	20271.9	1397.1	0	65.9	46.4	58	9.4	7.7	41				
species	7	11952.3	1461.9	0	61.1	47.2	65	2.5	4.4	49				
avg/sum	2.1	163450.1	28083.4	0	81.8	28.3	34.3	19.3	12.6	29.0			85.3	all but unassigned
avg/sum	1.5	239094.6	28083.4	0	84.0	24.8	30.1	29.4	11.0	25.5			89.5	all with unassigned

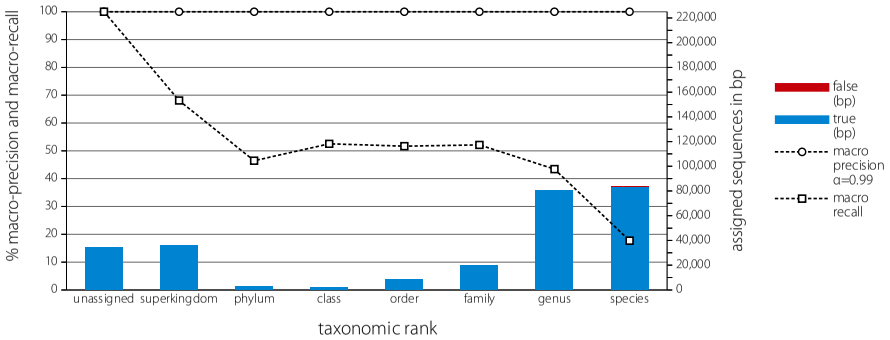
taxator-tk binning of simulated metagenome with 49 species



Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(b) all reference scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	34453	0	0	100.0	0.0	1	100.0	0.0	1	105775	0	100.0	root+superkingdom
superkingdom	1	35661	0	0	100.0	0.0	2	68.1	8.4	2				
phylum	2	2897	0	0	100.0	0.0	17	46.5	28.4	20	13098	0	100.0	phylum+class+order
class	3	1947	0	0	100.0	0.0	21	52.5	29.1	23				
order	4	8254	0	0	100.0	0.0	29	51.7	30.6	32				
family	5	19632	0	0	100.0	0.0	32	52.1	31.5	36	183965	1	100.0	family+genus+species
genus	6	80667	0	0	100.0	0.0	34	43.4	34.0	41				
species	7	83666	1	0	100.0	0.0	34	17.7	30.5	49				
avg/sum	4.4	232724	1	0	100.0	0.0	24.1	47.4	27.5	29.0			100.0	all but unassigned
avg/sum	3.6	267177	1	0	100.0	0.0	21.3	54.0	24.1	25.5			100.0	all with unassigned

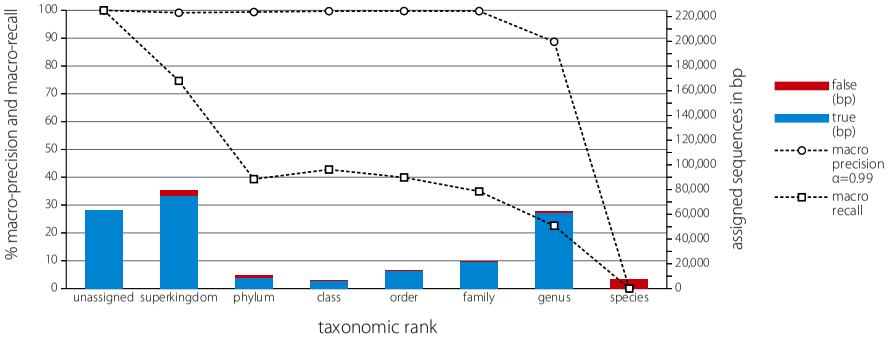
taxator-tk binning of simulated metagenome with 49 species



Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(c) new species scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	63523	0	0	100.0	0.0	1	100.0	0.0	1	214561	4247	98.1	root+superkingdom
superkingdom	1	75519	4247	0	99.1	0.9	2	74.7	16.4	2				
phylum	2	8526	1834	0	99.4	1.4	15	39.3	28.6	20	29377	2671	91.7	phylum+class+order
class	3	6516	515	0	99.7	0.5	18	42.8	27.4	23				
order	4	14335	322	0	99.7	0.5	25	39.9	26.9	32				
family	5	21470	246	0	99.7	0.8	28	34.9	28.6	36	82706	9135	90.1	family+genus+species
genus	6	61236	1365	0	88.7	30.3	26	22.6	28.3	41				
species	7	0	7524	0	0.0	0.0	48	0.0	0.0	49				
avg/sum	3.4	187602	16053	0	83.8	4.9	23.1	36.3	22.3	29.0			92.1	all but unassigned
avg/sum	2.6	251125	16053	0	85.8	4.3	20.4	44.3	19.5	25.5			94.0	all with unassigned

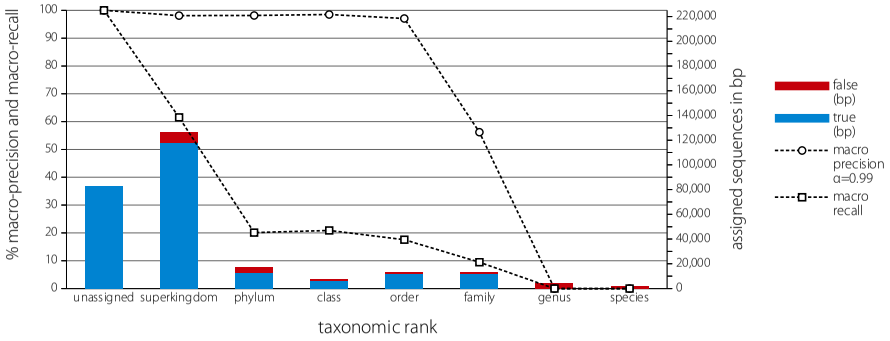
taxator-tk binning of simulated metagenome with 49 species



Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(d) new genus scenario**

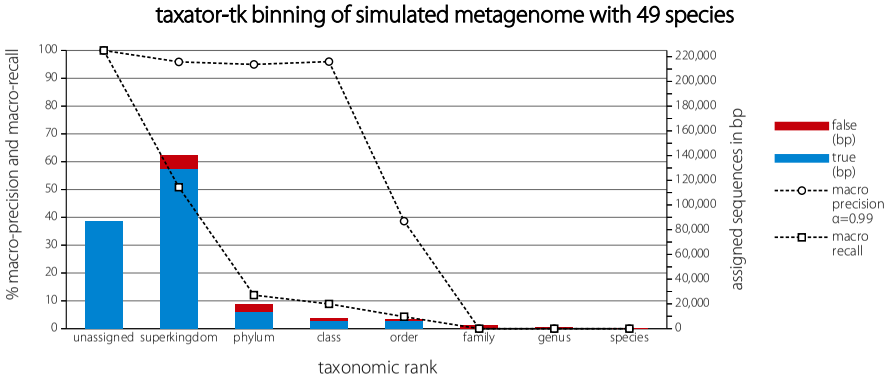
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	82737	0	0	100.0	0.0	1	100.0	0.0	1	318017	8730	97.3	root+superkingdom
superkingdom	1	117640	8730	0	98.1	1.4	2	61.5	31.7	2				
phylum	2	12404	4567	0	98.2	3.3	13	20.1	21.3	20	31213	7247	81.2	phylum+class+order
class	3	6555	1439	0	98.5	1.9	16	20.9	21.8	23				
order	4	12254	1241	0	97.1	4.9	22	17.6	20.2	32				
family	5	11752	1508	0	56.2	46.1	33	9.5	16.7	36	11752	7859	59.9	family+genus+species
genus	6	0	4633	0	0.0	0.0	52	0.0	0.0	41				
species	7	0	1718	0	0.0	0.0	49	0.0	0.0	49				
avg/sum	1.8	160605	23836	0	64.0	8.2	26.7	18.5	16.0	29.0			87.1	all but unassigned
avg/sum	1.3	243342	23836	0	68.5	7.2	23.5	28.7	14.0	25.5			91.1	all with unassigned

taxator-tk binning of simulated metagenome with 49 species



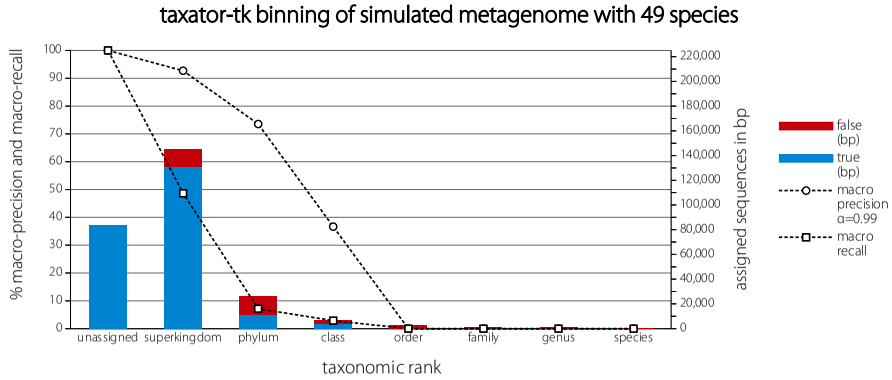
Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(e) new family scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	86887	0	0	100.0	0.0	1	100.0	0.0	1	345265	10849	97.0	root+superkingdom
superkingdom	1	129189	10849	0	95.9	2.0	2	50.8	42.7	2				
phylum	2	13258	6399	0	95.0	6.2	8	12.1	19.5	20	25853	9802	72.5	phylum+class+order
class	3	6610	1866	0	96.0	4.4	12	8.9	14.0	23				
order	4	5985	1537	0	38.6	46.6	27	4.3	8.3	32				
family	5	0	2768	0	0.0	0.0	48	0.0	0.0	36	0	4598	0.0	family+genus+species
genus	6	0	1474	0	0.0	0.0	85	0.0	0.0	41				
species	7	0	356	0	0.0	0.0	81	0.0	0.0	49				
avg/sum	1.4	155042	25249	0	46.5	8.5	37.6	10.9	12.1	29.0			86.0	all but unassigned
avg/sum	1.0	241929	25249	0	53.2	7.4	33.0	22.0	10.6	25.5			90.5	all with unassigned



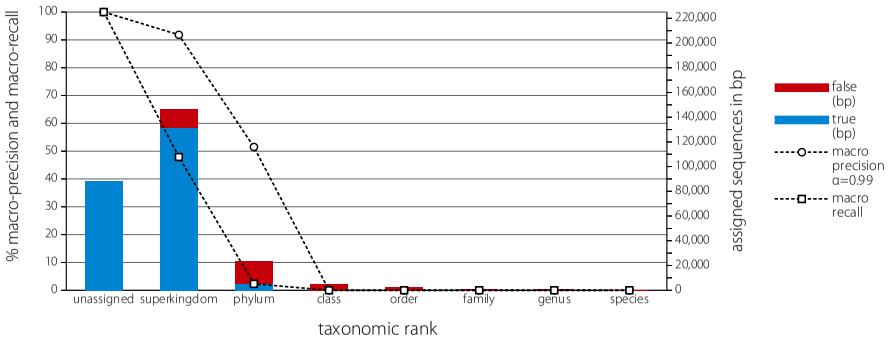
Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(f) new order scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	83674	0	0	100.0	0.0	1	100.0	0.0	1	345694	14172	96.1	root+superkingdom
superkingdom	1	131010	14172	0	92.7	0.0	1	48.7	45.1	2				
phylum	2	11454	14889	0	73.6	25.5	6	7.2	16.0	20				
class	3	3968	3071	0	36.7	40.1	16	2.9	7.2	23	15422	20404	43.0	phylum+class+order
order	4	0	2444	0	0.0	0.0	38	0.0	0.0	32				
family	5	0	1364	0	0.0	0.0	70	0.0	0.0	36				
genus	6	0	901	0	0.0	0.0	103	0.0	0.0	41	0	2496	0.0	family+genus+species
species	7	0	231	0	0.0	0.0	83	0.0	0.0	49				
avg/sum	1.3	146432	37072	0	29.0	9.4	45.3	8.4	9.7	29.0			79.8	all but unassigned
avg/sum	0.9	230106	37072	0	37.9	8.2	39.8	19.8	8.5	25.5			86.1	all with unassigned



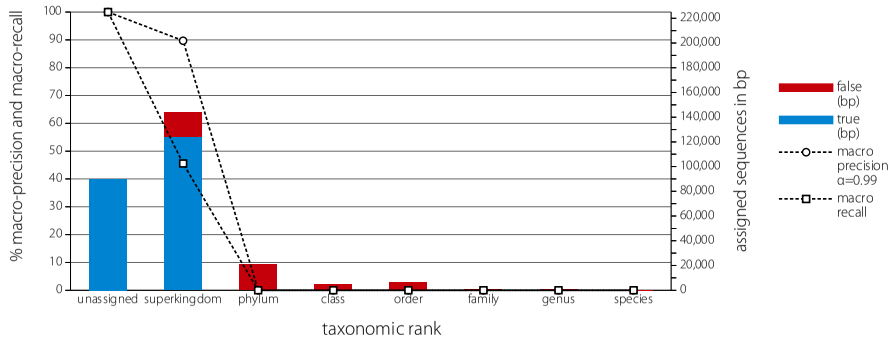
Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(g) new class scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	88280	0	0	100.0	0.0	1	100.0	0.0	1	352286	14601	96.0	root+superkingdom
superkingdom	1	132003	14601	0	91.8	0.0	1	47.9	44.8	2				
phylum	2	5302	17992	0	51.5	32.3	5	2.3	5.5	20	5302	25163	17.4	phylum+class+order
class	3	0	4839	0	0.0	0.0	17	0.0	0.0	23				
order	4	0	2332	0	0.0	0.0	41	0.0	0.0	32				
family	5	0	961	0	0.0	0.0	73	0.0	0.0	36				
genus	6	0	672	0	0.0	0.0	107	0.0	0.0	41	0	1829	0.0	family+genus+species
species	7	0	196	0	0.0	0.0	74	0.0	0.0	49				
avg/sum	1.3	137305	41593	0	20.5	4.6	45.4	7.2	7.2	29.0			76.8	all but unassigned
avg/sum	0.8	225585	41593	0	30.4	4.0	39.9	18.8	6.3	25.5			84.4	all with unassigned

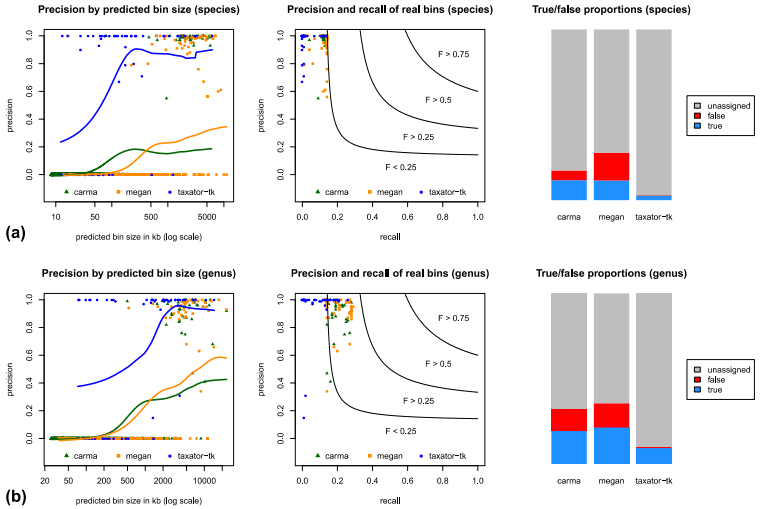


Supplementary Figure S13 - Taxator-tk binning of simulated metagenome with 49 species (simArt49e) **(h) new phylum scenario**

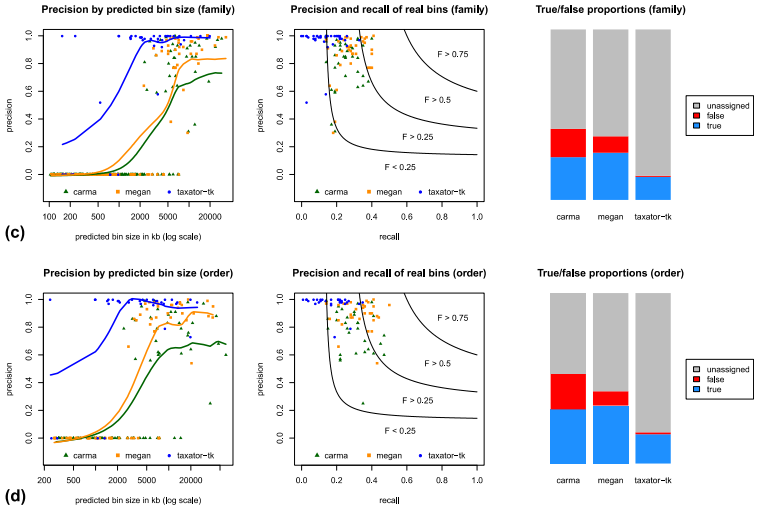
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	89957	0	0	100.0	0.0	1	100.0	0.0	1	338839	19454	94.6	root+superkingdom
superkingdom	1	124441	19454	0	89.7	0.0	1	45.6	45.0	2				
phylum	2	0	20771	0	0.0	0.0	6	0.0	0.0	20	0	31673	0.0	phylum+class+order
class	3	0	4678	0	0.0	0.0	15	0.0	0.0	23				
order	4	0	6224	0	0.0	0.0	27	0.0	0.0	32				
family	5	0	711	0	0.0	0.0	94	0.0	0.0	36				
genus	6	0	735	0	0.0	0.0	141	0.0	0.0	41	0	1653	0.0	family+genus+species
species	7	0	207	0	0.0	0.0	100	0.0	0.0	49				
avg/sum	1.3	124441	52780	0	12.8	0.0	54.9	6.5	6.4	29.0			70.2	all but unassigned
avg/sum	0.9	214398	52780	0	23.7	0.0	48.1	18.2	5.6	25.5			80.2	all with unassigned



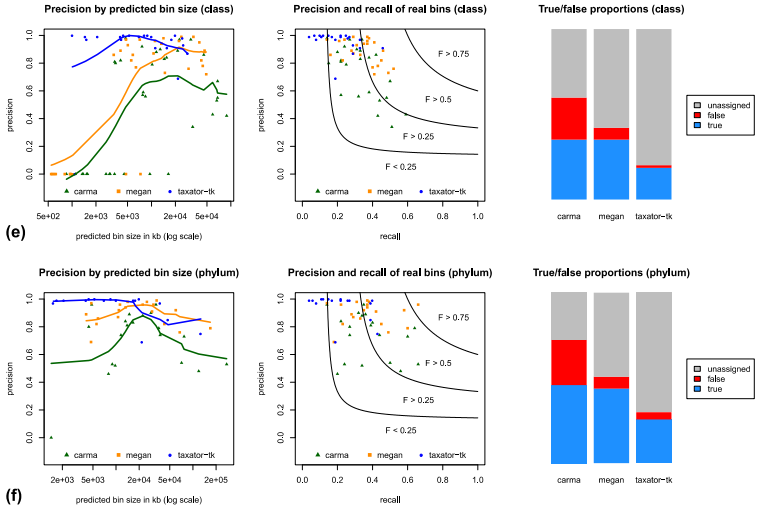
Supplementary Figure S14: Bin precision plots for 49 species simulated metagenomic sample (simArt49e)



Supplementary Figure S14: Bin precision plots for 49 species simulated metagenomic sample (simArt49e)



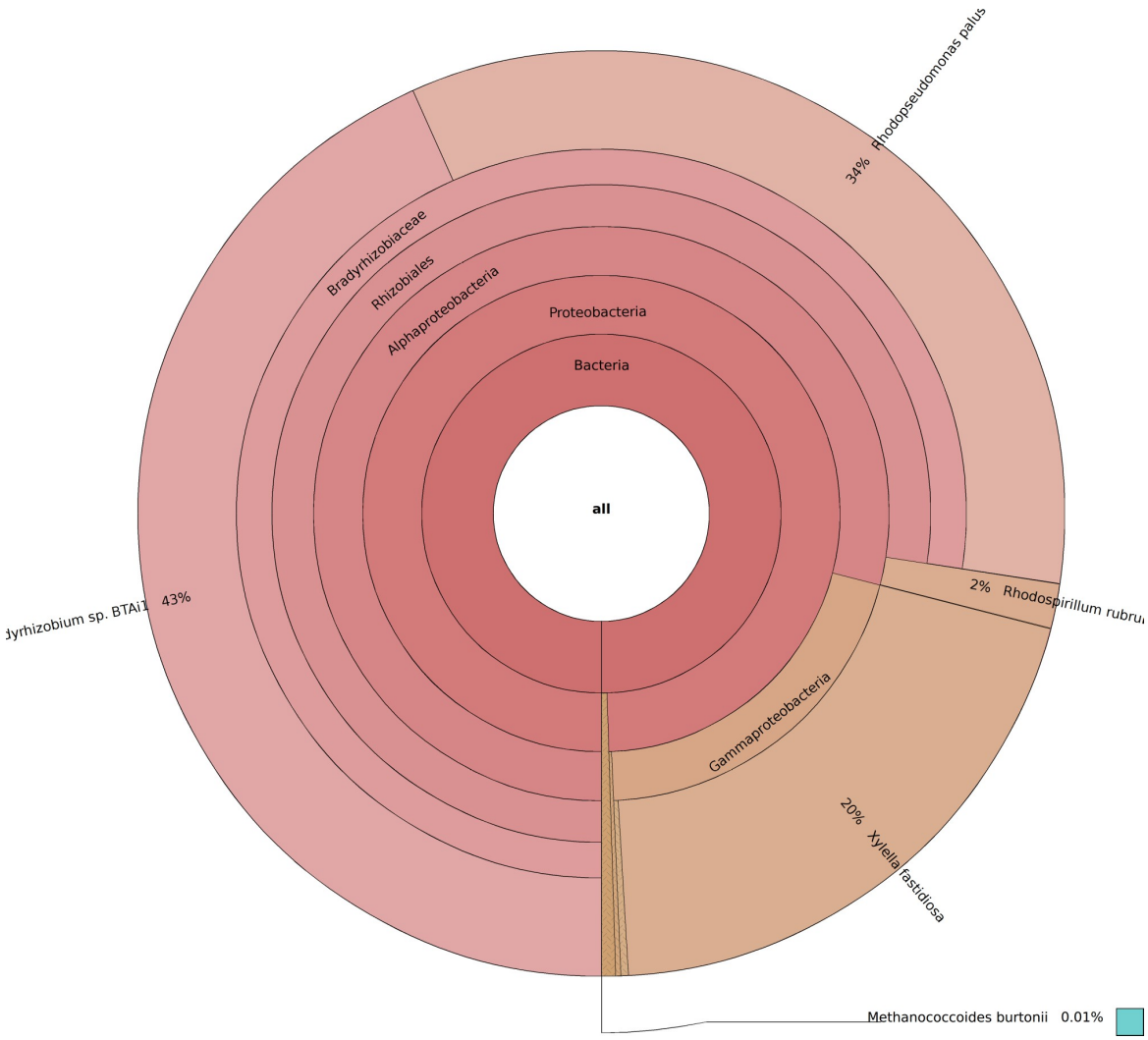
Supplementary Figure S14: Bin precision plots for 49 species simulated metagenomic sample (simArt49e)



Supplementary Figure S14: Bin precision plots for 49 species simulated metagenomic sample (simArt49e)

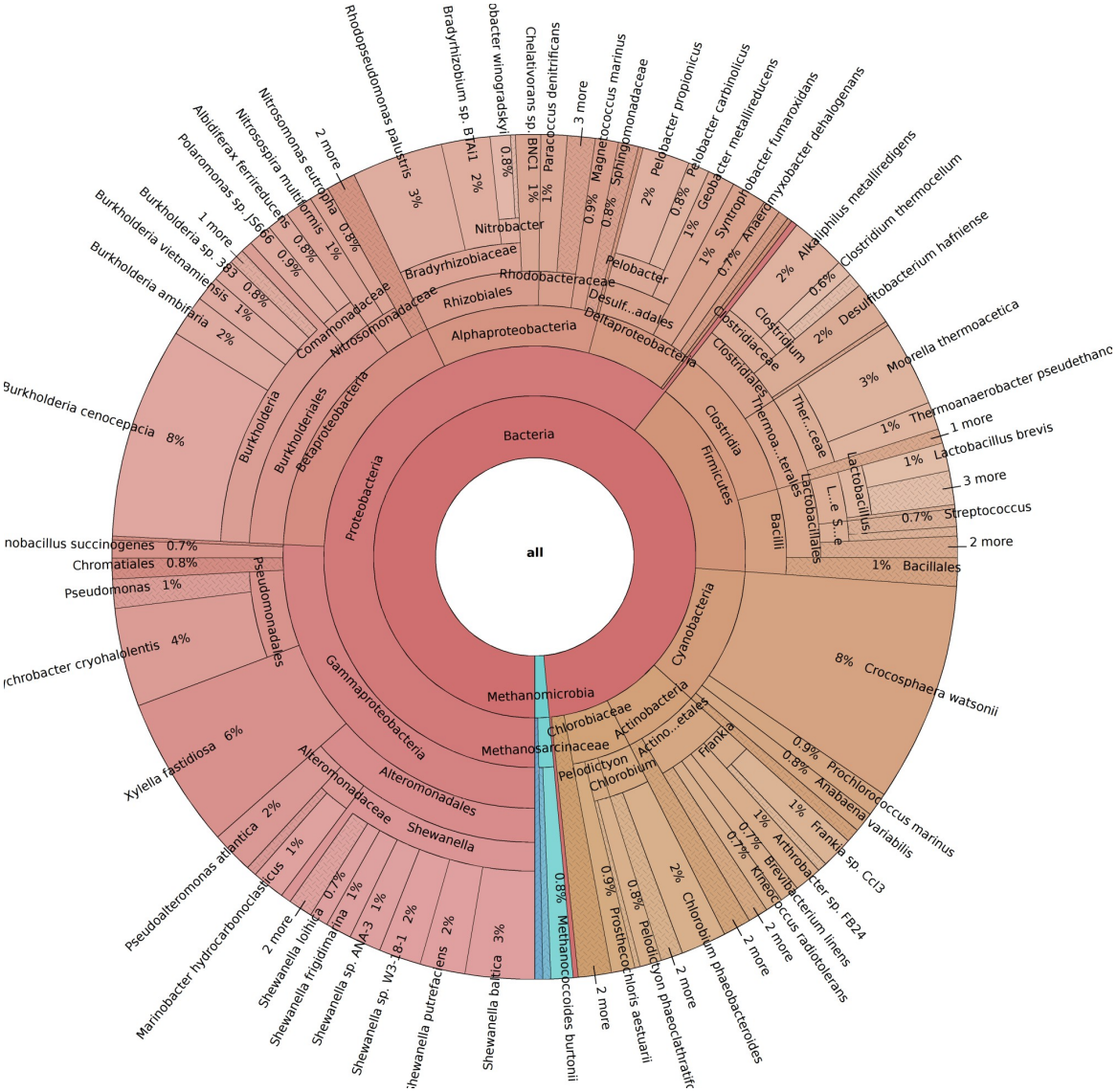
Comparison of assignment quality of *CARMA3*, *MEGAN4* and *taxator-tk* for a simulated metagenome sample from a 49 species microbial community. Values are shown for the summary scenario (sum of all seven cross-validation scenarios), for assignments to the (a) species, (b) genus, (c) family, (d) order, (e) class and (f) phylum ranks, respectively. The first of each panels shows the precision and size for every predicted bin (after removing low abundance bins). The colored line shows a smoothed k-nearest-neighbor estimate of the mean precision as a function of predicted bin size using the R function *wapply* (width=0.3) followed by smooth.spline (df=10). The second panel for each rank shows bin precisions relative to recall. The F-score partitioning helps to identify similar quality bins if precision and recall are equally weighted, however we consider precision more important than recall. The third panel illustrates the total number of true (blue) and false (red) and unassigned (gray) portion of assignments at the respective ranks. Note that partially incorrect assignments are considered incorrect for the low ranking false part of the assignment and correct for the higher ranks.

Supplementary Figure S15: Taxonomic composition of SimMC/AMD



Taxonomic composition of the FAMeS simulated metagenome sample SimMC/AMD using Krona (Ondov et *al*., 2011). An interactive version can be found in the supplementary files (SimMC.krona.html). Abundance is measured in terms of accumulated contigs lengths.

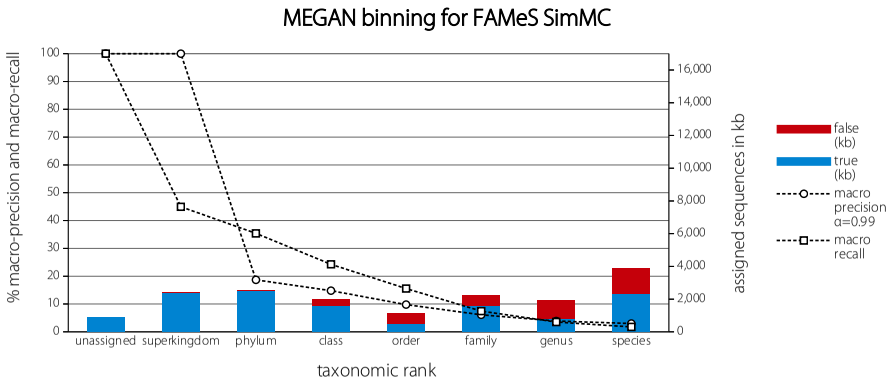
Supplementary Figure S16: Taxonomic composition of SimHC/soil



Taxonomic composition of the FAMeS simulated metagenome sample SimHC/soil using Krona (Ondov et *al.*, 2011). An interactive version can be found in the supplementary files (SimHC.krona.html). Abundance is measured in terms of accumulated contigs lengths.

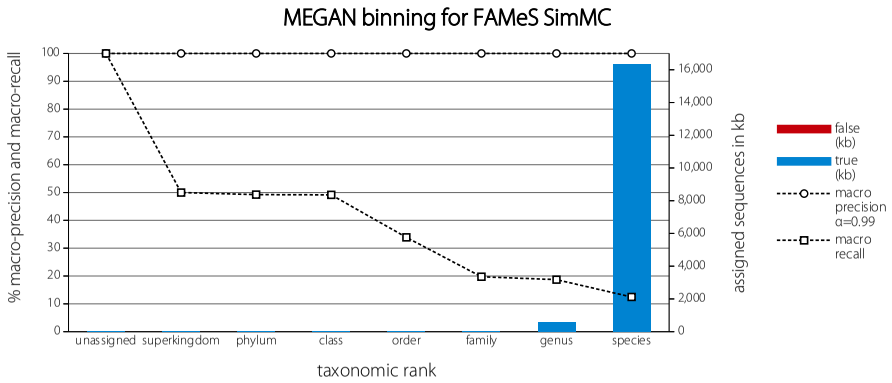
Supplementary Figure S17 - MEGAN binning for FAMeS SimMC **(a) summary scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	877.9	0.0	0	100.0	0.0	1	100.0	0.0	1	5735.3	7.5	99.9	root+superkingdom
superkingdom	1	2428.7	7.5	0	100.0	0.0	1	45.0	45.0	2				
phylum	2	2508.3	60.0	0	18.7	32.6	8	35.4	23.9	8	4604.0	1095.1	80.8	phylum+class+order
class	3	1611.6	389.1	0	14.8	29.4	17	24.3	19.7	12				
order	4	484.1	646.1	0	9.8	23.8	39	15.6	16.5	23				
family	5	1590.7	617.3	0	6.1	21.5	69	7.5	13.0	30	4734.4	3292.7	59.0	family+genus+species
genus	6	811.4	1102.6	0	3.9	18.0	131	3.5	7.2	37				
species	7	2332.3	1572.8	0	3.0	16.5	188	1.8	4.7	47				
avg/sum	3.3	11767.1	4395.3	0	22.3	20.2	64.7	19.0	18.6	22.7			72.8	all but unassigned
avg/sum	3.1	12645.0	4395.3	0	32.0	17.7	56.8	29.1	16.2	20.0			74.2	all with unassigned



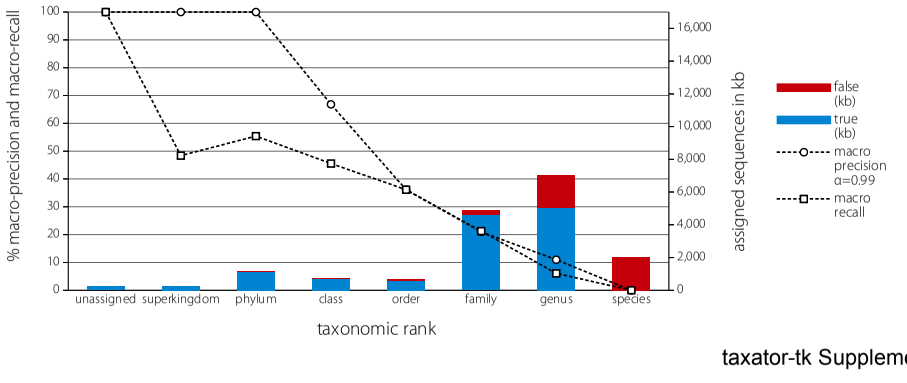
Supplementary Figure S17 - MEGAN binning for FAMeS SimMC **(b) all reference scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	2.03	0	0	100.0	0.0	1	100.0	0.0	1	20.91	0	100.0	root+superkingdom
superkingdom	1	9.44	0	0	100.0	0.0	1	50.0	50.0	2				
phylum	2	21.2	0	0	100.0	0.0	1	49.3	49.3	8				
class	3	34.76	0	0	100.0	0.0	2	49.2	49.2	12	72.15	0	100.0	phylum+class+order
order	4	16.19	0	0	100.0	0.0	3	33.9	46.5	23				
family	5	28.28	0	0	100.0	0.0	3	19.7	39.5	30				
genus	6	602.48	0	0	100.0	0.0	4	18.7	38.7	37	16956.66	0	100.0	family+genus+species
species	7	16325.9	0	0	100.0	0.0	4	12.5	32.6	47				
avg/sum	5.6	17038.25	0	0	100.0	0.0	2.6	33.3	43.7	22.7			100.0	all but unassigned
avg/sum	5.6	17040.28	0	0	100.0	0.0	2.4	41.7	38.2	20.0			100.0	all with unassigned



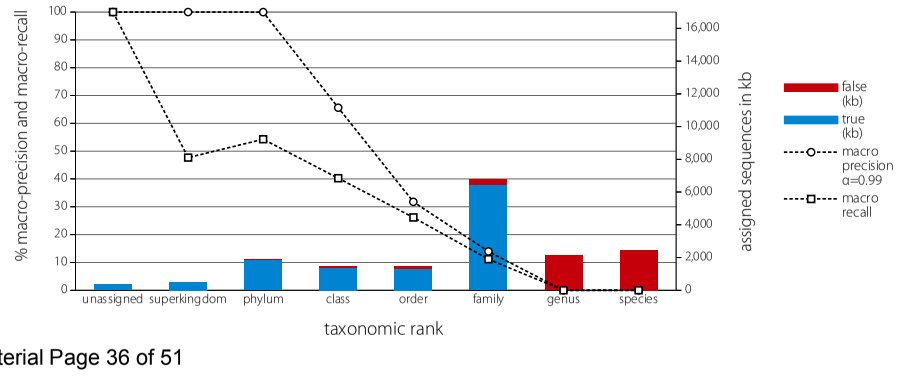
Supplementary Figure S17 - MEGAN binning for FAMeS SimMC **(c) new species scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	234.56	0	0	100.0	0.0	1	100.0	0.0	1	781.74	0	100.0	root+superkingdom
superkingdom	1	273.59	0	0	100.0	0.0	1	48.4	48.4	2				
phylum	2	1162.26	2.62	0	100.0	0.0	1	55.4	44.1	8	2423.08	125.55	95.1	phylum+class+order
class	3	683.93	59.82	0	66.8	44.7	3	45.6	37.9	12				
order	4	576.89	63.11	0	36.1	40.3	10	36.2	39.5	23				
family	5	4640.58	256.49	0	21.0	38.4	18	21.3	36.6	30	9717.71	4265.77	69.5	family+genus+species
genus	6	5077.13	1966.57	0	11.0	28.8	24	6.1	18.7	37				
species	7	0	2042.71	0	0.0	0.0	32	0.0	0.0	47				
avg/sum	5.0	12414.38	4391.32	0	47.8	21.7	12.7	30.4	32.2	22.7			73.9	all but unassigned
avg/sum	4.9	12648.94	4391.32	0	54.4	19.0	11.3	39.1	28.2	20.0			74.2	all with unassigned

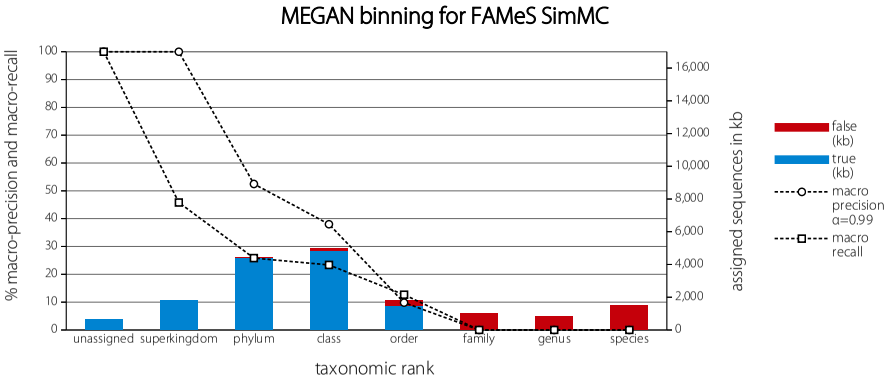


Supplementary Figure S17 - MEGAN binning for FAMeS SimMC **(d) new genus scenario**

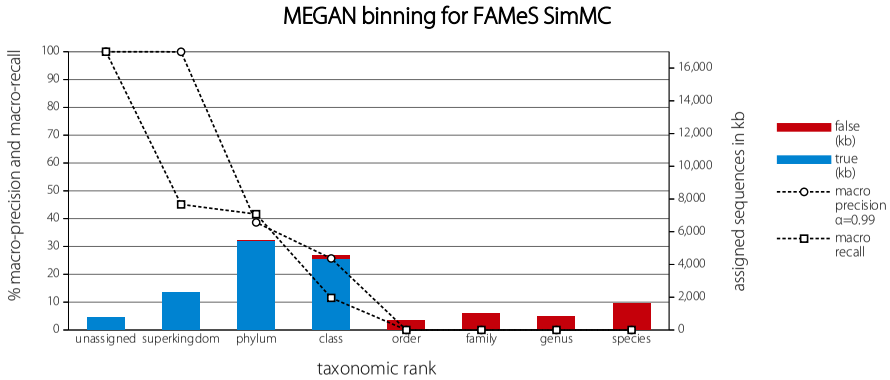
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	358.62	0	0	100.0	0.0	1	100.0	0.0	1	1411.58	0	100.0	root+superkingdom
superkingdom	1	526.48	0	0	100.0	0.0	1	47.7	47.7	2				
phylum	2	1889.35	2.62	0	100.0	0.0	1	54.3	43.3	8	4564.44	220.44	95.4	phylum+class+order
class	3	1360.44	89.37	0	65.6	45.6	3	40.3	35.8	12				
order	4	1314.65	128.45	0	31.8	40.5	11	26.2	31.9	23				
family	5	6466.24	303.88	0	13.9	32.2	17	11.2	28.1	30	6466.24	4904.05	56.9	family+genus+species
genus	6	0	2126.87	0	0.0	0.0	39	0.0	0.0	37				
species	7	0	2473.3	0	0.0	0.0	45	0.0	0.0	47				
avg/sum	4.3	11557.16	5124.49	0	44.5	16.9	16.7	25.7	26.7	22.7			69.3	all but unassigned
avg/sum	4.2	11915.78	5124.49	0	51.4	14.8	14.8	34.9	23.3	20.0			69.9	all with unassigned



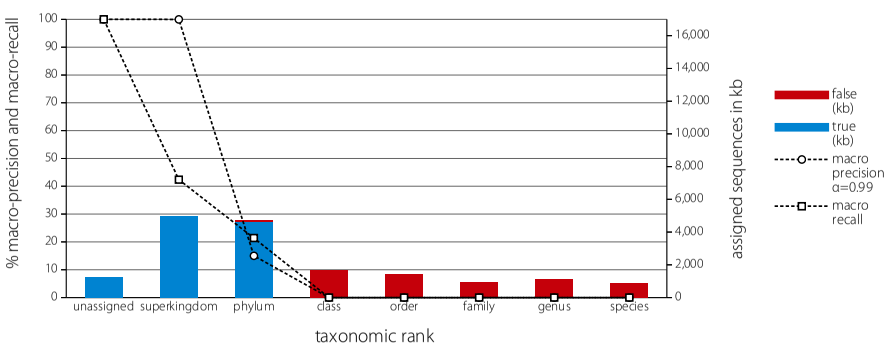
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	663.34	0	0	100.0	0.0	1	100.0	0.0	1	4214.02	0	100.0	root+superkingdom
superkingdom	1	1775.34	0	0	100.0	0.0	1	45.8	45.8	2				
phylum	2	4398.98	13.81	0	52.5	47.5	2	25.8	33.8	8	10748.33	470.17	95.8	phylum+class+order
class	3	4868.55	130.75	0	38.0	42.5	5	23.4	24.1	12				
order	4	1480.8	325.61	0	9.8	23.2	18	12.7	21.0	23				
family	5	0	1031.59	0	0.0	0.0	28	0.0	0.0	30				
genus	6	0	839.84	0	0.0	0.0	47	0.0	0.0	37	0	3383.11	0.0	family+genus+species
species	7	0	1511.68	0	0.0	0.0	47	0.0	0.0	47				
avg/sum	2.9	12523.67	3853.28	0	28.6	16.2	21.1	15.4	17.8	22.7			76.5	all but unassigned
avg/sum	2.8	13187.01	3853.28	0	37.5	14.1	18.6	26.0	15.6	20.0			77.4	all with unassigned



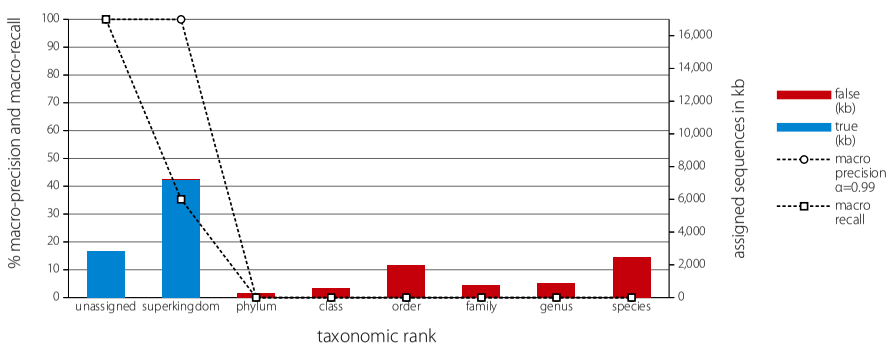
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	767.74	0	0	100.0	0.0	1	100.0	0.0	1	5311.66	0	100.0	root+superkingdom
superkingdom	1	2271.96	0	0	100.0	0.0	1	45.1	45.1	2				
phylum	2	5432.33	39.29	0	38.6	43.7	3	41.6	40.0	8	9765.97	793.72	92.5	phylum+class+order
class	3	4333.64	193	0	25.7	36.9	7	11.5	13.9	12				
order	4	0	561.43	0	0.0	0.0	21	0.0	0.0	23				
family	5	0	1006.4	0	0.0	0.0	30	0.0	0.0	30				
genus	6	0	817.67	0	0.0	0.0	43	0.0	0.0	37	0	3440.89	0.0	family+genus+species
species	7	0	1616.82	0	0.0	0.0	39	0.0	0.0	47				
avg/sum	2.7	12037.93	4234.61	0	23.5	11.5	20.6	14.0	14.1	22.7			74.0	all but unassigned
avg/sum	2.5	12805.67	4234.61	0	33.0	10.1	18.1	24.8	12.4	20.0			75.1	all with unassigned



rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1274.66	0	0	100.0	0.0	1	100.0	0.0	1	11193.42	0	100.0	root+superkingdom
superkingdom	1	4959.38	0	0	100.0	0.0	1	42.4	42.4	2				
phylum	2	4654.01	87.88	0	15.0	34.7	7	21.4	34.5	8	4654.01	3194.98	59.3	phylum+class+order
class	3	0	1661.84	0	0.0	0.0	12	0.0	0.0	12				
order	4	0	1445.26	0	0.0	0.0	31	0.0	0.0	23				
family	5	0	960.19	0	0.0	0.0	40	0.0	0.0	30				
genus	6	0	1106.84	0	0.0	0.0	57	0.0	0.0	37	0	2957.25	0.0	family+genus+species
species	7	0	890.22	0	0.0	0.0	46	0.0	0.0	47				
avg/sum	2.5	9613.39	6152.23	0	16.4	5.0	27.7	9.1	11.0	22.7			61.0	all but unassigned
avg/sum	2.3	10888.05	6152.23	0	26.9	4.3	24.4	20.5	9.6	20.0			63.9	all with unassigned

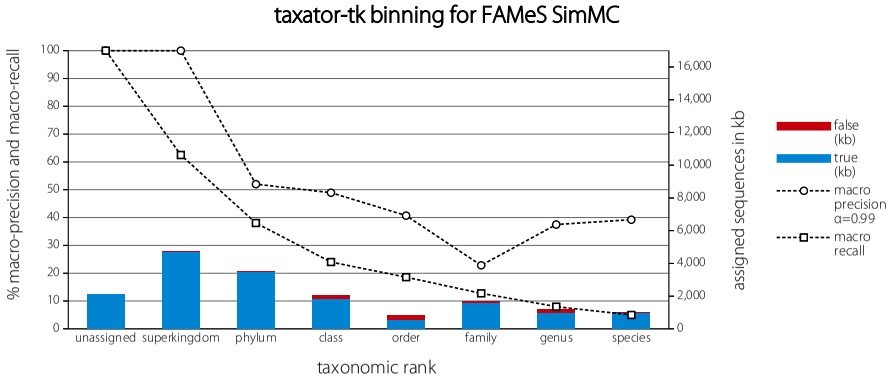


rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	2844.15	0	0	100.0	0.0	1	100.0	0.0	1	17213.77	52.43	99.7	root+superkingdom
superkingdom	1	7184.81	52.43	0	100.0	0.0	1	35.3	35.3	2				
phylum	2	0	273.68	0	0.0	0.0	14	0.0	0.0	8				
class	3	0	588.78	0	0.0	0.0	26	0.0	0.0	12	0	2861.02	0.0	phylum+class+order
order	4	0	1998.56	0	0.0	0.0	40	0.0	0.0	23				
family	5	0	762.4	0	0.0	0.0	61	0.0	0.0	30				
genus	6	0	860.34	0	0.0	0.0	72	0.0	0.0	37	0	4097.87	0.0	family+genus+species
species	7	0	2475.13	0	0.0	0.0	69	0.0	0.0	47				
avg/sum	2.3	7184.81	7011.32	0	14.3	0.0	40.4	5.0	5.0	22.7			50.6	all but unassigned
avg/sum	1.8	10028.96	7011.32	0	25.0	0.0	35.5	16.9	4.4	20.0			58.9	all with unassigned



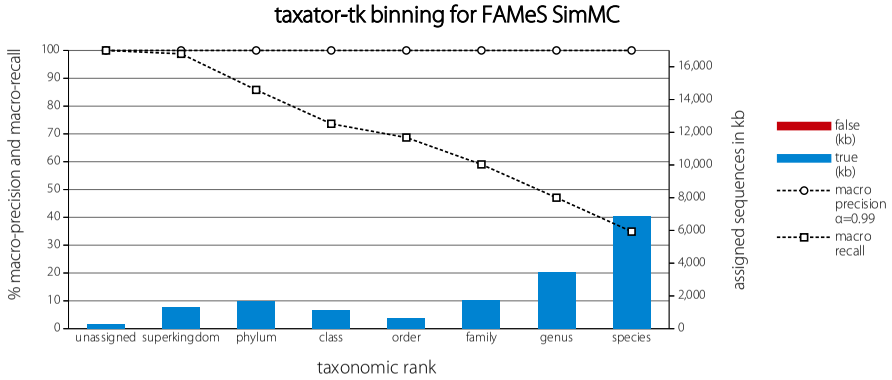
Supplementary Figure S18 - Taxator-tk binning for FAMeS SimMC **(a) summary scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	2083.8	0.0	0	100.0	0.0	1	100.0	0.0	1	11492.5	0.6	100.0	root+superkingdom
superkingdom	1	4704.3	0.6	0	100.0	0.0	1	62.5	19.6	2				
phylum	2	3460.1	26.9	0	52.0	36.2	4	38.1	13.3	8	5881.5	466.4	92.7	phylum+class+order
class	3	1860.6	182.2	0	49.0	47.6	4	24.0	14.2	12				
order	4	560.8	257.2	0	40.7	41.6	15	18.6	12.0	23				
family	5	1573.3	89.4	0	22.8	38.3	19	12.8	10.8	30	3564.0	339.7	91.3	family+genus+species
genus	6	1012.7	196.9	0	37.5	45.7	19	8.0	7.9	37				
species	7	978.0	53.3	0	39.2	48.4	54	5.0	6.3	47				
avg/sum	2.6	14149.8	806.6	0	48.7	36.8	16.6	24.1	12.0	22.7			94.6	all but unassigned
avg/sum	2.3	16233.6	806.6	0	55.2	32.2	14.6	33.6	10.5	20.0			95.3	all with unassigned



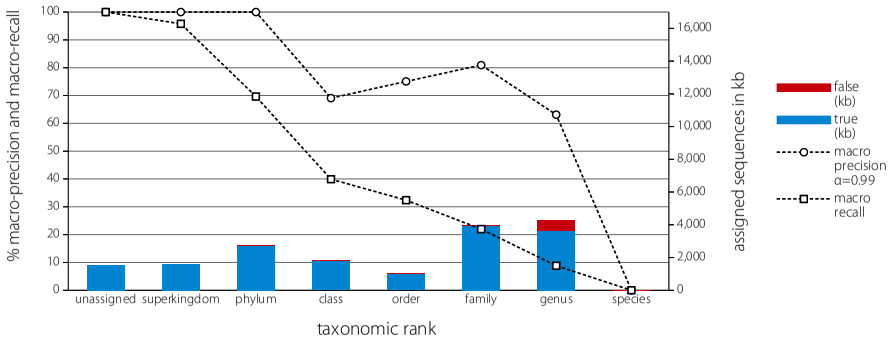
Supplementary Figure S18 - Taxator-tk binning for FAMeS SimMC **(b) all reference scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	251.07	0	0	100.0	0.0	1	100.0	0.0	1	2857.85	0	100.0	root+superkingdom
superkingdom	1	1303.39	0	0	100.0	0.0	1	98.8	1.2	2				
phylum	2	1673.66	0	0	100.0	0.0	1	85.8	15.8	8				
class	3	1129.62	0	0	100.0	0.0	2	73.6	22.1	12	3450.59	0	100.0	phylum+class+order
order	4	647.31	0	0	100.0	0.0	4	68.7	29.4	23				
family	5	1728.38	0	0	100.0	0.0	3	59.0	42.0	30				
genus	6	3460.93	0	0	100.0	0.0	5	47.1	43.6	37	12035.22	0	100.0	family+genus+species
species	7	6845.91	0	0	100.0	0.0	5	34.9	44.2	47				
avg/sum	4.0	16789.2	0	0	100.0	0.0	3.0	66.8	28.3	22.7			100.0	all but unassigned
avg/sum	3.9	17040.27	0	0	100.0	0.0	2.8	71.0	24.8	20.0			100.0	all with unassigned



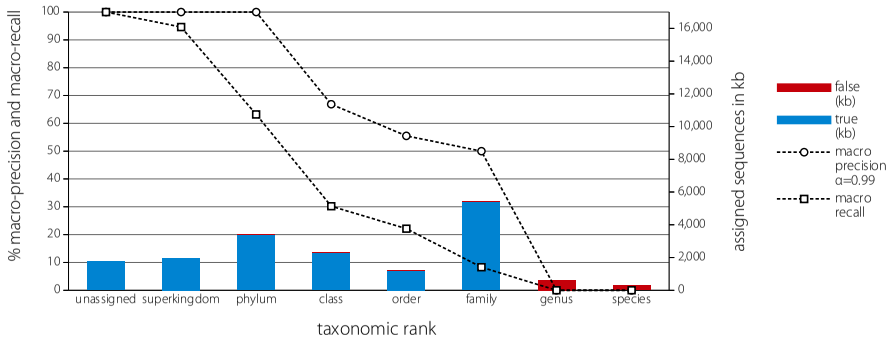
Supplementary Figure S18 - Taxator-tk binning for FAMeS SimMC **(c) new species scenario**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1558.82	0	0	100.0	0.0	1	100.0	0.0	1	4785.94	0	100.0	root+superkingdom
superkingdom	1	1613.56	0	0	100.0	0.0	1	95.8	4.2	2				
phylum	2	2761.05	2.62	0	100.0	0.0	1	69.6	30.2	8				
class	3	1806.58	27.87	0	69.1	42.1	3	39.9	35.3	12	5592.14	50.24	99.1	phylum+class+order
order	4	1024.51	19.75	0	75.1	35.6	7	32.4	34.2	23				
family	5	3915.66	33.96	0	81.0	36.6	6	22.0	34.7	30				
genus	6	3628.14	630.72	0	63.1	44.8	3	8.8	24.1	37	7543.8	681.73	91.7	family+genus+species
species	7	0	17.05	0	0.0	0.0	11	0.0	0.0	47				
avg/sum	4.0	14749.5	731.97	0	69.7	22.7	4.6	38.4	23.2	22.7			95.3	all but unassigned
avg/sum	3.6	16308.32	731.97	0	73.5	19.9	4.1	46.1	20.3	20.0			95.7	all with unassigned

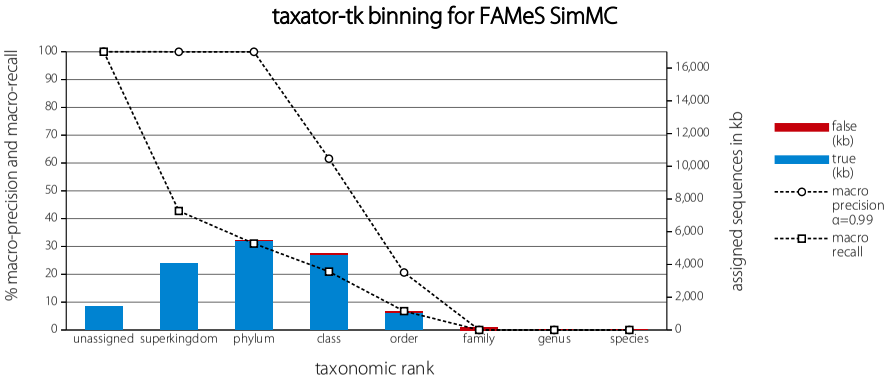


Supplementary Figure S18 - Taxator-tk binning for FAMeS SimMC **(d) new genus scenario**

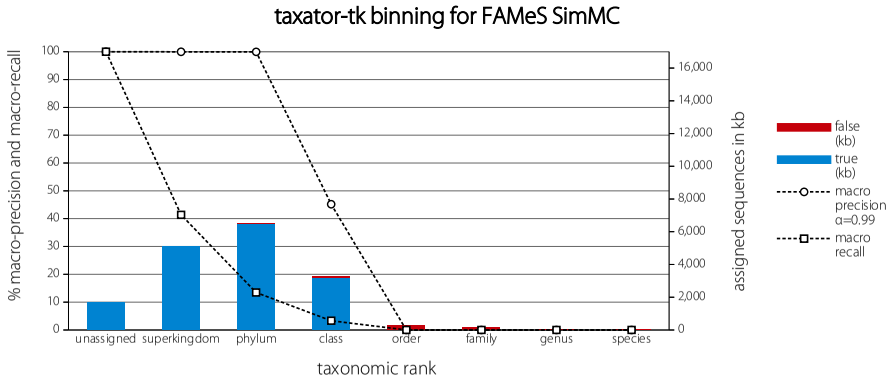
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1745.53	0	0	100.0	0.0	1	100.0	0.0	1	5698.63	0	100.0	root+superkingdom
superkingdom	1	1976.55	0	0	100.0	0.0	1	94.6	5.4	2				
phylum	2	3398.87	2.62	0	100.0	0.0	1	63.2	32.7	8				
class	3	2312.46	32.37	0	66.9	44.9	3	30.2	32.2	12	6900.35	75.58	98.9	phylum+class+order
order	4	1189.02	40.59	0	55.5	42.8	7	22.2	32.0	23				
family	5	5368.74	54.74	0	50.0	46.3	7	8.3	22.3	30				
genus	6	0	636.48	0	0.0	0.0	15	0.0	0.0	37	5368.74	973.52	84.7	family+genus+species
species	7	0	282.3	0	0.0	0.0	9	0.0	0.0	47				
avg/sum	3.4	14245.64	1049.1	0	53.2	19.2	6.1	31.2	17.8	22.7			93.1	all but unassigned
avg/sum	3.1	15991.17	1049.1	0	59.0	16.8	5.5	39.8	15.6	20.0			93.8	all with unassigned



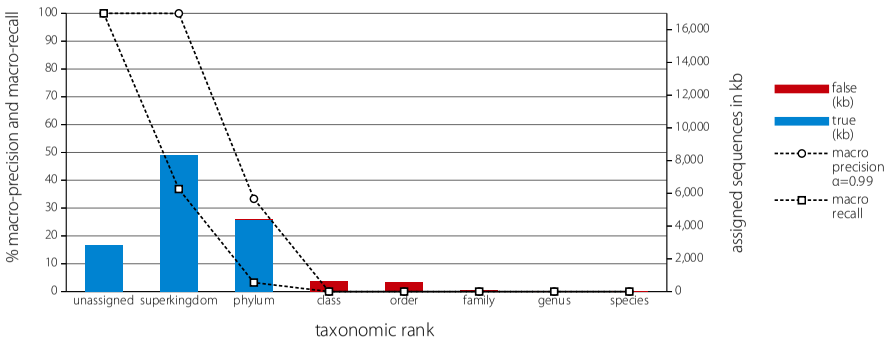
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1444.21	0	0	100.0	0.0	1	100.0	0.0	1	9543.73	0	100.0	root+superkingdom
superkingdom	1	4049.76	0	0	100.0	0.0	1	42.8	42.8	2				
phylum	2	5463.82	11.04	0	100.0	0.0	1	31.1	36.5	8	11135.95	187.88	98.3	phylum+class+order
class	3	4607.11	80.28	0	61.6	43.9	3	21.0	31.0	12				
order	4	1065.02	96.56	0	20.6	37.4	18	6.8	17.0	23				
family	5	0	179.84	0	0.0	0.0	21	0.0	0.0	30	0	222.49	0.0	family+genus+species
genus	6	0	32.8	0	0.0	0.0	14	0.0	0.0	37				
species	7	0	9.85	0	0.0	0.0	7	0.0	0.0	47				
avg/sum	2.2	15185.71	410.37	0	40.3	11.6	9.3	14.5	18.2	22.7			97.4	all but unassigned
avg/sum	2.0	16629.92	410.37	0	47.8	10.2	8.3	25.2	15.9	20.0			97.6	all with unassigned



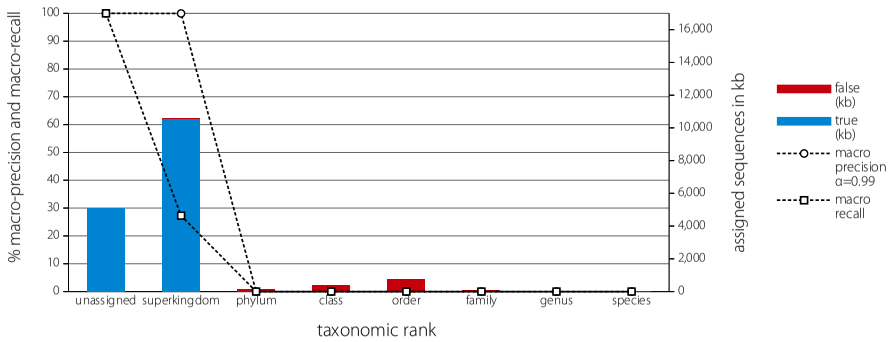
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1665.24	0	0	100.0	0.0	1	100.0	0.0	1	11800.28	0	100.0	root+superkingdom
superkingdom	1	5067.52	0	0	100.0	0.0	1	41.4	41.4	2				
phylum	2	6525.5	14.12	0	100.0	0.0	1	13.5	20.3	8	9694.14	388.89	96.1	phylum+class+order
class	3	3168.64	86.66	0	45.2	45.6	4	3.3	6.6	12				
order	4	0	288.11	0	0.0	0.0	19	0.0	0.0	23				
family	5	0	169.88	0	0.0	0.0	17	0.0	0.0	30	0	224.49	0.0	family+genus+species
genus	6	0	40.01	0	0.0	0.0	14	0.0	0.0	37				
species	7	0	14.6	0	0.0	0.0	9	0.0	0.0	47				
avg/sum	2.0	14761.66	613.38	0	35.0	6.5	9.3	8.3	9.8	22.7			96.0	all but unassigned
avg/sum	1.8	16426.9	613.38	0	43.1	5.7	8.3	19.8	8.5	20.0			96.4	all with unassigned



rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	2853.36	0	0	100.0	0.0	1	100.0	0.0	1	19566.78	0	100.0	root+superkingdom
superkingdom	1	8356.71	0	0	100.0	0.0	1	36.9	36.9	2				
phylum	2	4397.58	25.84	0	33.3	47.1	3	3.2	8.5	8	4397.58	1282.66	77.4	phylum+class+order
class	3	0	659.76	0	0.0	0.0	11	0.0	0.0	12				
order	4	0	597.06	0	0.0	0.0	18	0.0	0.0	23				
family	5	0	108.29	0	0.0	0.0	21	0.0	0.0	30	0	149.96	0.0	family+genus+species
genus	6	0	23.82	0	0.0	0.0	14	0.0	0.0	37				
species	7	0	17.85	0	0.0	0.0	9	0.0	0.0	47				
avg/sum	1.6	12754.29	1432.62	0	19.0	6.7	11.0	5.7	6.5	22.7			89.9	all but unassigned
avg/sum	1.3	15607.65	1432.62	0	29.2	5.9	9.8	17.5	5.7	20.0			91.6	all with unassigned

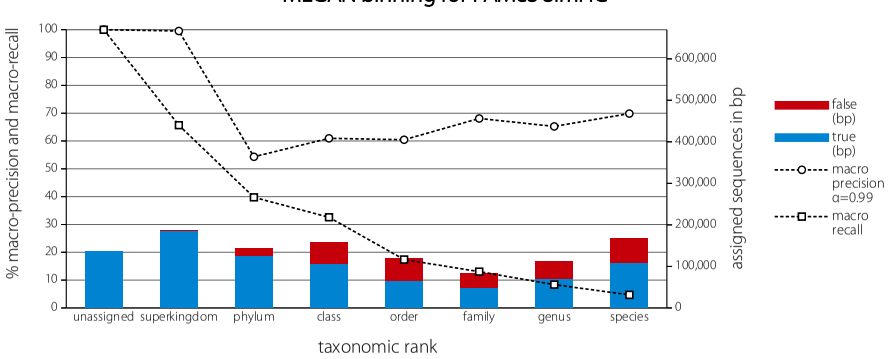


rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	5068.27	0	0	100.0	0.0	1	100.0	0.0	1	26194.15	4.05	100.0	root+superkingdom
superkingdom	1	10562.94	4.05	0	100.0	0.0	1	27.3	27.3	2				
phylum	2	0	132.26	0	0.0	0.0	10	0.0	0.0	8				
class	3	0	388.44	0	0.0	0.0	18	0.0	0.0	12	0	1279.35	0.0	phylum+class+order
order	4	0	758.65	0	0.0	0.0	19	0.0	0.0	23				
family	5	0	79.35	0	0.0	0.0	24	0.0	0.0	30				
genus	6	0	14.55	0	0.0	0.0	17	0.0	0.0	37	0	125.67	0.0	family+genus+species
species	7	0	31.77	0	0.0	0.0	14	0.0	0.0	47				
avg/sum	1.3	10562.94	1409.07	0	14.3	0.0	14.7	3.9	3.9	22.7			88.2	all but unassigned
avg/sum	0.9	15631.21	1409.07	0	25.0	0.0	13.0	15.9	3.4	20.0			91.7	all with unassigned



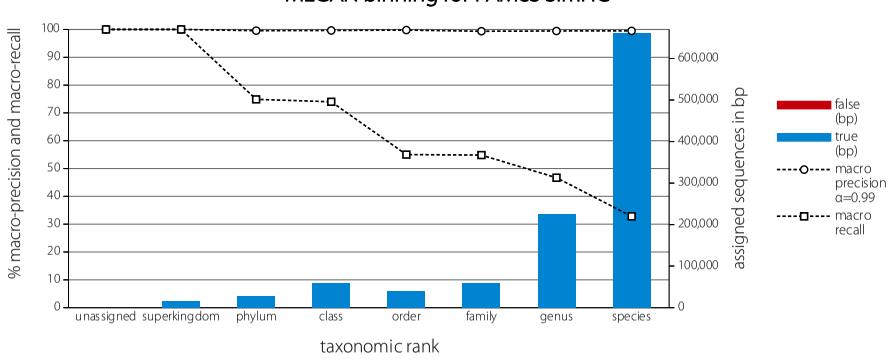
Supplementary Figure S19 - MEGAN binning for FAMeS SimHC **(a) summary scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	135097.7	0.0	0.0	100.0	0.0	1	100.0	0.0	1	504932.0	1240.4	99.8	root+superkingdom
superkingdom	1	184917.1	1240.4	0.0	99.5	0.0	1	65.7	21.3	2				
phylum	2	126186.3	17996.0	0.0	54.3	45.2	10	39.8	24.2	8	297237.9	124491.1	70.5	phylum+class+order
class	3	105637.4	52554.0	2704.9	61.0	36.9	12	32.6	19.7	12				
order	4	65414.1	53941.1	0.0	60.5	41.5	27	17.4	17.2	36				
family	5	47775.7	34408.6	0.0	68.1	38.6	36	13.1	13.3	52	228155.2	135050.2	62.8	family+genus+species
genus	6	70368.1	42132.0	382.3	65.3	43.0	47	8.4	9.4	72				
species	7	110011.3	58509.7	356.5	69.9	45.6	47	4.7	6.5	96				
avg/sum	3.1	710310.2	260781.8	3443.6	68.4	35.8	25.7	25.9	15.9	39.7			73.1	all but unassigned
avg/sum	2.7	845407.9	260781.8	3443.6	72.3	31.3	22.6	35.2	14.0	34.9			76.4	all with unassigned



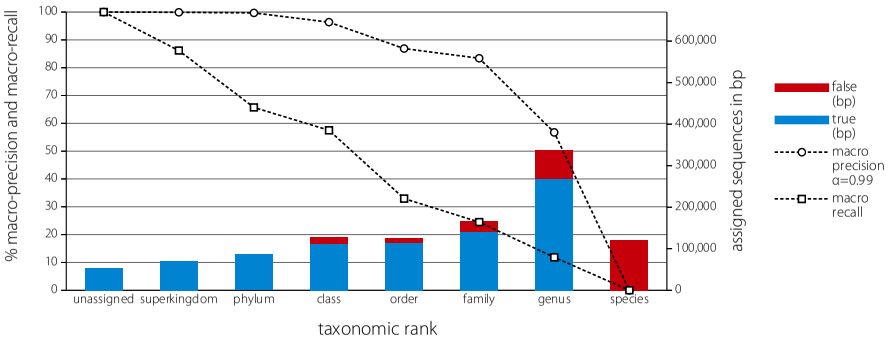
Supplementary Figure S19 - MEGAN binning for FAMeS SimHC **(b) all reference scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	0	0	0	100.0	0.0	1	100.0	0.0	1	29660	0	100.0	root+superkingdom
superkingdom	1	14830	0	0	100.0	0.0	2	100.0	0.0	2				
phylum	2	27071	0	0	99.6	1.0	6	74.9	43.2	8	123849	0	100.0	phylum+class+order
class	3	58344	0	0	99.6	0.8	9	74.0	42.8	12				
order	4	38434	0	0	99.8	0.7	20	55.0	49.3	36				
family	5	58139	0	0	99.4	3.1	29	54.8	49.1	52	942014	0	100.0	family+genus+species
genus	6	223807	0	2676	99.5	2.8	34	46.7	49.5	72				
species	7	660068	0	2139	99.5	2.9	33	32.8	45.8	96				
avg/sum	4.8	1080693	0	4815	99.6	1.6	19.0	62.6	39.9	39.7			100.0	all but unassigned
avg/sum	4.8	1080693	0	4815	99.7	1.4	16.8	67.3	35.0	34.9			100.0	all with unassigned



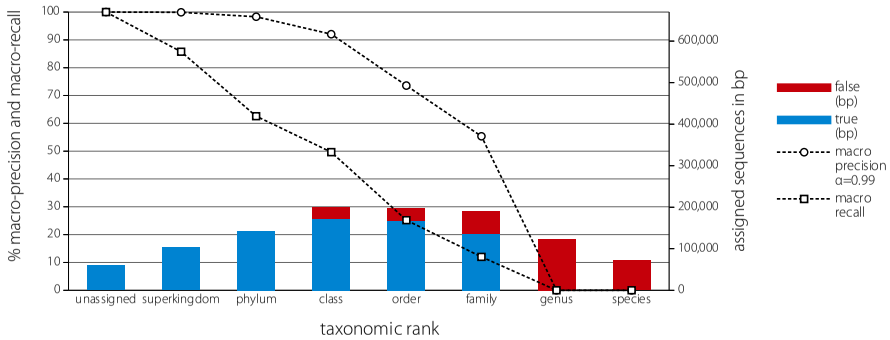
Supplementary Figure S19 - MEGAN binning for FAMeS SimHC **(c) new species scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	53833	0	0	100.0	0.0	1	100.0	0.0	1	196059	0	100.0	root+superkingdom
superkingdom	1	71113	0	0	99.9	0.1	2	86.2	8.4	2				
phylum	2	85694	0	0	99.7	0.5	6	65.7	38.2	8	311019	25259	92.5	phylum+class+order
class	3	111601	14697	0	96.4	4.6	9	57.5	34.8	12				
order	4	113724	10562	0	86.9	29.7	18	33.0	38.4	36				
family	5	140308	25633	0	83.4	27.0	20	24.5	35.6	52	409078	215206	65.5	family+genus+species
genus	6	268770	68288	0	56.8	44.8	17	11.9	29.1	72				
species	7	0	121285	0	0.0	0.0	8	0.0	0.0	96				
avg/sum	4.4	791210	240465	0	74.7	15.2	11.4	39.8	26.4	39.7			76.7	all but unassigned
avg/sum	4.1	845043	240465	0	77.9	13.3	10.1	47.3	23.1	34.9			77.8	all with unassigned

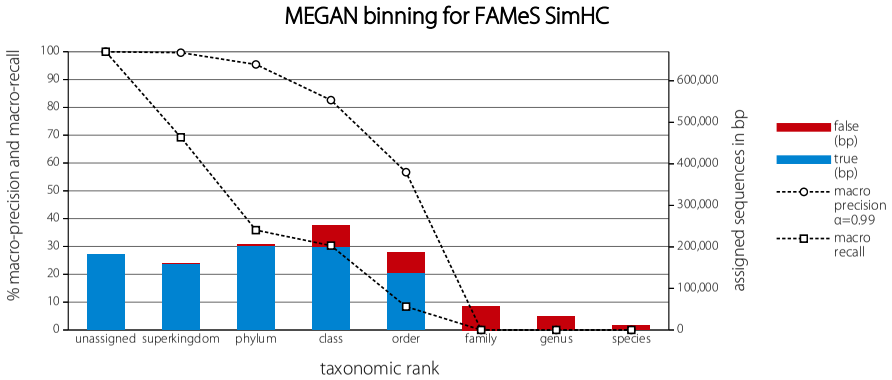


Supplementary Figure S19 - MEGAN binning for FAMeS SimHC **(d) new genus scenario**

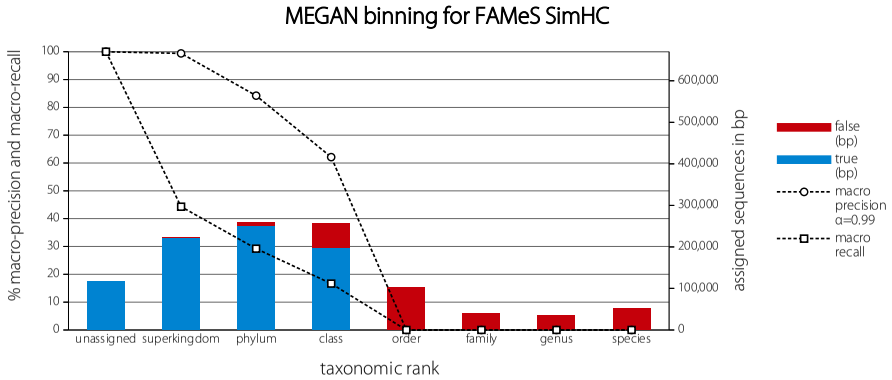
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	60008	0	0	100.0	0.0	1	100.0	0.0	1	266134	0	100.0	root+superkingdom
superkingdom	1	103063	0	0	99.9	0.1	2	85.7	8.0	2				
phylum	2	142075	0	0	98.3	2.3	6	62.6	36.7	8	481779	57207	89.4	phylum+class+order
class	3	171151	27602	0	92.0	6.7	9	49.7	31.2	12				
order	4	168553	29605	0	73.6	34.2	18	25.2	32.4	36				
family	5	135983	53126	0	55.4	39.5	16	12.0	25.2	52	135983	247468	35.5	family+genus+species
genus	6	0	123176	0	0.0	0.0	12	0.0	0.0	72				
species	7	0	71166	0	0.0	0.0	5	0.0	0.0	96				
avg/sum	3.6	720825	304675	0	59.9	11.8	9.7	33.6	19.1	39.7			70.3	all but unassigned
avg/sum	3.4	780833	304675	0	64.9	10.4	8.6	41.9	16.7	34.9			71.9	all with unassigned



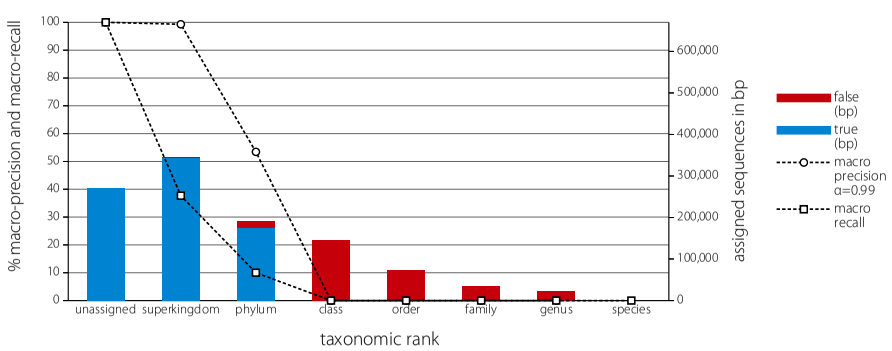
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	180596	0	0	100.0	0.0	1	100.0	0.0	1	498492	1776	99.6	root+superkingdom
superkingdom	1	158948	1776	0	99.7	0.3	2	69.3	13.7	2				
phylum	2	202392	3264	0	95.4	4.8	5	35.9	33.9	8	540433	103443	83.9	phylum+class+order
class	3	200853	52078	0	82.6	9.8	7	30.3	27.9	12				
order	4	137188	48101	0	56.7	33.0	11	8.4	17.2	36				
family	5	0	57702	0	0.0	0.0	8	0.0	0.0	52	0	100312	0.0	family+genus+species
genus	6	0	31996	0	0.0	0.0	3	0.0	0.0	72				
species	7	0	10614	0	0.0	0.0	1	0.0	0.0	96				
avg/sum	2.9	699381	205531	0	47.8	6.8	5.3	20.5	13.2	39.7			77.3	all but unassigned
avg/sum	2.4	879977	205531	0	54.3	6.0	4.8	30.5	11.6	34.9			81.1	all with unassigned



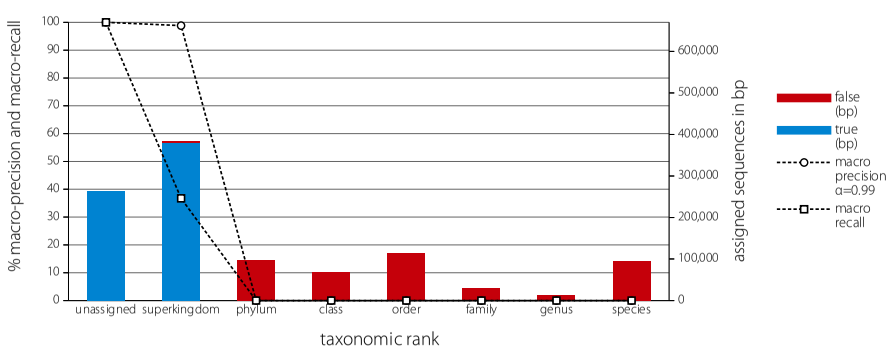
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	115421	0	0	100.0	0.0	1	100.0	0.0	1	558349	1776	99.7	root+superkingdom
superkingdom	1	221464	1776	0	99.4	0.0	1	44.3	44.3	2				
phylum	2	250454	8635	0	84.2	18.2	4	29.2	34.1	8	447967	169429	72.6	phylum+class+order
class	3	197513	58615	1390	62.1	16.6	7	16.7	18.1	12				
order	4	0	102179	0	0.0	0.0	12	0.0	0.0	36				
family	5	0	40254	0	0.0	0.0	6	0.0	0.0	52	0	128061	0.0	family+genus+species
genus	6	0	35656	0	0.0	0.0	5	0.0	0.0	72				
species	7	0	52151	0	0.0	0.0	2	0.0	0.0	96				
avg/sum	2.5	669431	299266	1390	35.1	5.0	5.3	12.9	13.8	39.7			69.1	all but unassigned
avg/sum	2.3	784852	299266	1390	43.2	4.3	4.8	23.8	12.1	34.9			72.4	all with unassigned



rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	271642	0	0	100.0	0.0	1	100.0	0.0	1	958974	1776	99.8	root+superkingdom
superkingdom	1	343666	1776	0	99.3	0.0	1	37.7	37.7	2				
phylum	2	175618	16711	0	53.5	38.2	3	10.0	18.1	8	175618	235397	42.7	phylum+class+order
class	3	0	146037	0	0.0	0.0	7	0.0	0.0	12				
order	4	0	72649	0	0.0	0.0	9	0.0	0.0	36				
family	5	0	34445	0	0.0	0.0	5	0.0	0.0	52	0	57409	0.0	family+genus+species
genus	6	0	22964	0	0.0	0.0	2	0.0	0.0	72				
species	7							0.0	0.0	96				
avg/sum	2.2	519284	294582	0	25.5	6.4	4.5	6.8	8.0	39.7			63.8	all but unassigned
avg/sum	1.6	790926	294582	0	36.1	5.5	4.0	18.5	7.0	34.9			72.9	all with unassigned

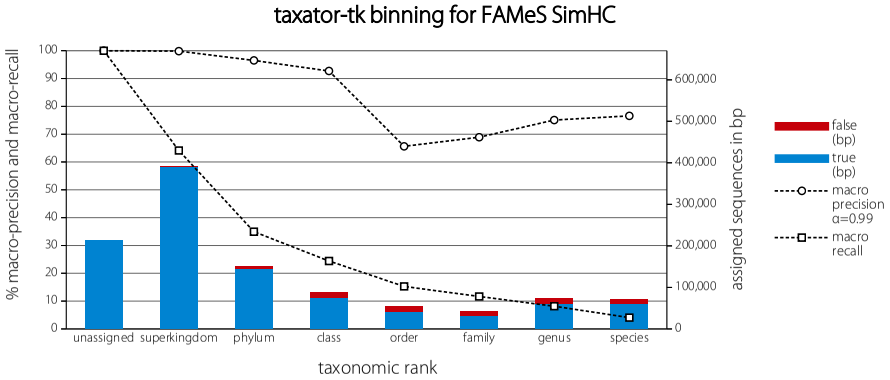


rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec	description
unassigned	0	264184	0	0	100.0	0.0	1	100.0	0.0	1	1026856	3355	99.7	root+superkingdom
superkingdom	1	381336	3355	0	98.8	0.0	1	36.7	36.7	2				
phylum	2	0	97362	0	0.0	0.0	12	0.0	0.0	8	0	280703	0.0	phylum+class+order
class	3	0	68849	17544	0.0	0.0	11	0.0	0.0	12				
order	4	0	114492	0	0.0	0.0	10	0.0	0.0	36				
family	5	0	29700	0	0.0	0.0	6	0.0	0.0	52	0	138386	0.0	family+genus+species
genus	6	0	12844	0	0.0	0.0	4	0.0	0.0	72				
species	7	0	95842	0	0.0	0.0	5	0.0	0.0	96				
avg/sum	2.1	381336	422444	17544	14.1	0.0	7.0	5.2	5.2	39.7			47.4	all but unassigned
avg/sum	1.5	645520	422444	17544	24.9	0.0	6.3	17.1	4.6	34.9			60.4	all with unassigned



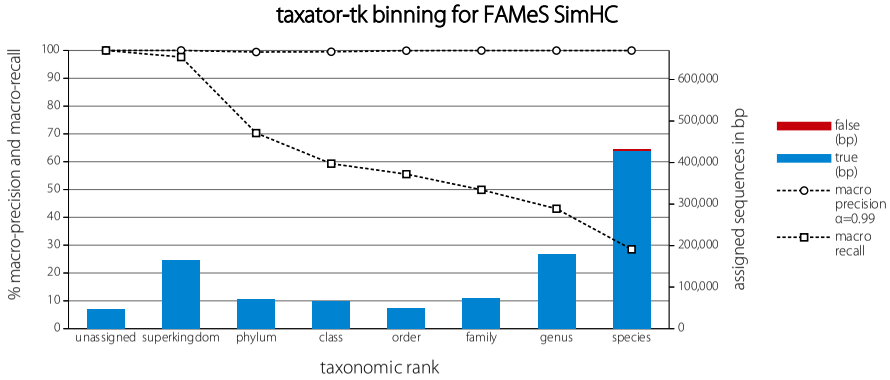
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(a) summary scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	213863.4	0.0	0.0	100.0	0.0	1	100.0	0.0	1	993072.9	1240.4	99.9	root+superkingdom
superkingdom	1	389604.7	1240.4	0.0	99.9	0.1	2	64.1	14.9	2				
phylum	2	144608.1	5934.9	0.0	96.5	5.1	7	34.9	7.9	8	259408.7	32094.3	89.0	phylum+class+order
class	3	74893.3	12043.4	757.0	92.7	8.4	11	24.4	10.3	12				
order	4	39907.3	14116.0	0.0	65.6	44.8	47	15.3	10.0	36				
family	5	31822.9	11110.7	0.0	68.9	43.0	58	11.7	9.7	52	153059.7	34715.1	81.5	family+genus+species
genus	6	59831.9	13687.3	382.3	75.1	40.3	68	8.2	7.5	72				
species	7	61405.0	9917.1	382.3	76.6	41.5	66	4.1	4.8	96				
avg/sum	2.3	802073.1	68049.9	1521.6	82.2	26.2	37.0	23.2	9.3	39.7			92.2	all but unassigned
avg/sum	1.8	1015936.6	68049.9	1521.6	84.4	22.9	32.5	32.8	8.1	34.9			93.7	all with unassigned



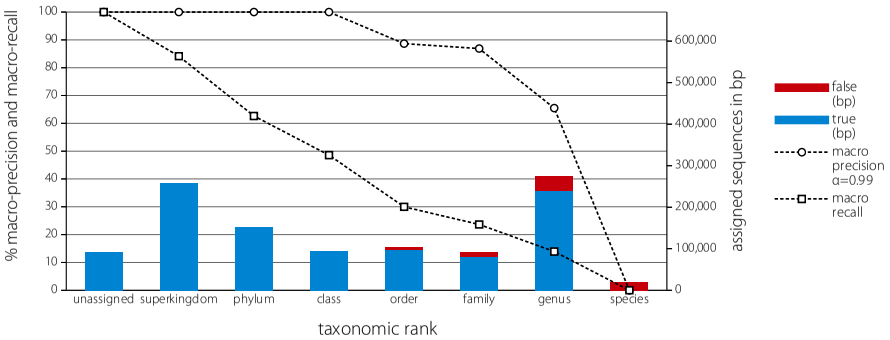
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(b) all reference scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	47885		0	100.0	0.0	1	100.0	0.0	1	375435	0	100.0	root+superkingdom
superkingdom	1	163775	(bp)	0	100.0	0.0	2	97.6	2.4	2				
phylum	2	70870	0	0	99.4	1.2	6	70.3	28.5	8				
class	3	65908	0	2139	99.5	1.1	9	59.3	28.9	12	185117	0	100.0	phylum+class+order
order	4	48339	0	0	99.9	0.4	32	55.5	30.6	36				
family	5	72448	0	0	99.9	0.4	43	49.9	34.0	52				
genus	6	179113	0	2676	100.0	0.3	55	43.1	35.7	72	681396	2520	99.6	family+genus+species
species	7	429835	2520	0	100.0	0.3	52	28.5	33.5	96				
avg/sum	3.6	1030288	2520	4815	99.8	0.5	28.4	57.8	27.7	39.7			99.8	all but unassigned
avg/sum	3.3	1078173	2520	4815	99.8	0.5	25.0	63.0	24.2	34.9			99.8	all with unassigned



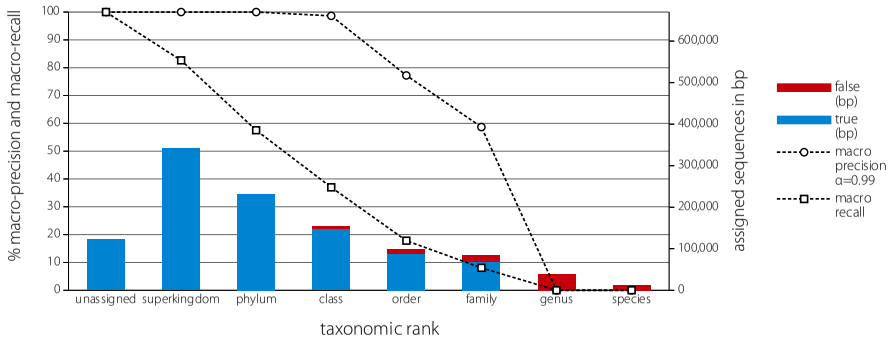
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(c) new species scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	92565		0	100.0	0.0	1	100.0	0.0	1	608487	0	100.0	root+superkingdom
superkingdom	1	257961	(bp)	0	100.0	0.0	2	84.1	6.4	2				
phylum	2	152494	0	0	100.0	0.0	7	62.6	26.6	8	343148	6930	98.0	phylum+class+order
class	3	94075	0	0	100.0	0.0	10	48.6	28.5	12				
order	4	96579	6930	0	88.6	28.8	26	30.0	30.1	36				
family	5	80651	10331	0	86.9	30.4	30	23.7	31.8	52	320361	64543	83.2	family+genus+species
genus	6	239710	34526	0	65.5	46.1	31	13.9	26.1	72				
species	7	0	19686	0	0.0	0.0	9	0.0	0.0	96				
avg/sum	3.5	921470	71473	0	77.3	15.0	16.4	37.6	21.3	39.7			92.8	all but unassigned
avg/sum	3.2	1014035	71473	0	80.1	13.2	14.5	45.4	18.7	34.9			93.4	all with unassigned



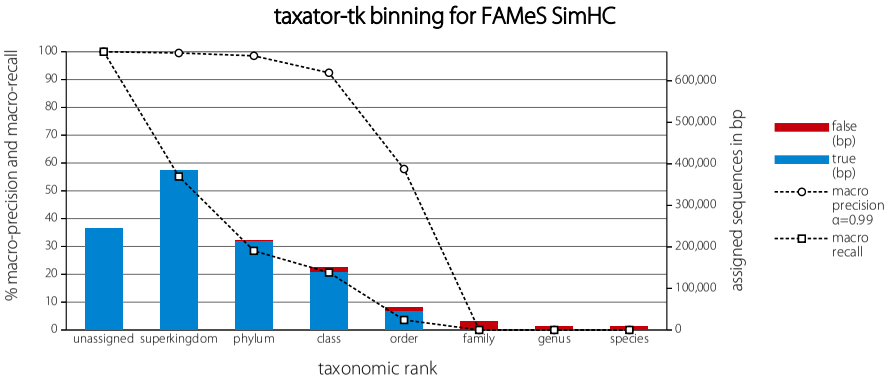
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(d) new genus scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	121975		0	100.0	0.0	1	100.0	0.0	1	804007	0	100.0	root+superkingdom
superkingdom	1	341016	(bp)	0	100.0	0.0	2	82.7	4.9	2				
phylum	2	232318	0	0	100.0	0.0	7	57.5	14.0	8	468720	17552	96.4	phylum+class+order
class	3	148392	5857	0	98.6	2.6	11	37.0	20.9	12				
order	4	88010	11695	0	77.2	38.5	23	17.8	24.3	36				
family	5	69661	14936	0	58.7	46.8	20	8.1	19.0	52				
genus	6	0	39229	0	0.0	0.0	14	0.0	0.0	72	69661	66584	51.1	family+genus+species
species	7	0	12419	0	0.0	0.0	7	0.0	0.0	96				
avg/sum	2.4	879397	84136	0	62.1	12.6	12.0	29.0	11.9	39.7			91.3	all but unassigned
avg/sum	2.2	1001372	84136	0	66.8	11.0	10.6	37.9	10.4	34.9			92.2	all with unassigned



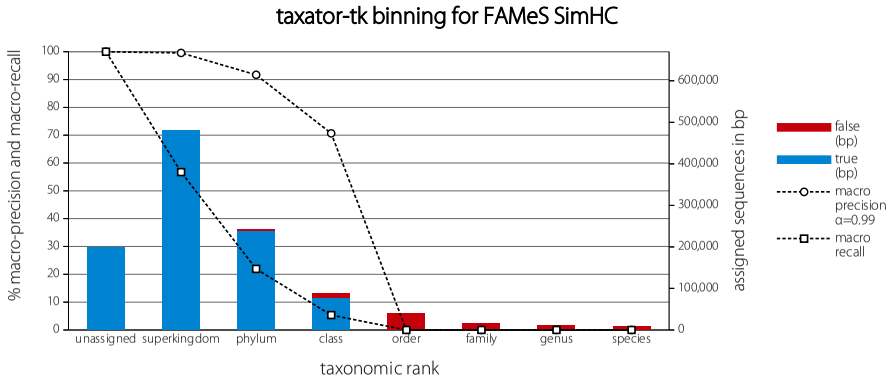
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(e) new family scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	243406		0	100.0	0.0	1	100.0	0.0	1	1010134	0	100.0	root+superkingdom
superkingdom	1	383364	(bp)	0	99.6	0.0	1	55.2	21.8	2				
phylum	2	215052	1898	0	98.5	1.9	7	28.4	17.2	8	400621	21498	94.9	phylum+class+order
class	3	139146	11207	0	92.4	11.7	10	20.6	14.9	12				
order	4	46423	8393	0	57.8	44.5	12	3.6	9.5	36				
family	5	0	19955	0	0.0	0.0	11	0.0	0.0	52	0	34843	0.0	family+genus+species
genus	6	0	7389	0	0.0	0.0	6	0.0	0.0	72				
species	7	0	7499	0	0.0	0.0	3	0.0	0.0	96				
avg/sum	2.0	783985	56341	0	49.8	8.3	7.1	15.4	9.1	39.7			93.3	all but unassigned
avg/sum	1.5	1027391	56341	0	56.0	7.3	6.4	26.0	7.9	34.9			94.8	all with unassigned



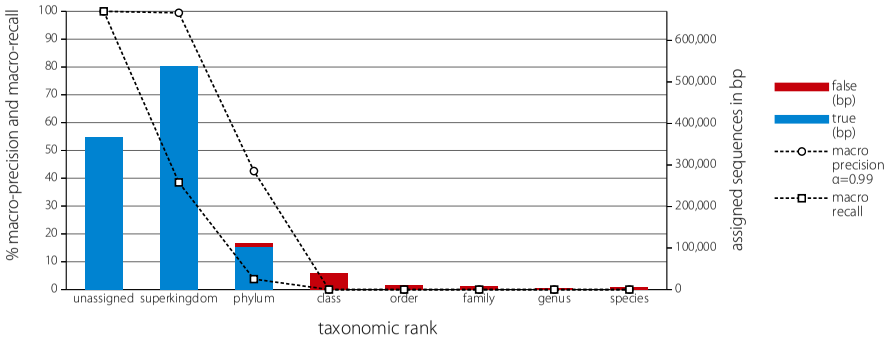
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(f) new order scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	197736		0	100.0	0.0	1	100.0	0.0	1	1157742	0	100.0	root+superkingdom
superkingdom	1	480003	(bp)	0	99.6	0.0	1	56.7	23.4	2				
phylum	2	238652	3399	0	91.7	12.7	5	21.9	17.7	8				
class	3	76732	11575	0	70.7	25.1	7	5.3	5.9	12	315384	53711	85.4	phylum+class+order
order	4	0	38737	0	0.0	0.0	13	0.0	0.0	36				
family	5	0	16635	0	0.0	0.0	9	0.0	0.0	52				
genus	6	0	11103	0	0.0	0.0	6	0.0	0.0	72	0	36898	0.0	family+genus+species
species	7	0	9160	0	0.0	0.0	3	0.0	0.0	96				
avg/sum	1.7	795387	90609	0	37.4	5.4	6.3	12.0	6.7	39.7			89.8	all but unassigned
avg/sum	1.4	993123	90609	0	45.2	4.7	5.6	23.0	5.9	34.9			91.6	all with unassigned



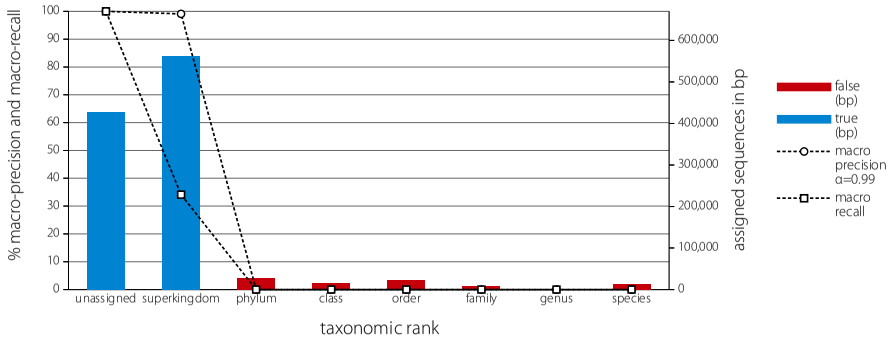
Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(g) new class scenario**

rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	366106		0	100.0	0.0	1	100.0	0.0	1	1442954	0	100.0	root+superkingdom
superkingdom	1	538424	(bp)	0	99.5	0.0	1	38.5	27.4	2				
phylum	2	102871	9200	0	42.6	42.8	4	3.8	7.3	8	102871	60229	63.1	phylum+class+order
class	3	0	39497	0	0.0	0.0	9	0.0	0.0	12				
order	4	0	11532	0	0.0	0.0	8	0.0	0.0	36				
family	5	0	7485	0	0.0	0.0	6	0.0	0.0	52				
genus	6	0	3564	0	0.0	0.0	2	0.0	0.0	72	0	16102	0.0	family+genus+species
species	7	0	5053	0	0.0	0.0	2	0.0	0.0	96				
avg/sum	1.4	641295	76331	0	20.3	6.1	4.6	6.0	5.0	39.7			89.4	all but unassigned
avg/sum	0.9	1007401	76331	0	30.3	5.3	4.1	17.8	4.3	34.9			93.0	all with unassigned



Supplementary Figure S20 - Taxator-tk binning for FAMeS SimHC **(h)new phylum scenario**

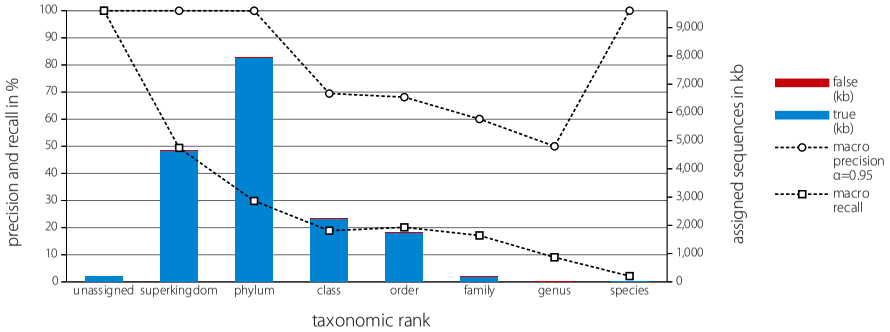
rank	depth	true (bp)	false (bp)	unknown (bp)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (bp)	sum false (bp)	overall prec.	description
unassigned	0	427371		0	100.0	0.0	1	100.0	0.0	1	1552751	0	100.0	root+superkingdom
superkingdom	1	562690	(bp)	0	99.1	0.0	1	34.1	23.0	2				
phylum	2	0	27047	0	0.0	0.0	11	0.0	0.0	8				
class	3	0	16168	3160	0.0	0.0	11	0.0	0.0	12	0	64740	0.0	phylum+class+order
order	4	0	21525	0	0.0	0.0	10	0.0	0.0	36				
family	5	0	8433	0	0.0	0.0	7	0.0	0.0	52				
genus	6	0	0	0	0.0	0.0	3	0.0	0.0	72	0	21516	0.0	family+genus+species
species	7	0	13083	2676	0.0	0.0	5	0.0	0.0	96				
avg/sum	1.2	562690	86256	5836	14.2	0.0	6.9	4.9	3.3	39.7			86.7	all but unassigned
avg/sum	0.7	990061	86256	5836	24.9	0.0	6.1	16.8	2.9	34.9			92.0	all with unassigned



Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(a) taxator-tk (nucleotide)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.95	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	199.2	0	0	100.0	0.0	1	100.0	0.0	1	9526.04	2.03	100.0	root+superkingdom
superkingdom	1	4663.42	2.03	0	100.0	0.0	1	49.4	49.4	2				
phylum	2	7936.17	2.11	0	100.0	0.0	1	29.9	36.1	8	11881.62	56.97	99.5	phylum+class+order
class	3	2215.89	25.89	0	69.5	42.7	3	18.9	28.7	12				
order	4	1729.56	28.97	0	68.1	45.1	6	20.1	32.0	22				
family	5	191.38	13.42	0	60.1	47.2	14	17.1	33.0	29	212.11	24.95	89.5	family+genus+species
genus	6	19	11.53	0	50.0	50.0	10	9.1	25.3	37				
species	7	1.73	0	0	100.0	0.0	1	2.1	14.4	47				
avg./sum	2.1	16757.15	83.95	0	78.2	26.4	5.1	20.9	31.3	22.4			99.5	all but unassigned
avg./sum	2.1	16956.35	83.95	0	81.0	23.1	4.6	30.8	27.4	19.8			99.5	all with unassigned

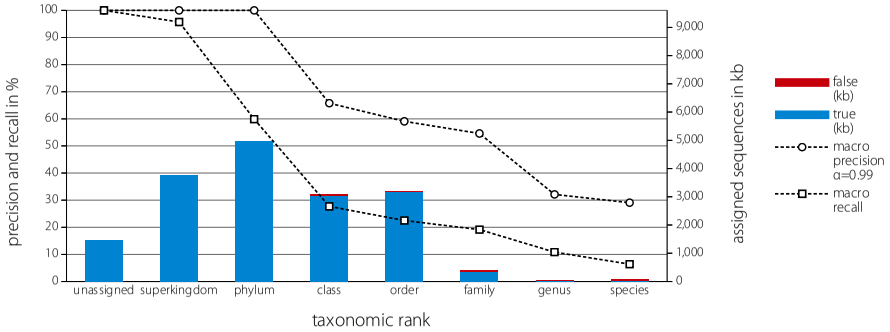
taxtor-tk binning for FAMeS SimMC scenario (Nature Methods 2011)
using nucleotide-level alignment



Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(b) taxator-tk (amino acid)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1461.8	0	0	100.0	0.0	1	100.0	0.0	1	9038.52	0	100.0	root+superkingdom
superkingdom	1	3788.36	0	0	100.0	0.0	1	95.7	4.3	2				
phylum	2	4989.57	0	0	100.0	0.0	1	59.9	33.3	8				
class	3	3035.46	58.8	0	65.8	46.5	3	27.7	30.1	12	11175.42	93.58	99.2	phylum+class+order
order	4	3150.39	34.78	0	59.1	44.7	7	22.5	32.4	22				
family	5	347.7	56.72	0	54.6	49.0	18	19.2	35.2	29				
genus	6	17.35	31.58	0	32.1	45.7	18	10.9	28.9	37	377.28	143.85	72.4	family+genus+species
species	7	12.23	55.55	0	29.1	44.0	13	6.4	24.4	47				
avg./sum	2.5	15341.06	237.43	0	63.0	32.8	8.7	34.6	26.9	22.4			98.5	all but unassigned
avg./sum	2.2	16802.86	237.43	0	67.6	28.7	7.8	42.8	23.6	19.8			98.6	all with unassigned

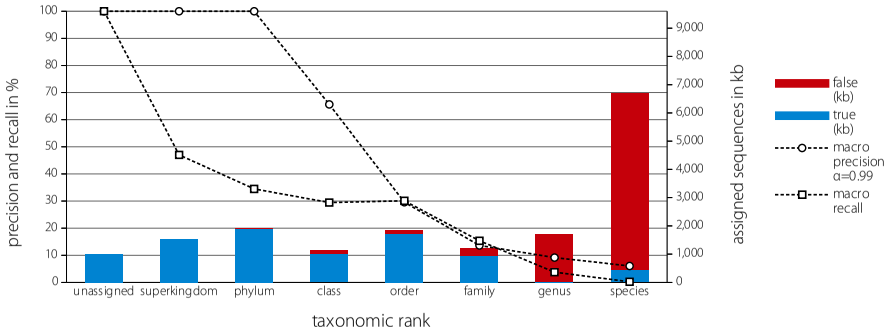
taxtor-tk binning for FAMeS SimMC scenario (Nature Methods 2011)
using protein-level alignment



Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(c) MEGAN4 (nucleotide)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1014.03	0	0	100.0	0.0	1	100.0	0.0	1	4082.47	0	100.0	root+superkingdom
superkingdom	1	1534.22	0	0	100.0	0.0	1	47.0	47.0	2				
phylum	2	1896.2	1.83	0	100.0	0.0	1	34.5	38.5	8				
class	3	1021.99	108.21	0	65.6	45.6	3	29.4	31.7	12	4643.84	228.86	95.3	phylum+class+order
order	4	1725.65	118.82	0	29.6	40.6	11	30.1	35.2	22				
family	5	935.47	266.12	0	13.6	32.0	17	15.3	30.8	29				
genus	6	18.97	1684.92	0	9.2	25.9	36	3.8	14.5	37	1411.9	8207.44	14.7	family+genus+species
species	7	457.46	6256.4	0	6.1	21.2	38	0.2	1.0	47				
avg./sum	3.5	7589.96	8436.3	0	46.3	23.6	15.3	22.9	28.4	22.4			47.4	all but unassigned
avg./sum	3.1	8603.99	8436.3	0	53.0	20.7	13.5	32.5	24.8	19.8			50.5	all with unassigned

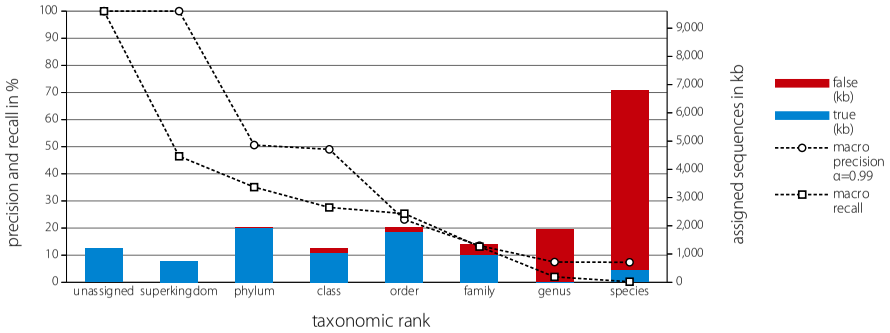
MEGAN4 binning for FAMeS SimMC scenario (Nature Methods 2011)
using nucleotide-level alignment



Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(d) MEGAN5 (nucleotide)**

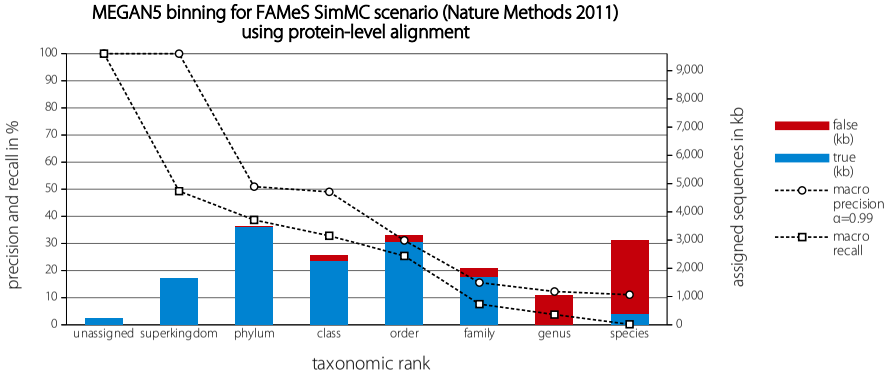
rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1198.19	0	0	100.0	0.0	1	100.0	0.0	1	2696.73	0	100.0	root+superkingdom
superkingdom	1	749.27	0	0	100.0	0.0	1	46.5	46.5	2				
phylum	2	1920.91	19.56	0	50.6	49.3	2	35.1	39.8	8				
class	3	1049.16	158.44	0	49.0	47.9	4	27.6	33.1	12	4741.87	344.53	93.2	phylum+class+order
order	4	1771.8	166.53	0	23.1	38.4	10	25.3	34.2	22				
family	5	961.81	367.01	0	13.5	32.0	17	13.2	29.5	29				
genus	6	2.33	1871.14	0	7.5	22.8	32	2.0	10.2	37	1421.6	8584.84	14.2	family+genus+species
species	7	457.46	6346.69	0	7.4	23.2	31	0.2	1.0	47				
avg./sum	3.7	6912.74	8929.37	0	35.9	30.5	13.9	21.4	27.7	22.4			43.6	all but unassigned
avg./sum	3.3	8110.93	8929.37	0	43.9	26.7	12.3	31.2	24.3	19.8			47.6	all with unassigned

MEGAN5 binning for FAMeS SimMC scenario (Nature Methods 2011)
using nucleotide-level alignment



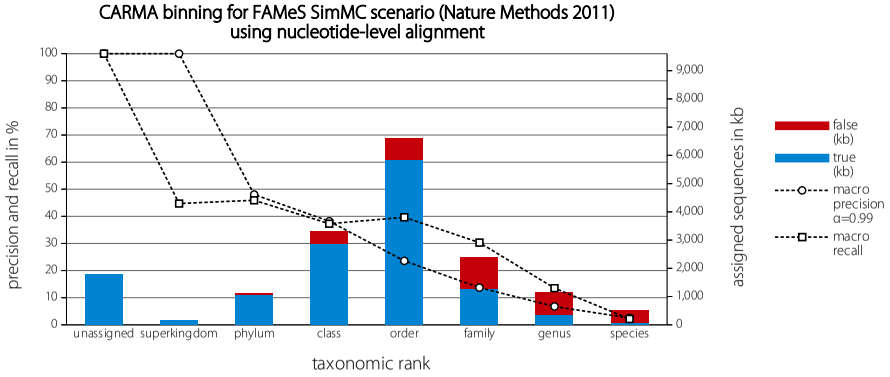
Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(e) MEGAN5 (amino acid)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	232.06	0	0	100.0	0.0	1	100.0	0.0	1	3526.24	0	100.0	root+superkingdom
superkingdom	1	1647.09	0	0	100.0	0.0	1	49.3	49.3	2				
phylum	2	3471.29	14	0	51.0	49.0	2	38.7	43.2	8	8695.34	423.54	95.4	phylum+class+order
class	3	2273.77	182.52	0	49.1	47.6	4	32.8	35.3	12				
order	4	2950.28	227.02	0	31.1	37.3	11	25.4	35.4	22				
family	5	1710.01	284.1	0	15.6	32.3	18	7.6	20.1	29	2114.54	3927.71	35.0	family+genus+species
genus	6	18.97	1024.71	0	12.3	29.1	29	3.8	14.5	37				
species	7	385.56	2618.9	0	11.1	30.2	25	0.2	0.8	47				
avg/sum	3.3	12456.97	4351.25	0	38.6	32.2	12.9	22.6	28.4	22.4			74.1	all but unassigned
avg/sum	3.2	12689.03	4351.25	0	46.3	28.2	11.4	32.2	24.8	19.8			74.5	all with unassigned



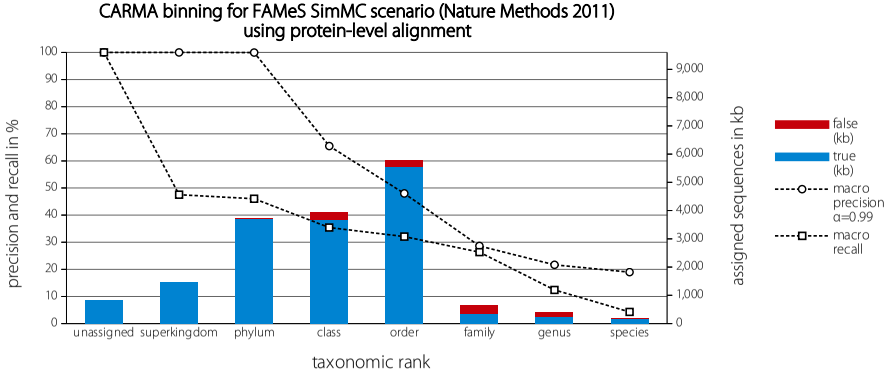
Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(f) CARMA (nucleotide)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	1788.99	0	0	100.0	0.0	1	100.0	0.0	1	2125.95	0	100.0	root+superkingdom
superkingdom	1	168.48	0	0	100.0	0.0	1	44.7	44.7	2				
phylum	2	1071.86	49.59	0	48.1	40.4	3	45.9	36.4	8	9749.83	1265.46	88.5	phylum+class+order
class	3	2853.36	446.35	0	38.2	46.0	5	37.3	34.0	12				
order	4	5824.61	769.52	0	23.6	35.1	19	39.7	37.0	22				
family	5	1266.67	1107.7	0	13.7	28.0	50	30.3	41.8	29	1719.92	2347.61	42.3	family+genus+species
genus	6	364.14	796.45	0	6.8	23.0	93	13.5	32.1	37				
species	7	89.11	443.46	0	2.5	15.1	135	2.2	14.4	47				
avg/sum	3.9	11638.23	3613.07	0	33.3	26.8	43.7	30.5	34.4	22.4			76.3	all but unassigned
avg/sum	3.5	13427.22	3613.07	0	41.6	23.4	38.4	39.2	30.1	19.8			78.8	all with unassigned



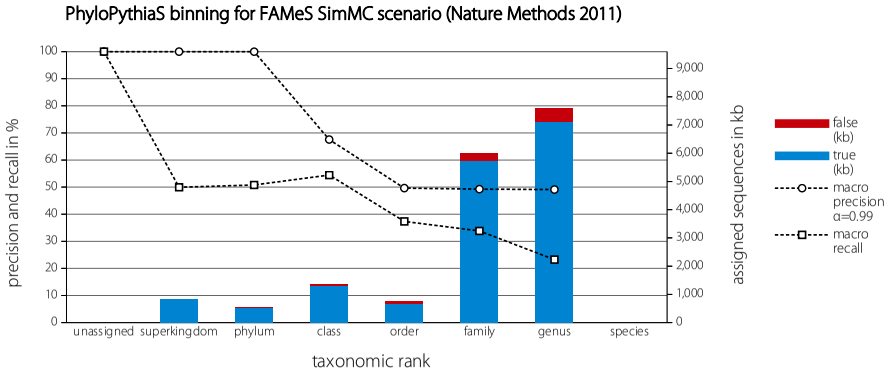
Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(g) CARMA (amino acid)**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	843.44	0	0	100.0	0.0	1	100.0	0.0	1	3809.88	0	100.0	root+superkingdom
superkingdom	1	1483.22	0	0	100.0	0.0	1	47.5	47.5	2				
phylum	2	3695.28	28.78	0	100.0	0.0	1	46.0	38.1	8	12906.86	538.72	96.0	phylum+class+order
class	3	3671.01	275.38	0	65.5	44.9	3	35.4	32.6	12				
order	4	5540.57	234.56	0	48.0	43.3	10	32.1	33.8	22				
family	5	345.44	315.09	0	28.6	41.3	32	26.3	39.1	29	753.71	514.34	59.4	family+genus+species
genus	6	237.82	174.47	0	21.7	39.9	36	12.4	29.7	37				
species	7	170.45	24.78	0	19.0	38.2	25	4.3	20.2	47				
avg/sum	3.1	15143.79	1053.06	0	54.7	29.7	15.4	29.2	34.4	22.4			93.5	all but unassigned
avg/sum	2.9	15987.23	1053.06	0	60.3	26.0	13.6	38.0	30.1	19.8			93.8	all with unassigned



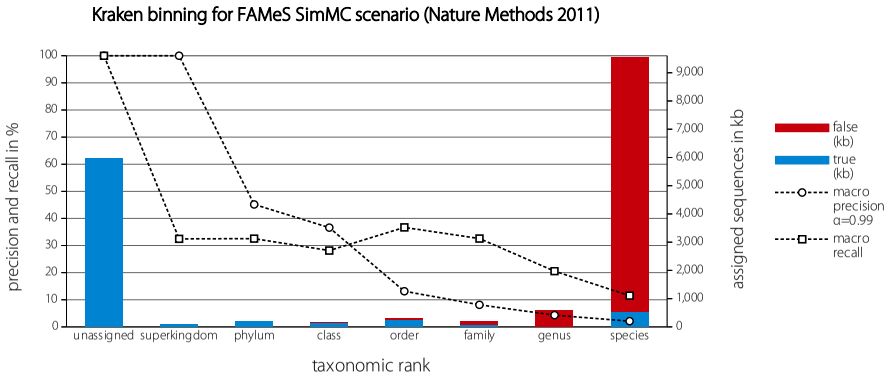
Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(h) PhyloPythiaS**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	0	0	0	100.0	0.0	1	100.0	0.0	1	1664.72	0	100.0	root+superkingdom
superkingdom	1	832.36	0	0	100.0	0.0	1	49.9	49.9	2				
phylum	2	517.19	25.6	0	100.0	0.0	1	50.8	41.1	8	2474.3	154.19	94.1	phylum+class+order
class	3	1297.42	52.21	0	67.6	45.5	3	54.5	40.0	12				
order	4	659.69	76.38	0	49.6	44.9	6	37.4	34.2	22				
family	5	5715.02	272.58	0	49.3	48.2	6	33.8	40.7	29	12831.92	747.52	94.5	family+genus+species
genus	6	7116.9	474.94	0	49.1	46.0	6	23.3	37.4	37				
species	7													
avg/sum	5.0	16138.58	901.71	0	69.2	30.8	3.8	41.6	40.6	18.3			94.7	all but unassigned
avg/sum	5.0	16138.58	901.71	0	73.6	26.4	3.4	49.9	34.8	15.9			94.7	all with unassigned



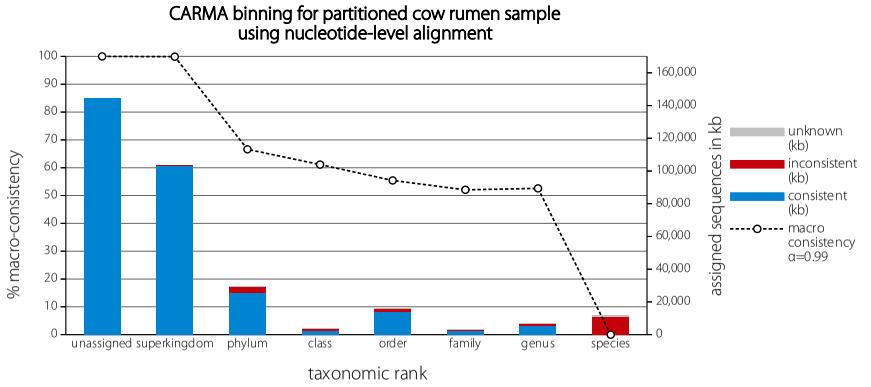
Supplementary Figure S21 - Binning for FAMeS SimMC scenario (Nature Methods 2011) **(i) Kraken**

rank	depth	true (kb)	false (kb)	unknown (kb)	macro precision α=0.99	stdev	pred. bins	macro recall	stdev	real bins	sum true (kb)	sum false (kb)	overall prec.	description
unassigned	0	5974.51	0	0	100.0	0.0	1	100.0	0.0	1	6130.53	0	100.0	root+superkingdom
superkingdom	1	78.01	0	0	100.0	0.0	1	32.5	32.5	2				
phylum	2	208.37	0	0	45.2	41.4	3	32.6	36.0	8				
class	3	144	17.5	0	36.6	44.9	5	28.2	33.5	12	591.64	79.44	88.2	phylum+class+order
order	4	239.27	61.94	0	13.2	28.5	21	36.7	38.6	22				
family	5	78.26	105.58	0	8.2	25.2	43	32.6	42.5	29				
genus	6	1.01	578.56	0	4.3	17.1	83	20.6	37.9	37	606.97	9709.72	5.9	family+genus+species
species	7	527.7	9025.58	0	2.1	13.1	123	11.5	30.9	47				
avg/sum	4.4	1276.62	9789.16	0	29.9	24.3	39.9	27.8	36.0	22.4			11.5	all but unassigned
avg/sum	0.9	7251.13	9789.16	0	38.7	21.3	35.0	36.8	31.5	19.8			42.6	all with unassigned



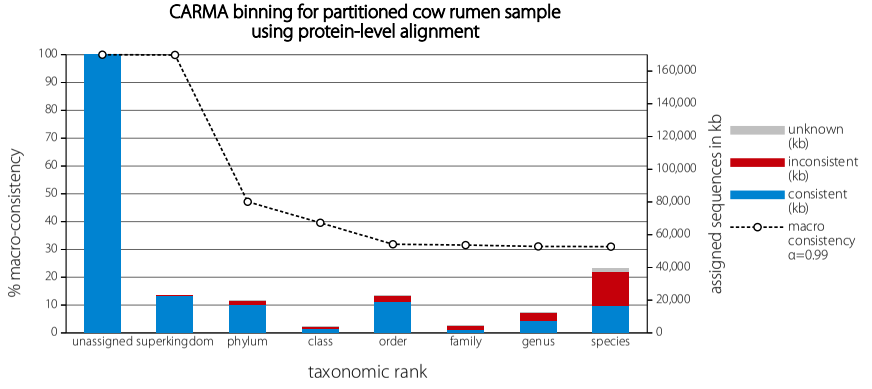
Supplementary Figure S22 - Binning for partitioned cow rumen sample **(a) CARMA (nucleotide)**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	144478	0	0	100.0	0.0	1	100.0	0.0	1	350414	154	100.0	root+superkingdom
superkingdom	1	102968	154	0	99.9	0.0	1	42.6	13.9	2				
phylum	2	25730	3872	22	66.6	20.9	13	13.0	5.8	30	41974	7032	85.7	phylum+class+order
class	3	2256	1350	54	61.1	18.7	28	11.1	4.7	52				
order	4	13988	1810	42	55.4	17.1	62	9.7	4.5	99				
family	5	2400	964	104	52.1	23.2	167	9.0	4.8	198	7952	12724	38.5	family+genus+species
genus	6	5552	1090	132	52.6	36.6	572	9.2	5.1	446				
species	7	0	10670	890	0.0	0.0	1254	0.0	0.0	926				
avg/sum	1.8	152894	19910	1244	55.4	16.6	299.6	13.5	5.5	250.4			88.5	all but unassigned
avg/sum	1.0	297372	19910	1244	61.0	14.5	262.3	24.3	4.8	219.3			93.7	all with unassigned



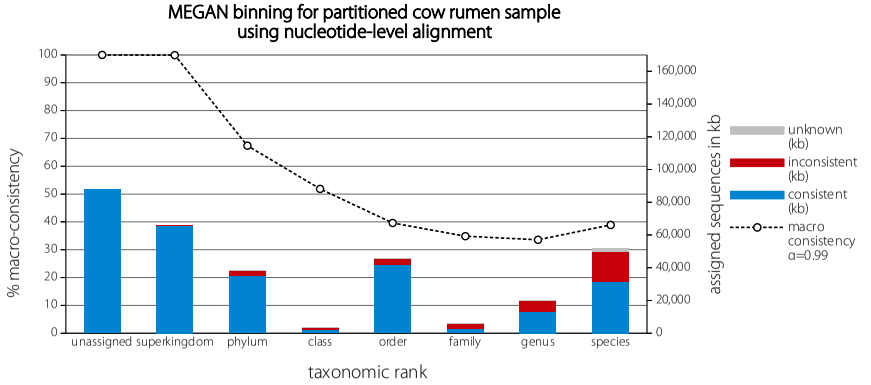
Supplementary Figure S22 - Binning for partitioned cow rumen sample **(b) CARMA (amino acid)**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	192034	0	0	100.0	0.0	1	100.0	0.0	1	238074	66	100.0	root+superkingdom
superkingdom	1	23020	66	0	99.9	0.0	1	33.2	7.1	2				
phylum	2	17210	2812	12	47.1	27.8	15	15.4	6.0	30	38912	7660	83.6	phylum+class+order
class	3	2620	1202	6	39.6	21.7	31	13.9	4.5	52				
order	4	19082	3646	14	31.9	17.3	71	12.3	4.3	104				
family	5	2190	2166	82	31.6	14.8	188	11.0	4.1	221	25938	28344	47.8	family+genus+species
genus	6	7412	4932	216	31.1	19.3	611	11.0	4.7	578				
species	7	16336	21246	2222	31.0	29.0	1956	10.4	5.0	1330				
avg/sum	3.0	87870	36070	2552	44.6	18.6	410.4	15.3	5.1	331.0			70.9	all but unassigned
avg/sum	0.9	279904	36070	2552	51.5	16.2	359.3	25.9	4.5	289.8			88.6	all with unassigned



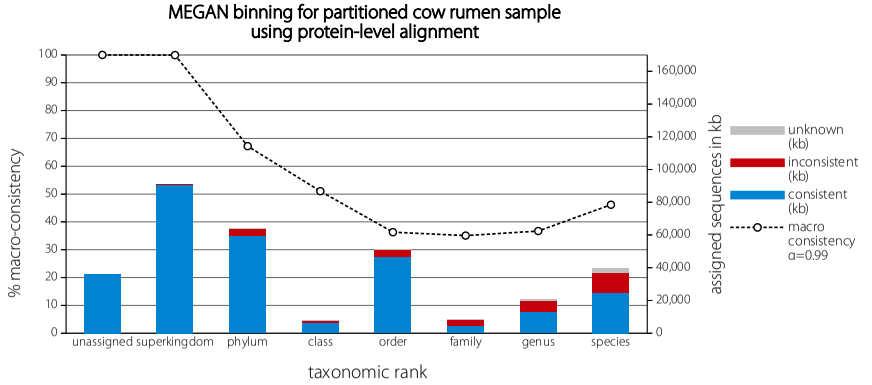
Supplementary Figure S22 - Binning for partitioned cow rumen sample **(c) MEGAN4 (nucleotide)**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	87760	0	0	100.0	0.0	1	100.0	0.0	1	218880	116	99.9	root+superkingdom
superkingdom	1	65560	116	0	99.9	0.0	1	43.8	26.2	3				
phylum	2	35352	2802	34	67.4	24.6	12	24.7	16.9	27	79676	7200	91.7	phylum+class+order
class	3	2242	1090	42	51.9	27.7	25	19.7	13.9	48				
order	4	42082	3308	66	39.6	26.3	51	15.5	11.9	88				
family	5	2802	3220	178	34.9	21.1	132	13.6	9.2	168	46888	28304	62.4	family+genus+species
genus	6	12764	6726	436	33.6	20.4	264	12.6	8.0	295				
species	7	31322	18358	2266	38.9	21.9	564	11.7	7.6	535				
avg/sum	2.7	192124	35620	3022	52.3	20.3	149.9	20.2	13.4	166.3			84.4	all but unassigned
avg/sum	1.8	279884	35620	3022	58.3	17.8	131.3	30.2	11.7	145.6			88.7	all with unassigned



Supplementary Figure S22 - Binning for partitioned cow rumen sample **(d) MEGAN5 (amino acid)**

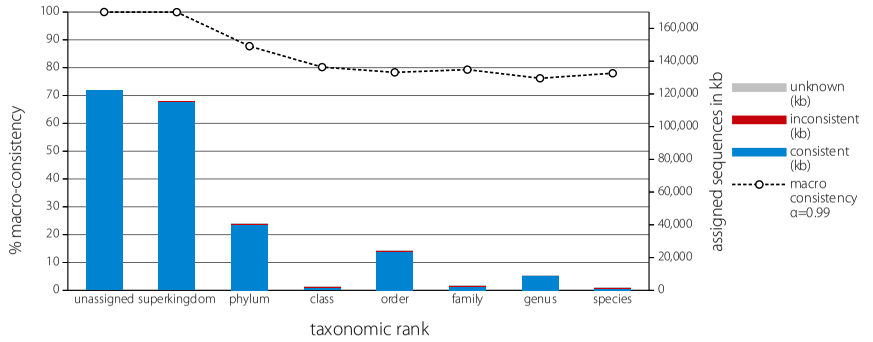
rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	36062	0	0	100.0	0.0	1	100.0	0.0	1	217142	220	99.9	root+superkingdom
superkingdom	1	90540	220	0	99.9	0.0	1	72.9	15.9	2				
phylum	2	59686	4074	26	67.2	25.6	12	25.3	17.8	26	112124	10496	91.4	phylum+class+order
class	3	6054	1614	24	51.0	27.9	25	20.1	13.4	44				
order	4	46384	4808	118	36.3	27.1	52	15.9	11.4	79				
family	5	4208	4130	358	35.1	21.7	119	13.1	8.7	140	41984	23208	64.4	family+genus+species
genus	6	13258	6756	778	36.7	20.3	203	12.3	7.9	218				
species	7	24518	12322	2588	46.2	20.5	347	11.5	8.6	343				
avg/sum	2.5	244648	33924	3892	53.2	20.4	108.4	24.4	11.9	121.7			87.8	all but unassigned
avg/sum	2.2	280710	33924	3892	59.1	17.9	95.0	33.9	10.5	106.6			89.2	all with unassigned



Supplementary Figure S22 - Binning for partitioned cow rumen sample **(e) taxator-tk (nucleotide)**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	122146	0	0	100.0	0.0	1	100.0	0.0	1	353002	62	100.0	root+superkingdom
superkingdom	1	115428	62	0	100.0	0.0	1	56.8	7.3	2				
phylum	2	39828	1152	4	87.7	16.9	7	16.4	12.3	22	65086	2212	96.7	phylum+class+order
class	3	1676	334	28	80.2	17.0	14	13.7	11.3	34				
order	4	23582	726	28	78.3	20.2	16	11.7	10.8	56				
family	5	2524	198	100	79.3	19.8	50	10.3	8.8	84	12810	440	96.7	family+genus+species
genus	6	8938	198	94	76.2	35.9	110	9.8	7.8	94				
species	7	1348	44	88	78.0	37.4	123	8.6	6.7	103				
avg/sum	1.9	193324	2714	342	82.8	21.0	45.9	18.2	9.3	56.4			98.6	all but unassigned
avg/sum	1.2	315470	2714	342	85.0	18.4	40.3	28.4	8.1	49.5			99.1	all with unassigned

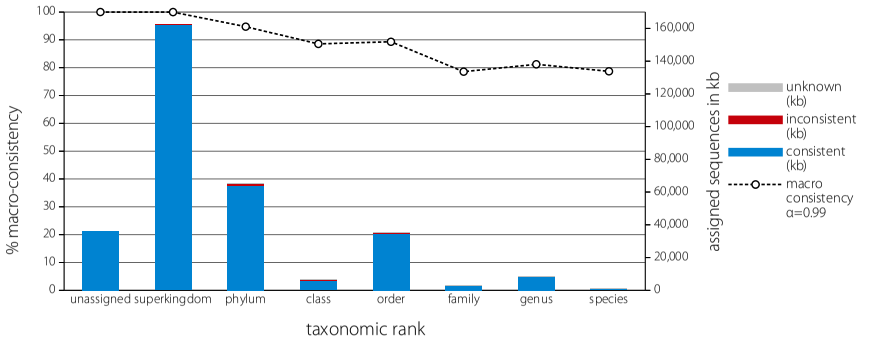
taxator-tk binning for partitioned cow rumen sample
using nucleotide-level alignment



Supplementary Figure S22 - Binning for partitioned cow rumen sample **(f) taxator-tk (amino acid)**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	36398	0	0	100.0	0.0	1	100.0	0.0	1	361102	144	100.0	root+superkingdom
superkingdom	1	162352	144	0	100.0	0.0	1	74.1	14.5	2				
phylum	2	63942	1604	2	94.7	6.2	5	20.9	15.7	17	104336	2614	97.6	phylum+class+order
class	3	5920	410	24	88.6	14.1	10	16.3	13.1	26				
order	4	34474	600	30	89.3	15.4	9	14.0	12.4	37				
family	5	2586	92	72	78.6	22.0	32	11.8	9.4	51	12066	328	97.4	family+genus+species
genus	6	8344	228	82	81.2	26.8	27	10.5	8.1	55				
species	7	1136	8	78	78.7	37.7	59	9.8	7.8	49				
avg/sum	1.8	278754	3086	288	87.3	17.5	20.4	22.5	11.6	33.9			98.9	all but unassigned
avg/sum	1.6	315152	3086	288	88.9	15.3	18.0	32.2	10.1	29.8			99.0	all with unassigned

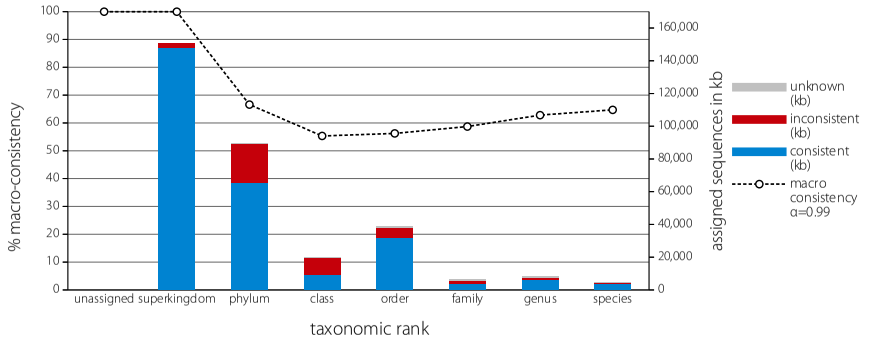
taxator-tk binning for partitioned cow rumen sample
using protein-level alignment



Supplementary Figure S22 - Binning for partitioned cow rumen sample **(g) PhyloPythiaS**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	0	0	0	100.0	0.0	1	100.0	0.0	1	296276	2810	99.1	root+superkingdom
superkingdom	1	148138	2810	0	100.0	0.0	1	81.6	17.5	2				
phylum	2	65136	24220	4	66.6	15.1	4	31.0	14.7	7	106338	40640	72.3	phylum+class+order
class	3	9468	9930	568	55.3	22.0	10	21.0	10.7	13				
order	4	31734	6490	828	56.2	21.0	19	12.8	4.9	25				
family	5	3990	1438	708	58.7	18.9	30	10.4	4.1	39	13842	3054	81.9	family+genus+species
genus	6	6144	1078	1072	62.8	18.5	33	9.2	4.2	45				
species	7	3708	538	524	64.7	29.0	64	8.0	4.6	67				
avg/sum	2.0	268318	46504	3704	66.3	17.8	23.0	24.9	8.7	28.3			85.2	all but unassigned
avg/sum	2.0	268318	46504	3704	70.6	15.6	20.3	34.3	7.6	24.9			85.2	all with unassigned

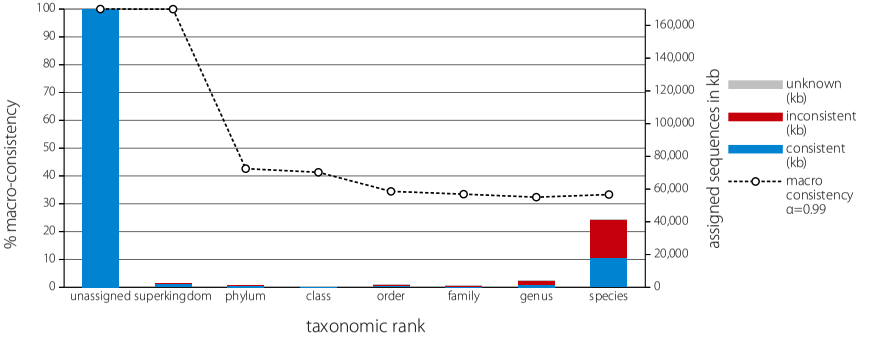
PhyloPythiaS binning for partitioned cow rumen sample



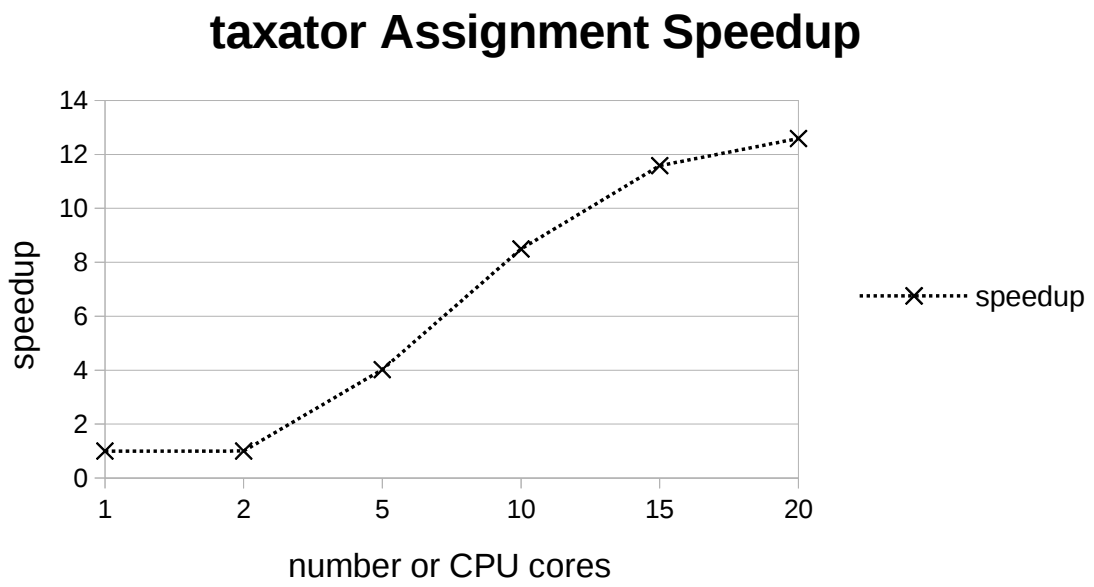
Supplementary Figure S22 - Binning for partitioned cow rumen sample **(h) Kraken**

rank	depth	consistent (kb)	inconsistent (kb)	unknown (kb)	macro consistency α=0.99	stdev	pred. bins	macro recall	stdev	cons. bins	sum true (kb)	sum false (kb)	overall consist.	description
unassigned	0	265616	0	0	100.0	0.0	1	100.0	0.0	1	270140	4	100.0	root+superkingdom
superkingdom	1	2262	4	0	99.9	0.0	1	21.0	1.2	2				
phylum	2	728	504	0	42.6	25.6	18	10.8	5.8	30	2034	1628	55.5	phylum+class+order
class	3	190	434	2	41.3	22.9	33	10.5	5.3	52				
order	4	1116	690	8	34.5	18.4	77	9.9	5.0	110				
family	5	254	684	8	33.5	15.3	195	9.3	4.3	233	19274	27218	41.5	family+genus+species
genus	6	1378	2736	34	32.4	19.5	661	9.7	4.6	640				
species	7	17642	23798	438	33.3	28.2	1953	9.9	5.2	1461				
avg/sum	3.9	23570	28850	490	45.4	18.6	419.7	11.6	4.5	361.1			45.0	all but unassigned
avg/sum	0.2	289186	28850	490	52.2	16.2	367.4	22.6	3.9	316.1			90.9	all with unassigned

Kraken binning for partitioned cow rumen sample



Supplementary Figure S23: Parallel speedup of program *taxator*



Execution time analysis with taxator for parallelized processing with multiple CPU cores. Taxonomic placement of sequence segments with taxator on input alignments for sequences of length 1000 bp (*syn1000* data-set aligned against *mRefSeq47* with *LAST*). The speedup was calculated using wall clock time for a parallelized run relative to serial execution with one CPU thread. With multiple threads, there is always one producer thread (consumer-producer model). Thus for more than two threads, multiple consumers work on the input data in parallel. An approximate linear scale-up was observed up to 15 threads and saturation effects appear when using 20 CPU cores on our system.

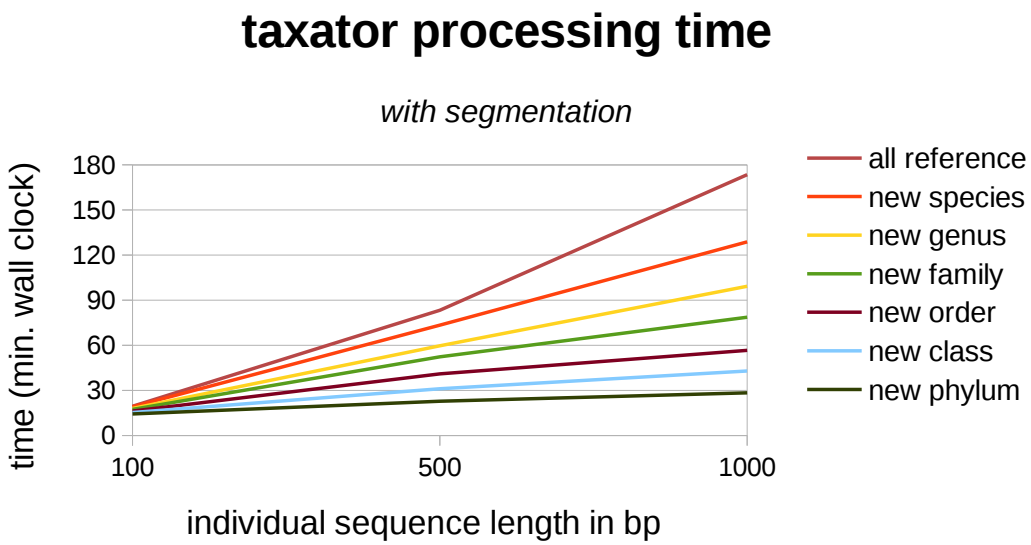
$speedup = \frac{T_1}{T_p}$, *with*

T_1*: serial execution time*

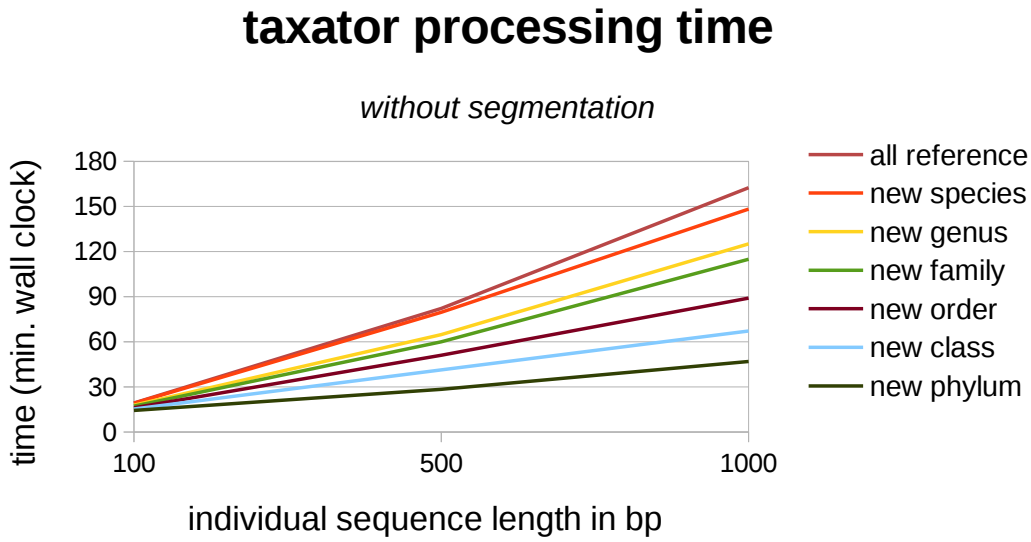
T_p*: execution time using p threads and CPU cores*

Supplementary Figure S24: Effect of input sequence length and segmentation on *taxator-tk* processing time.

(a)



(b)



We processed approximately the same number of sequences of length 100, 500 and 1000 bp with *taxator-tk* (*syn100*,*syn500*,*syn1000*), once with the segmentation procedure being enabled (a) and once with segmentation disabled (b). The run-time increases for both cases are approximately linear with the input length, where the slope depends on the completeness of the reference sequence data. With all reference data available, the run-time increases more than linear, as there is no segmentation of queries during computations. For all other cases, segmentation substantially decreases the execution time.

Supplementary Figure S25: Example GFF3 output of taxator

```
##gff-version 3
contig_0 taxator-tk sequence_feature 102   121  1     .      .     seqlen=1012;tax=1224:19;ival=0.5
contig_0 taxator-tk sequence_feature 155   194  0.91  .      .     seqlen=1012;tax=2:32-1;ival=0.8
contig_0 taxator-tk sequence_feature 201   220  1     .      .     seqlen=1012;tax=40324:20-2;ival=0
contig_0 taxator-tk sequence_feature 225   243  1     .      .     seqlen=1012;tax=316277:19-1;ival=1
contig_0 taxator-tk sequence_feature 246   301  1     .      .     seqlen=1012;tax=731:38-1224;ival=0.72
contig_0 taxator-tk sequence_feature 326   471  1     .      .     seqlen=1012;tax=338:87-1224;ival=0.98
contig_0 taxator-tk sequence_feature 486   554  0.60  .      .     seqlen=1012;tax=1224:59-2;ival=0.67
contig_0 taxator-tk sequence_feature 555   616  0.63  .      .     seqlen=1012;tax=32008:43-1224;ival=0.86
contig_0 taxator-tk sequence_feature 633   651  1     .      .     seqlen=1012;tax=876:19-1;ival=1
contig_0 taxator-tk sequence_feature 670   745  0.89  .      .     seqlen=1012;tax=31998:60-1;ival=0.89
contig_0 taxator-tk sequence_feature 786   809  0.89  .      .     seqlen=1012;tax=256618:23-2;ival=0.2
contig_0 taxator-tk sequence_feature 886   932  1     .      .     seqlen=1012;tax=644:33-2;ival=0.67
contig_0 taxator-tk sequence_feature 958   980  1     .      .     seqlen=1012;tax=347:22-1;ival=1
```

Query identifier	Generator	Type	Begin	End	Score	Strand	Phase	Query length Taxonomic range and support Interpolation value

Query segment assignments calculated by the program *taxator* (version 1.1.1) are generated in standard GFF3 format. Each tab-separated field holds the information which is named in the bottom description. The score measures the assignment quality and is under ongoing improvement. Strand and phase contain a dot as placeholder as they are invalid GFF3 fields for this output. The last column holds data in a key-value scheme and includes the query sequence length, a taxonomic prediction range of the form low:support-high where low/specific (node X in Fig. 2a) and high/general (node R in Fig. 2a) are NCBI taxon IDs. The included interpolation value ranging from zero (low) and one (high) can be used to determine an approximate position in the given taxonomic range. As it might become necessary for post-processing applications such as whole sequence binning, more information can be added in the last column while preserving backward compatibility.

Appendix B

Supplementary Material for "A Probabilistic Model to Recover Genomes in Shotgun Metagenomics"

Supplementary Material

A probabilistic model to recover individual genomes from metagenomes

Johannes Dröge[1], Alexander Schönhuth[2], and Alice C. McHardy[3]

[1] *Helmholtz Centre for Infection Research, Braunschweig, science@fungs.de*
[2] *Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, a.schoenhuth@cwi.nl*
[3] *Helmholtz Centre for Infection Research, Braunschweig, alice.mchardy@helmholtz-hzi.de*

2017-10-25

Supporting Data

The simulated contigs, features files and scripts to reproduce the results are deposited under:

DOI:10.5281/zenodo.201076

The CAMI reference sequence data and corresponding taxonomy used in this article is available as refpack "microbial-full_20150430" for the software taxator-tk.

Supplementary Methods

Poisson approximation for absolute abundance

When sequencing reads have been mapped to the contigs, we can quantify the number of reads that covers each position of each contig. This is the vector x with $len(\boldsymbol{x}) = L$. We model the positional read coverage using a Poisson event model and assume that the positions are independent according to the Lander-Waterman statistics so that the joint likelihood is a product of positional likelihoods. Additionally, we scale the likelihood to a single event by taking the geometric mean. After simplification, the formula almost looks like the the Poisson over the mean contig coverage.

$$\mathcal{L}(\theta \mid \boldsymbol{x}) = \sqrt[L]{\prod_{i=1}^{L} \frac{\theta^{x_i}}{x_i!} e^{-\theta}} = \left(\frac{\prod_{i=1}^{L} \theta^{x_i}}{\prod_{i=1}^{L} x_i!} e^{-\theta L} \right)^{\frac{1}{L}} = \frac{\bar{\theta}}{\sqrt[L]{\prod_{i=1}^{L} x_i!}} e^{-\theta} \tag{1}$$

The data term in the denominator is a constant factor which is not dependent on θ. It is the geometric mean over the $x_i!$ values which we approximate using the arithmetic mean $\bar{x}$ of the positional contig coverage values.

$$\sqrt[L]{\prod_{i=1}^{L} x_i!} \approx \left(\frac{1}{L} \sum_{i=1}^{L} x_i \right)! = \bar{x}! \tag{2}$$

The approximation is good if the variance of the x_i is low. We use the approximation to avoid to handle other values than the mean which is usually computed. Since the term is a data constant, it is

irrelevant for model comparison where only θ differs among the genomes. The approximated likelihood using mean values is the standard Poisson formula.

$$\mathcal{L}'(\theta \mid \boldsymbol{x}) = \frac{\theta^{\bar{x}}}{\bar{x}!} e^{-\theta} \tag{3}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell'(\theta \mid \boldsymbol{x}) = -\log \bar{x}! + \bar{x} \log \theta - \theta \tag{4}$$

MLE for Poisson

The multi-sample log-likelihood is the weighted sum over the sample log-likelihoods using mean vector $\boldsymbol{a_i}$ with length $len(\boldsymbol{a_i}) = M$. This corresponds to the geometric mean in the exponential likelihood formula.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{a_i}) = \frac{1}{M} \sum_{j=1}^{M} -\log a_{i,j}! + a_{i,j} \cdot \log \theta_j - \theta_j \tag{5}$$

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ on the training data a. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{a_i}) = \sum_{i=1}^{N} w_i \cdot \frac{1}{M} \sum_{j=1}^{M} -\log a_{i,j}! + a_{i,j} \log \theta_j - \theta_j \tag{6}$$

The partial derivative of f with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial f}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{a_{i,j}}{\theta_j} - 1 \right) \tag{7}$$

We find the zeros of f to determine the MLE $\hat{\theta}_j$.

$$\sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{a_{i,j}}{\theta_j} - 1 \right) = 0 \Leftrightarrow \sum_{i=1}^{N} \frac{w_i \, a_{i,j}}{\theta_j} = \sum_{i=1}^{N} w_i \Leftrightarrow \theta_j = \frac{\sum_{i=1}^{N} w_i \, a_{i,j}}{\sum_{i=1}^{N} w_i} \tag{8}$$

We see that the estimates for θ_j maximize the joint log-likelihood because the second partial derivative with respect to θ_j is always negative.

$$\frac{\partial^2 f}{\partial \theta_j^2} = -\sum_{i=1}^{N} \frac{w_i \, a_{i,j}}{M \, \theta_j^2} \tag{9}$$

Binomial approximation for relative abundance

Similarly to the Poisson approximation for absolute abundance, we derive the Binomial approximation via a product of positional Binomials. Vector $\boldsymbol{x}$ with length $len(\boldsymbol{x}) = L$ holds the positional read coverage of a contig with length L for one sample and vector s with same length holds the sum of positional read counts for the position i of the contig across all samples. There must be more than one sample to apply this model. We write the likelihood normalized to a single event as

$$\begin{aligned}
\mathcal{L}(\theta \mid \boldsymbol{x}) &= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i} \theta^{x_i} (1-\theta)^{(s_i - x_i)}} \\
&= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} \cdot \sqrt[L]{\prod_{i=1}^{L} \theta^{x_i}} \cdot \sqrt[L]{\prod_{i=1}^{L} (1-\theta)^{(s_i - x_i)}} \\
&= \sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} \cdot \theta^{\bar{x}} \cdot (1-\theta)^{(\bar{s} - \bar{x})}
\end{aligned} \tag{10}$$

The geometric mean of positional binomial coefficients (first term) is again a constant factor which is not dependent on θ. We approximate this term using the arithmetic mean.

$$\begin{aligned}
\sqrt[L]{\prod_{i=1}^{L} \binom{s_i}{x_i}} &= \frac{\sqrt[L]{\prod_{i=1}^{L} s_i!}}{\sqrt[L]{\prod_{i=1}^{L} x_i!} \cdot \sqrt[L]{\prod_{i=1}^{L} (s_i - x_i)!}} \\
&\approx \frac{\frac{1}{L}\sum_{i=1}^{L} s_i!}{\frac{1}{L}\sum_{i=1}^{L} x_i! \cdot \frac{1}{L}\sum_{i=1}^{L} (s_i - x_i)!} \\
&\approx \frac{\frac{1}{L}\sum_{i=1}^{L} s_i!}{\frac{1}{L}\sum_{i=1}^{L} x_i! \cdot \left(\frac{1}{L}\sum_{i=1}^{L} s_i - \frac{1}{L}\sum_{i=1}^{L} x_i\right)!} = \binom{\bar{s}}{\bar{x}}
\end{aligned} \tag{11}$$

The approximation is good if the differences in the coefficients are small. We use the approximation to avoid to handle other values than the mean which is usually computed. Since the term is a data constant, it is irrelevant for model comparison where only θ differs among the genomes. The approximated likelihood using mean values is the standard Binomial formula.

$$\mathcal{L}'(\theta \mid \boldsymbol{x}) = \binom{\bar{s}}{\bar{x}} \theta^{\bar{x}} (1-\theta)^{(\bar{s} - \bar{x})} \tag{12}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell'(\theta \mid \boldsymbol{x}) = \log \binom{\bar{s}}{\bar{x}} + \bar{x} \log \theta + (\bar{s} - \bar{x}) \log(1-\theta) \tag{13}$$

MLE for Binomial

The multi-sample log-likelihood is the weighted sum over the sample log-likelihoods using mean vector $\boldsymbol{r_i}$ with length $len(\boldsymbol{r_i}) = M$. This corresponds to the geometric mean in the exponential likelihood formula.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{r_i}) = \frac{1}{M} \sum_{j=1}^{M} \log \binom{R_i}{r_{i,j}} + r_{i,j} \log \theta_j + (R_i - r_{i,j}) \log(1 - \theta_j) \tag{14}$$

R_i is the sum of the abundance vector $\boldsymbol{r_i}$.

$$R_i = \sum_{j=1}^{M} r_{i,j} \tag{15}$$

Because both R_i and $r_{i,j}$ can be real numbers, we need to generalize the binomial coefficient to positive real numbers via the gamma function Γ.

$$\log \binom{n}{k} = \log \Gamma(n+1) - \log \Gamma(k+1) - \log \Gamma(n-k+1) \tag{16}$$

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ of the training data r. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$\begin{aligned} f(\boldsymbol{\theta}) &= \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{r_i}) \\ &= \sum_{i=1}^{N} w_i \cdot \frac{1}{M} \sum_{j=1}^{M} \log \binom{R_i}{r_{i,j}} + r_{i,j} \log \theta_j + (R_i - r_{i,j}) \log(1 - \theta_j) \end{aligned} \tag{17}$$

The partial derivative of f with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial f}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{r_{i,j}}{\theta_j} - \frac{R_i - r_{i,j}}{1 - \theta_j} \right) \tag{18}$$

We find the zeros of f to determine the MLE $\hat{\theta}_j$.

$$\begin{aligned} &\sum_{i=1}^{N} \frac{w_i}{M} \left(\frac{r_{i,j}}{\theta_j} - \frac{R_i - r_{i,j}}{1 - \theta_j} \right) = 0 \\ \Leftrightarrow &(1 - \theta_j) \sum_{i=1}^{N} w_i r_{i,j} = \theta_j \left(\sum_{i=1}^{N} w_i R_i - \sum_{i=1}^{N} w_i r_{i,j} \right) \\ \Leftrightarrow &\frac{1}{\theta_j} \sum_{i=1}^{N} w_i r_{i,j} = \sum_{i=1}^{N} w_i R_i \\ \Leftrightarrow &\theta_j = \frac{\sum_{i=1}^{N} w_i r_{i,j}}{\sum_{i=1}^{N} w_i R_i} \end{aligned} \tag{19}$$

We see that the estimates for θ_j maximize the joint log-likelihood because the second partial derivative with respect to θ_j is negative for our estimates $\hat{\theta}_j$ for all $j \in \{1...M\}$.

$$\frac{\partial^2 f}{\partial \theta_j^2} = -\frac{R_i\theta_j^2 - 2r_{i,j}\theta_j + r_{i,j}}{(\theta_j - 1)^2 \theta_j^2} \tag{20}$$

$$-\frac{R_i\hat{\theta}_j^2 - 2r_{i,j}\hat{\theta}_j + r_{i,j}}{\left(\hat{\theta}_j - 1\right)^2 \hat{\theta}_j^2} < 0 \Leftrightarrow \sum_{i=1}^{N} w_i r_{i,j} < \sum_{i=1}^{N} w_i R_i \tag{21}$$

The last inequality is true by definition of R_i (assuming $r_{i,j} \neq R_i$ for simplicity).

Frequency model for nucleotide composition

The frequency model assumes independence of features so that the likelihood can be written as a product of likelihoods for all features. The feature vector $\boldsymbol{x}$ for a contig contains nucleotide features such as all the absolute counts for all possible 5-mers. The length $len(\boldsymbol{x})$ is M. The total sum of counts for the contig is S.

$$S = \sum_{i=1}^{M} x_i \tag{22}$$

The likelihood is normalized to a single event via the geometric mean.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sqrt[S]{\prod_{i=1}^{M} \theta_i^{x_i}} = \prod_{i=1}^{M} \theta_i^{\frac{x_i}{S}} = \prod_{i=1}^{M} \theta_i^{x_i'} \tag{23}$$

Therefore, we directly use the normalized features.

$$x_i' = \frac{x_i}{\sum_{j=1}^{M} x_j} \tag{24}$$

The log-likelihood is used in the MGLEX implementation for computational reasons. It is directly visible that the calculation is linear in the input.

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x'}) = \sum_{i=1}^{M} x_i' \log \theta_i \tag{25}$$

MLE for frequency model

We select $\boldsymbol{\theta}$ to maximize the joint log-likelihood $f(\boldsymbol{\theta})$ on the training data c. The joint likelihood is a weighted sum of the log-likelihood values of all N contigs. Each contig's weight w_i is the contig length.

$$f(\boldsymbol{\theta}) = \sum_{i=1}^{N} w_i \cdot \ell(\boldsymbol{\theta} \mid \boldsymbol{c_i}) = \sum_{i=1}^{N} w_i \cdot \sum_{j=1}^{M} c_{i,j} \log \theta_j \tag{26}$$

We consider the constraint that $sum(\boldsymbol{\theta}) = 1$ because these are relative frequencies in each genome.

$$\sum_{j=1}^{M} \theta_j = 1 \tag{27}$$

Using the Lagrange method, we set up a function to maximize the joint data log-likelihood $f(\boldsymbol{\theta})$ under the given constraint.

$$\Lambda(\boldsymbol{\theta}, \lambda) = f(\boldsymbol{\theta}) + \lambda \left(\left(\sum_{j=1}^{M} \theta_j \right) - 1 \right) \tag{28}$$

The partial derivative of Λ with respect to θ_j for all $j \in \{1...M\}$ is given by

$$\frac{\partial \Lambda}{\partial \theta_j} = \sum_{i=1}^{N} \frac{w_i \, c_{i,j}}{\theta_j} + \lambda \tag{29}$$

We find the zeros of Λ to determine the MLE $\hat{\theta}_j$.

$$\frac{\partial \Lambda}{\partial \theta_j} = 0 \Leftrightarrow \theta_j = \frac{\sum_{i=1}^{N} w_i \, c_{i,j}}{-\lambda} \tag{30}$$

Substituting θ_j in Suppl. Equation 27 gives

$$-\lambda = \sum_{i=1}^{N} w_i \sum_{j=1}^{M} c_{i,j} = \sum_{i=1}^{N} w_i \tag{31}$$

The last simplification works because we work with normalized features that sum to one. Finally, we substitute $-\lambda$ in (1) for the MLE.

$$\hat{\theta}_j = \frac{\sum_{i=1}^{N} w_i \, c_{i,j}}{\sum_{i=1}^{N} w_i} \tag{32}$$

Multi-layer frequency model for sequence similarity

We adapted the simple frequency model to weighted taxa by transforming the associated weights (i.e. alignments scores) into a set of sparse vectors x_l, one for each taxonomic rank. There are L such layers. The model likelihood is a product of observation probabilities, like in the standard simple model, but the layers are also connected by multiplication.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{l=1}^{L} \prod_{j=1}^{len(\boldsymbol{x_l})} \theta_{l,j}^{x_{l,j}} \tag{33}$$

The small difference to the simple model in the previous section is that there are no sequence length weights and that the feature vectors are not normalized. The multiplication of layers is a simplification

because we know that taxonomic ranks are not independent. However, the model proved to be simple and effective for our purposes.

MLE for multi-layer frequency model

Once the assumption of layer independence has been made, the problem simplifies to L independent simple frequency models with separate feature vectors and model parameters. The MLE derivation for each of these models is equivalent to the previous section. T_l is the number of features on level l.

$$\hat{\theta}_l = \frac{\sum_{i=1}^{N} t_{i,l}}{\sum_{j=1}^{T_l} \sum_{i=1}^{N} t_{i,l}} \tag{34}$$

Metagenome simulation

We chose genomes according to the CAMI2015 (www.cami-challenge.org) medium complexity toy dataset which contained 450 different strains. Because some of the strains were simulated and had no accessible genome data, we reduced the dataset to 400 genomes with corresponding accessions. These comprised both finished and draft genomes. We sampled the abundance distributions from a lognormal with expectation value one and variance one, which produced abundance value in an reasonable range and formed relative abundance by normalization (Supplementary Table 1, column S1). We derived three secondary samples (Supplementary Table 1, columns S2, S3, S4) by separately applying continuous (exponential) growth to a randomly chosen set of genomes which each constituted 100 genomes (25%) in the primary sample using the following formula.

$$abundance'(\text{genome}) = abundance(\text{gnome}) \cdot 2^{growth\_rate(\text{genome})} \tag{35}$$

We modeled the change of the community composition in reaction to variation of environmental parameters, for instance if the growth medium is altered with no space restrictions then community members will grow according to their genomic potential. In our simplified growth model we choose the growth rate uniformly at random between one and ten regardless of the actual genome. We generated three secondary abundance profiles using the described procedure. We then simulated HiSeq Illumina reads for each sample using the ART simulator with read length 150 bp, insert size 270 bp and insert size standard deviation 27 bp. This corresponds to a common experimental setting because the reads are likely to overlap in the read assembly step. We chose a large yield of 15 Gb per sample to also cover genomes with low sample abundance (see Supplementary Table 1).

Feature generation

All features are represented as separate text files, which can be compressed. Each line corresponds to a sequence but does not contain sequence identifiers. Therefore, it is required that the number and order of lines are identical in all features files.

Sequence weights

We used the following GNU awk v4.0.1 script to calculate the length of each FASTA entry which we saved as `contigs.seqlen`.

```
#!/usr/bin/awk -f
BEGIN { id="\000" } # > not allowed in FASTA header
/^>/ {
  if( id != "\000" ) {
    printf "%s\t%s\n", id, sum;
  }
  id=substr( $0, 2 );
  sum = 0;
}
! /^>/ { sum+=length($0) }
END { printf "%s\t%s\n", id, sum }
```

5-mer frequencies

We derived 5-mer frequencies for the gzip-compressed FASTA sequences using the program fasta2kmerS using the following GNU Bash syntax

```
zcat contigs.fna.gz |
fasta2kmersS -i <(cat) -f >(cat) -j 5 -k 5 -s 0 -h 0 -n 0 |
tr '\t' ' ' > contigs.kmc
```

Taxonomic annotation

We generated alignments using NCBI BLAST+/*blastn* v2.2.28+ in taxator-tk tabular format and filtered out all species level alignments using program *alignments-filter* from taxator-tk v1.3.3 which effectively removes the genomes of the same species from the reference sequences. Next we ran the program *taxator* with the LCA algorithm using only the best hits and processed the resulting GFF3 file. We used the alignment score as weight for each taxon and combined the annotations for each contig. Finally, we shortened the taxon paths using numbers and applied the described accumulation scheme to project alignment score onto higher-level taxa (see Table 1).

Average read coverage

We aligned each sample's simulated read data to the artificial contigs with Bowtie v2.2.7 and converted the resulting SAM files to sorted BAM

```
bowtie2-build contigs.fna contigs.bowtie2
bowtie2 -x contigs.bowtie2 -1 forward.fq.gz -2 reverse.fq.gz |
samtools view -@ 5 -b - < input.sam | samtools sort -@ 5 - out
```

and then calculated the average read coverage using BedTools v2.25 and GNU awk v4.0.1

```
genomeCoverageBed -ibam out.sorted.bam -g contigs.seqlen -d -split |
awk 'BEGIN{IFS=OFS=FS="\t"}
     {if($1 == last){ s+=$3; c+=1;}
     else{if(s){print last, s/c; s=$3}; c=1; last=$1}}
     END{print last, s/c}' > out.twocol.cov
```

Contigs which recruited no reads are omitted by BedTools, therefore zero values must be added afterwards by comparison to the sequence length file. Finally, we merged the coverage columns in Bash using

```
paste -d ' ' <(cut -f 2 < 1.twocol.cov) <(cut -f 2 < 2.twocol.cov) [...] > out.cov
```

Performance measures

In order to evaluate the quality of the predictions and to pick the optimal β parameter for the posterior estimation, MGLEX implements two measures: a mean squared error (MSE) and the mean pairwise coclustering (MPC) probability. Both require as input a label probability matrix which defines to which genome (column) each sequence (row) belongs, in terms of probabilities. In our simulation, the genome column corresponding to the source genome contained a one, all other columns a zero. A prediction probability matrix of the same form is required for comparison. In the case of ML predictions, this matrix also contains only ones and zeros and continuous values for the posterior estimation. Because sequences typically have different lengths, the user must provide a file with the sequence lengths (see AWK script for sequence weight file generation).

Mean squared error (MSE)

The mean squared error is the square root of the average squared difference between the label and the prediction matrix per contig (a value between zero and one). It is weighted by the length of the sequence.

$$\mathrm{MSE} = \sqrt{\frac{1}{4\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \sum_{j=1}^{M} (L_{i,j} - P_{i,j})^2} \tag{36}$$

Here, N is the number of sequences, M the number of genomes, w is a vector with the sequence lengths, L the label probability matrix and P the prediction probability matrix.

Mean pairwise coclustering (MPC)

The mean pairwise coclustering probability reports how likely a pair of sequences chosen from any genome among the real genomes, are found in the same predicted genome. The MPC averages over both, the pairs in the genomes and the genomes, regardless of their size. Since all sequences in our evaluations have the same length, we report the unweighted version of the MPC. The MPC is a probability between zero and one. It is easier to interpret than the MSE but requires more computation because it needs to consider all possible sequence pairs.

$$\mathrm{MPC} = \frac{1}{|C|} \sum_{i=1}^{|C|} \left(\frac{1}{|C_i|(|C_i|-1)} \sum_{\substack{s_1, s_2 \in C_i \\ s_1 \neq s_2}} p(s_1|C_i) p(s_2|C_i) \right) \tag{37}$$

Here, the i^{th} genome is a set C_i which contains sequences s_i and C is a set which contains all genomes C_i.

Genome bin posterior

We calculate the bin posterior of a contig over the genome bins by normalization of the different likelihood values for each of the considered bins, so that their values sum to one. We assume, that

the bin posterior is uniform over all G genome bins, so there is no additional weighting, for instance by genome size. $\mathcal{L}(\text{genome} \mid \text{contig})$ is a vector which holds the likelihood of a specific contig for every genome bin. Then, the posterior is given by

$$P(\text{genome} \mid \text{contig}) = \frac{\mathcal{L}(\text{genome} \mid \text{contig})}{\sum_{n=1}^{G} \mathcal{L}(\text{genome}_n \mid \text{contig})} \tag{38}$$

Relative likelihood bin comparison

We derived a percentage similarity quantity S for two genome bins A and B, based on mixture likelihoods.

$$\mathrm{S}(A,B) = \sqrt[Z]{\prod_{i=1}^{N} \left(\frac{2\, L_i(\theta_A)\, L_i(\theta_B)}{L_i^2(\theta_A) + L_i^2(\theta_B)} \right)^{\frac{L_i^2(\theta_A)+L_i^2(\theta_B)}{L_i(\theta_A)+L_i(\theta_b)}}} \tag{39}$$

with normalization constant

$$Z = \sum_{i=1}^{N} \frac{L_i^2(\theta_A) + L_i^2(\theta_B)}{L_i(\theta_A) + L_i(\theta_b)} \tag{40}$$

Interestingly, when we interpret this quantity as a probability, a connection to the Kullback-Leibler divergence D_{KL}, also called relative entropy, can be constructed. The Boltzmann formula (Suppl. Equation 41) establishes a general connection between entropy H and probability P.

$$\mathrm{H} = \log \mathrm{P} \tag{41}$$

When we substitute the probability P in Suppl. Equation 41 with $\mathrm{S}(A,B)$ from Suppl. Equation 39, we get

$$\begin{aligned} \mathrm{H}(A,B) &= -\frac{1}{Z} \sum_{i=1}^{N} \left(\tfrac{L_A^2+L_B^2}{L_A+L_B} \right) \log \frac{L_i^2(\theta_A) + L_i^2(\theta_B)}{2\, L_i(\theta_A)\, L_i(\theta_B)} \\ &= -\frac{1}{Z} D_{\mathrm{KL}}(\hat{L} \| L_{swap}) \end{aligned} \tag{42}$$

Suppl. Equation 42 is the negative Kullback-Leibler divergence over the sample data, which measures the loss of information when the suboptimal model with swapped parameters is used instead of the MLE parameter model, divided by the summed likelihood of the observed data.

Program versions

The results in this paper are based on MGLEX v.0.1.1. For the generation of taxonomic annotation, we used the BLAST alignment pipeline in taxator-tk v1.3.3e with refpack microbial-full_20150430, which includes reference nucleotide sequences and a corresponding version of the NCBI taxonomy. We also compared the submodel performance on simulated data with external programs. These are Centrifuge v1.0.3b, a sequence classifier based on sequence similarity, and NBC v1.1, a similar program based on short k-mers (nucleotide composition).

Supplementary Tables

Supplementary Table 1: Taxa in the simulated dataset and corresponding relative abundances for the primary sample S1 and the three secondary samples S2, S3 and S4.

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Acaryochloris CCMEE 5410	0.27	0.07	0.08	0.08
Acetobacteraceae bacterium AT-5844	0.04	0.01	0.01	0.01
Acholeplasma laidlawii PG-8A	0.12	0.79	0.04	0.04
Acidaminococcus fermentans DSM 20731	0.16	0.04	0.05	0.05
Acidaminococcus BV3L6	0.29	0.08	0.21	0.92
Acidovorax ebreus TPSY	0.09	0.03	0.03	0.03
Acidovorax KKS102	0.21	0.96	1.23	0.06
Aciduliprofundum MAR08-339	1.12	0.31	0.34	0.34
Acinetobacter baumannii AB_TG2028	0.83	1.08	0.25	0.25
Acinetobacter baumannii Naval-113	0.13	0.25	0.18	0.04
Acinetobacter baumannii ZWS1122	0.05	0.06	0.01	0.01
Acinetobacter genomosp. 13TU NCTC 8102	0.06	0.02	0.02	0.12
Acinetobacter johnsonii ANC 3681	0.02	0.00	0.00	0.13
Acinetobacter nosocomialis 28F	0.07	0.02	0.02	0.02
Acinetobacter schindleri NIPH 900	0.01	0.00	0.05	0.06
Acinetobacter schindleri TG19614	0.08	0.20	0.34	0.32
Acinetobacter CIP 64.7	0.25	0.07	0.08	0.08
Actinobacillus minor NM305	0.23	0.06	0.07	0.07
Actinoplanes SE50/110	0.51	0.14	0.16	0.15
Actinopolyspora mortivallis DSM 44261	0.19	0.05	0.06	0.06
Aeromonas MDS8	0.16	0.04	0.05	0.29
Aggregatibacter actinomycetemcomitans AAS4A	0.02	0.00	0.01	0.01
Aggregatibacter actinomycetemcomitans SCC393	0.06	0.02	0.02	0.02
Alicyclobacillus acidocaldarius Tc-4-1	0.02	0.01	0.02	0.01
Alistipes CAG:53	0.14	0.04	0.28	0.04
Alloprevotella rava F0323	0.26	0.07	0.08	0.08
alpha proteobacterium LLX12A	0.07	0.02	0.02	0.02
alpha proteobacterium SCGC AAA015-O19	0.04	0.15	0.01	0.35
alpha proteobacterium SCGC AAA536-G10	0.62	0.17	0.19	5.38
Alteromonas macleodii `Ionian Sea U8'	0.05	0.04	0.01	0.01
Amphibacillus xylanus NBRC 15112	0.10	0.03	0.09	0.03
Amycolatopsis mediterranei U32	0.07	0.02	0.02	0.02
Anaerococcus hydrogenalis ACS-025-V-Sch4	0.03	0.01	0.06	0.01
Anaerococcus hydrogenalis DSM 7454	0.18	0.05	0.06	0.06
Anaplasma marginale Florida	0.01	0.00	0.01	0.00
Anaplasma marginale Gypsy Plains	0.74	0.20	0.23	4.79
Anaplasma marginale St. Maries	0.52	0.14	4.64	0.16
Anoxybacillus SK3-4	0.15	0.04	0.05	0.05
Arthrobacter FB24	0.14	0.04	0.04	0.04
Arthrobacter TB 23	0.25	0.07	0.08	0.08
Azospirillum CAG:239	0.06	0.02	0.02	0.08
Bacillus amyloliquefaciens DC-12	0.34	0.09	0.11	0.10
Bacillus anthracis A0193	0.54	3.44	1.17	3.68

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Bacillus anthracis A1055	0.16	0.10	0.05	0.05
Bacillus cereus Rock1-15	0.04	0.01	0.04	0.04
Bacillus cereus Rock4-2	0.30	0.23	0.09	0.09
Bacillus cereus VD014	0.56	0.15	0.17	0.17
Bacillus pumilus ATCC 7061	0.14	0.37	0.04	0.04
Bacillus 37MA	0.14	0.04	0.20	0.04
Bacillus EGD-AK10	0.24	0.06	0.07	0.07
Bacillus WBUNB004	0.31	0.08	0.09	0.23
Bacillus WBUNB009	0.37	0.10	0.11	0.11
Bacillus subtilis gtP20b	0.14	0.04	0.98	0.15
Bacillus subtilis S1-4	0.46	0.13	0.14	0.14
Bacillus subtilis 6051-HGW	0.10	0.03	0.03	0.03
Bacillus thuringiensis BGSC 4CC1	0.12	0.03	0.04	0.04
Bacteriovorax DB6_IX	0.11	0.03	0.03	0.03
Bacteroides faecis CAG:32	0.06	0.05	0.33	0.02
Bacteroides fragilis CAG:558	0.08	0.06	0.03	0.03
Bacteroides 4_1_36	0.20	0.05	0.29	0.06
Bacteroides CAG:443	0.27	0.07	0.08	0.08
Bacteroides CAG:714	0.04	0.01	0.01	0.03
Beijerinckia indica ATCC 9039	0.06	0.02	0.22	0.07
Bifidobacterium longum CAG:69	0.02	0.02	0.01	0.01
Bizionia argentinensis JUB59	0.31	0.09	0.10	0.27
Bordetella bronchiseptica Bbr77	0.17	0.05	0.05	0.05
Borrelia burgdorferi 29805	0.48	0.13	0.15	0.15
Brachyspira hampsonii 30599	0.10	0.03	0.03	0.03
Bradyrhizobium DFCI-1	0.11	0.06	0.03	0.03
Bradyrhizobium S23321	0.30	2.08	0.09	0.09
Bradyrhizobium WSM2793	0.03	0.06	0.01	0.01
Brevibacillus laterosporus PE36	0.05	0.14	0.02	0.39
Brevibacterium casei S18	0.40	0.54	0.12	0.12
Brevibacterium mcbrellneri ATCC 49030	0.58	3.30	0.77	0.18
Brevundimonas abyssalis TAR-001	0.37	2.08	0.11	0.11
Brevundimonas BAL3	0.18	0.05	0.06	0.06
Brucella abortus 68-3396P	0.22	0.06	0.07	0.07
Brucella abortus NI274	0.17	0.25	0.20	0.05
Burkholderia bryophila 376MFSha3.1	0.04	0.01	0.01	0.01
Burkholderia mallei 2002721280	0.25	0.07	0.08	1.08
Burkholderia pseudomallei 668	0.16	1.10	0.05	0.05
Burkholderia pseudomallei DM98	0.13	0.04	0.04	0.04
Burkholderia CCGE1001	0.09	0.03	0.90	0.03
Burkholderia WSM4176	0.05	0.01	0.02	0.02
butyrate-producing bacterium SM4/1	0.27	2.17	0.08	0.08
Butyrivibrio crossotus CAG:259	0.07	0.02	0.06	0.02
Caldicellulosiruptor bescii DSM 6725	0.51	0.14	0.16	0.16
Caldivirga maquilingensis IC-167	0.06	0.02	0.02	0.02
Candidatus Accumulibacter phosphatis UW-1	0.25	0.07	0.08	0.08
Candidatus Photodesmus katoptron Akat1	0.20	0.57	0.06	0.22
Candidatus Poribacteria WGA-A3	0.06	0.02	0.02	0.02

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Candidatus Saccharibacteria RAAC3_TM7_1	0.34	0.75	0.10	0.10
Capnocytophaga F0502	0.08	0.02	0.02	0.02
Carnobacterium WN1359	0.29	0.29	0.09	0.09
Catellicoccus marimammalium M35/04/3	0.33	0.57	0.10	0.36
Chitinophaga pinensis DSM 2588	0.24	0.06	0.33	0.07
Chlamydia psittaci WC	0.05	0.01	0.02	0.20
Chlamydia trachomatis IU888	0.02	0.00	0.02	0.01
Chlamydia trachomatis L2b/Ams2	0.05	0.01	0.01	0.04
Chlamydia trachomatis RC-J/953	0.72	0.20	0.22	0.22
Chloroflexi bacterium oral isolate Chl1-2	0.35	0.09	0.11	0.11
Chloroflexi bacterium SCGC AB-629-P13	0.32	0.09	0.10	1.65
Citrobacter rodentium ICC168	0.08	0.02	0.02	0.02
Citrobacter KTE151	0.04	0.04	0.01	0.25
Clostridium acetobutylicum EA 2018	0.18	0.05	0.06	1.13
Clostridium carboxidivorans P7	0.23	0.25	2.18	0.07
Clostridium ATCC BAA-442	0.32	0.09	0.10	1.42
Clostridium CAG:269	0.38	0.10	0.83	1.36
Clostridium CAG:452	0.21	0.06	0.06	0.06
Clostridium CAG:567	0.44	0.12	0.14	0.53
Clostridium SY8519	0.10	0.03	0.03	0.03
Clostridium tyrobutyricum DSM 2637/ATCC 25755/JCM 11008	0.88	0.24	4.86	0.27
Collimonas fungivorans Ter331	0.19	0.05	0.33	0.06
Coprococcus comes CAG:19	0.10	0.03	0.09	0.03
Corynebacterium pseudotuberculosis 316	0.02	0.00	0.00	0.00
Corynebacterium pseudotuberculosis Cp162	0.08	0.02	0.02	0.20
Corynebacterium pseudotuberculosis I19	0.07	0.02	0.02	0.02
Corynebacterium KPL1855	0.82	4.06	0.25	0.60
Corynebacterium KPL1859	0.09	0.09	0.23	0.03
Corynebacterium KPL1998	0.09	0.03	0.03	0.26
Cronobacter sakazakii 701	0.11	0.03	0.03	0.03
Cupriavidus basilensis B-8	0.11	0.10	0.03	0.03
Cyanothece CCY0110	0.08	0.02	0.03	0.02
Cyclobacterium qasimii M12-11B	0.13	0.04	0.04	0.50
Desulfococcus oleovorans Hxd3	0.10	0.03	0.12	0.03
Desulfovibrio aespoeensis Aspo-2	0.22	0.06	0.07	0.07
Desulfurivibrio alkaliphilus AHT2	0.18	0.05	0.05	0.05
Dictyoglomus turgidum DSM 6724	0.39	0.11	0.12	0.12
Eggerthia catenaformis OT 569/DSM 20559	0.33	0.09	0.10	0.40
Emticicia oligotrophica DSM 17448	0.31	0.09	0.10	0.10
Enterobacter R4-368	0.07	0.02	0.02	0.02
Enterococcus flavescens ATCC 49996	0.08	0.02	0.02	0.02
Enterococcus GMD4E	1.14	0.31	0.35	0.35
Enterovibrio norvegicus FF-162	0.49	0.13	0.15	0.15
Erysipelotrichaceae bacterium 5_2_54FAA	0.27	0.15	0.08	0.08
Erythrobacter litoralis HTCC2594	0.86	0.23	0.26	0.26
Exiguobacterium pavilionensis RW-2	0.11	0.36	0.03	0.03
Facklamia ignava CCUG 37419	0.58	0.51	0.39	0.18
Faecalibacterium prausnitzii A2-165	0.08	0.02	0.02	0.02

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Finegoldia magna BVS033A4	0.07	0.02	0.07	0.02
Firmicutes bacterium ASF500	0.09	0.02	0.03	0.03
Firmicutes bacterium CAG:170	0.17	0.05	0.05	0.05
Fischerella thermalis PCC 7521	0.13	0.04	0.04	0.14
Flavobacteriaceae bacterium S85	0.41	0.11	0.13	0.13
Flavobacterium B17	0.14	0.04	0.04	0.04
Formosa AK20	0.64	0.18	0.20	0.20
Francisella tularensis 80700075	0.08	0.07	0.03	0.12
Frankia alni ACN14a	0.33	0.09	0.10	0.73
gamma proteobacterium IMCC2047	0.09	0.02	0.03	0.03
Gardnerella vaginalis 0288E	0.04	0.01	0.01	0.01
Gardnerella vaginalis 1500E	0.27	0.07	0.08	0.08
Geobacillus JF8	0.11	0.72	0.04	0.03
Gillisia marina	0.41	0.11	0.13	0.13
Glaciecola polaris LMG 21857	0.26	0.07	0.08	0.08
Glaciecola 4H-3-7+YE-5	0.33	0.21	0.10	0.10
Gordonia effusa NBRC 100432	0.09	0.03	0.45	0.03
Gordonia sihwensis NBRC 108236	0.12	0.24	0.04	0.11
Haemophilus aegyptius ATCC 11116	0.48	0.13	0.15	0.15
Haemophilus somnus 129PT	0.31	0.09	0.10	0.10
Haemophilus sputorum HK 2154	0.94	0.26	0.29	0.29
Haloferax BAB2207	0.43	0.12	0.13	0.13
Halomonas KM-1	0.13	0.03	0.04	1.02
Halorhabdus utahensis DSM 12940	0.01	0.00	0.04	0.00
Haloterrigena limicola JCM 13563	0.04	0.01	0.01	0.01
Helicobacter hepaticus ATCC 51449	0.38	0.10	3.30	0.39
Herbaspirillum B39	0.41	0.11	0.12	0.12
Ignavibacterium album JCM 16511	0.39	0.11	0.12	0.12
Isoptericola variabilis 225	0.20	0.05	0.06	0.06
Janibacter HTCC2649	0.43	0.12	0.13	0.95
Kingella kingae PYKK081	0.19	0.05	0.12	0.06
Klebsiella pneumoniae UHKPC01	0.18	1.40	0.06	0.06
Klebsiella pneumoniae UHKPC02	0.14	0.04	1.04	0.04
Klebsiella pneumoniae UHKPC40	0.19	0.05	0.06	1.46
Ktedonobacter racemifer DSM 44963	0.14	0.04	0.04	0.04
Laceyella sacchari 1-1	0.08	0.02	0.02	0.21
Lachnospiraceae bacterium 2_1_46FAA	0.54	1.60	0.17	0.16
Lachnospiraceae bacterium 3-2	0.33	0.09	0.79	0.63
Lachnospiraceae bacterium 5_1_57FAA	0.04	0.22	0.01	0.37
Lachnospiraceae oral taxon 107 str. F0167	0.19	0.35	0.06	0.06
Lactobacillus acidipiscis KCTC 13900	0.04	0.01	0.01	0.01
Lactobacillus acidophilus 30SC	0.38	0.10	0.77	0.11
Lactobacillus acidophilus ATCC 4796	0.04	0.01	0.01	0.01
Lactobacillus casei 21/1	0.22	0.43	0.07	0.07
Lactobacillus casei Lpc-37	0.21	0.06	0.06	0.06
Lactobacillus delbrueckii ATCC BAA-365	0.03	0.01	0.01	0.01
Lactobacillus delbrueckii DSM 20072	0.32	0.09	1.90	2.76
Lactobacillus fermentum CECT 5716	0.42	0.11	0.13	0.13

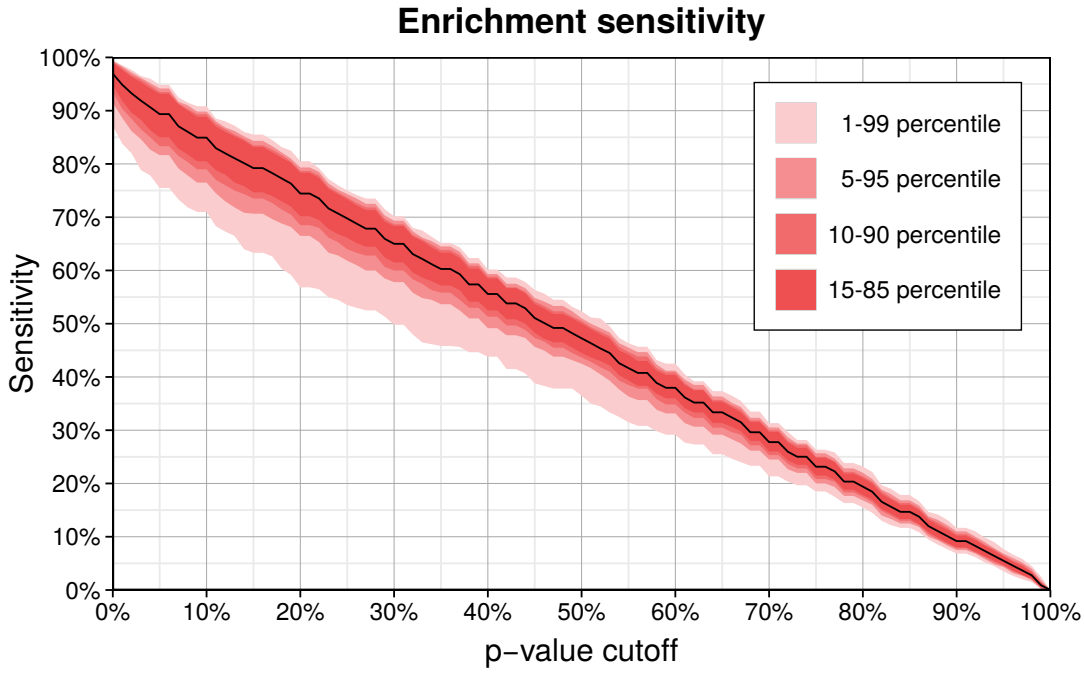
Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Lactobacillus helveticus CNRZ32	0.08	0.20	0.02	0.02
Lactobacillus helveticus R0052	0.10	0.03	0.53	0.03
Lactobacillus iners ATCC 55195	0.06	0.02	0.02	0.06
Lactobacillus iners LactinV 01V1-a	0.11	0.03	0.03	0.03
Lactobacillus plantarum 2165	0.85	0.23	0.26	0.54
Lactobacillus reuteri CF48-3A	0.66	0.18	0.20	0.20
Lactobacillus reuteri MM4-1A	0.23	1.14	0.07	0.07
Lactobacillus salivarius GJ-24	0.42	0.12	0.13	1.37
Lactobacillus ASF360	0.28	0.08	1.93	0.09
Legionella pneumophila str. 121004	0.05	0.06	0.01	0.29
Leifsonia xyli subxyli str. CTCB07	0.04	0.01	0.01	0.01
Leptospira borgpetersenii 200801910	0.20	0.05	0.06	0.06
Leptospira borgpetersenii 200901122	0.24	0.17	0.38	0.16
Leptospira interrogans Fiocruz R154	0.15	0.04	0.05	0.05
Leptospira interrogans L1207	0.11	0.03	0.04	0.03
Leptospira santarosai Oregon	0.59	3.58	0.18	0.18
Leptospira santarosai 2000027870	0.12	0.03	0.24	0.04
Leptospira santarosai HAI1380	0.13	0.04	0.04	0.04
Leuconostoc argentinum KCTC 3773	0.13	0.04	0.09	0.53
Leuconostoc citreum LBAE C10	0.02	0.01	0.20	0.07
Loktanella cinnabarina LL-001	0.23	0.06	0.07	0.07
Loktanella hongkongensis DSM 17492	0.18	0.05	0.06	0.53
Mannheimia haemolytica USDA-ARS-USMARC-183	0.19	0.05	0.06	0.06
marine gamma proteobacterium HTCC2080	0.10	0.06	0.30	0.03
Marinimicrobia bacterium SCGC AAA298-D23	0.26	0.65	0.08	0.08
Marinimicrobia bacterium SCGC AB-629-J13	0.26	0.07	0.08	0.08
Marinobacter EVN1	0.05	0.01	0.43	0.15
Megasphaera genomosp. type_1 str. 28L	0.14	0.04	0.47	0.04
Melissococcus plutonius DAT561	0.39	0.58	0.12	0.12
Mesoflavibacter zeaxanthinifaciens S86	0.11	0.03	0.21	0.03
Mesorhizobium LNHC229A00	0.16	1.18	0.11	0.05
Mesorhizobium LSHC416B00	0.04	0.01	0.01	0.01
Mesorhizobium LSJC264A00	0.04	0.08	0.01	0.01
Methanobrevibacter smithii TS146D	0.14	0.04	0.80	0.04
Methanobrevibacter smithii TS147C	0.14	0.04	0.04	0.11
Methanobrevibacter smithii TS95A	0.09	0.02	0.17	0.03
Methanocella arvoryzae MRE50	0.05	0.13	0.02	0.13
Methanosphaera stadtmanae DSM 3091	0.18	0.05	0.06	0.05
Methylobacterium extorquens PA1	0.10	0.78	0.03	0.03
Methyloglobulus morosus KoM1	0.19	0.05	0.06	1.18
Methylotenera versatilis 301	0.06	0.02	0.02	0.05
Methyloversatilis universalis EHg5	0.17	0.05	0.05	0.05
Microbacterium barkeri 2011-R4	0.12	0.03	0.04	0.04
Microbacterium 11MF	0.13	0.08	0.37	0.04
Microbacterium TS-1	0.10	0.03	0.03	0.03
Mobiluncus curtisii ATCC 43063	0.39	0.11	0.12	0.12
Mycobacterium abscessus 3A-0930-R	0.03	0.01	0.01	0.01
Mycobacterium abscessus 5S-0422	0.58	0.16	0.37	0.99

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Mycobacterium abscessus M139	0.26	0.07	0.08	0.08
Mycobacterium chubuense NBB4	0.18	0.05	0.06	0.05
Mycobacterium intracellulare MOTT-02	0.19	0.05	0.06	0.12
Mycoplasma gallisepticum NC08_2008.031-4-3P	0.05	0.02	0.02	0.02
Mycoplasma gallisepticum NY01_2001.047-5-1P	0.27	0.07	0.08	0.08
Neisseria gonorrhoeae PID18	0.13	0.03	0.04	0.04
Neisseria gonorrhoeae SK-92-679	0.13	0.04	0.04	0.04
Neisseria meningitidis NM1476	0.18	0.16	0.05	0.05
Neisseria meningitidis NM3223	0.16	0.05	0.05	0.15
Neisseria meningitidis NM604	0.27	0.07	0.08	0.08
Neisseria sicca 4320	0.09	0.23	0.03	0.03
Niabella aurantiaca DSM 17617	0.15	0.23	0.32	0.04
Nitrolancea hollandica Lb	0.36	0.78	1.24	0.11
Nocardia tenerifensis NBRC 101015	0.40	0.11	0.12	1.67
Nocardiopsis CNS639	0.74	0.20	0.23	0.23
Nonomuraea coxensis DSM 45129	0.69	1.26	0.21	1.07
Oceanicaulis HTCC2633	0.19	0.05	0.34	0.06
Oceanobacillus kimchii X50	0.74	2.17	0.23	0.86
Octadecabacter arcticus 238	0.06	0.02	0.02	0.02
Paenibacillus alvei TS-15	0.19	0.05	1.36	0.06
Paenibacillus larvae BRL-230010	0.03	0.01	0.01	0.01
Paenibacillus Aloe-11	0.04	0.01	0.01	0.01
Pantoea AS-PWVM4	0.09	0.02	0.03	0.03
Parabacteroides ASF519	0.19	0.05	0.06	0.85
Parascardovia denticolens IPLA 20019	0.42	0.11	2.48	3.38
Parasutterella excrementihominis CAG:233	0.72	0.20	0.22	1.60
Patulibacter americanus DSM 16676	0.07	0.02	0.02	0.55
Patulibacter medicamentivorans	0.45	0.12	1.09	0.14
Pediococcus acidilactici D3	0.07	0.05	0.02	0.02
Pelosinus fermentans A11	0.04	0.01	0.23	0.03
Peptoclostridium difficile P20	0.12	0.03	0.04	0.53
Peptoclostridium difficile P48	0.04	0.01	0.01	0.08
Peptoclostridium difficile P53	0.24	0.07	1.05	0.07
Polynucleobacter necessarius QLW-P1DMWA-1	0.09	0.02	0.03	0.03
Porphyromonas gingivalis JCVI SC001	0.17	0.05	0.23	1.25
Porphyromonas gingivalis W50	0.81	0.22	0.25	0.25
Porphyromonas macacae DSM 20710/JCM 13914	0.11	0.03	0.03	0.03
Prevotella salivae DSM 15606	0.04	0.01	0.01	0.01
Prevotella C561	0.03	0.19	0.01	0.01
Prevotella CAG:1185	0.39	0.11	3.30	0.12
Prevotella CAG:592	0.35	1.04	1.14	0.11
Prevotella CAG:617	0.32	0.09	0.10	0.91
Prevotella CAG:755	0.14	0.04	0.04	0.04
Prevotella CAG:873	0.07	0.02	0.02	0.02
Pseudomonas aeruginosa BWHPSA006	0.10	0.03	0.03	0.03
Pseudomonas aeruginosa LESB58	0.23	0.20	1.02	0.07
Pseudomonas aeruginosa PABL056	0.09	0.03	0.03	0.03
Pseudomonas mendocina ymp	0.10	0.03	0.03	0.18

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Pseudomonas CF161	0.07	0.02	0.02	0.02
Pseudomonas EGD-AK9	0.04	0.01	0.03	0.01
Pseudomonas M47T1	0.28	0.08	0.09	0.39
Pseudomonas TJI-51	0.03	0.01	0.01	0.01
Pseudomonas syringae pv. lachrymans M302278	0.25	0.70	1.23	0.08
Psychrobacter PRwf-1	0.03	0.01	0.01	0.01
Pyrobaculum aerophilum str. IM2	0.32	0.09	0.10	0.10
Pyrobaculum calidifontis JCM 11548	0.03	0.01	0.01	0.07
Pyrococcus furiosus COM1	0.36	0.10	0.11	0.11
Ralstonia solanacearum Po82	0.18	0.05	0.06	0.05
Renibacterium salmoninarum ATCC 33209	0.49	0.13	0.15	0.15
Rhizobium etli Brasil 5	0.02	0.01	0.01	0.01
Rhizobium phaseoli Ch24-10	0.08	0.33	0.02	0.02
Rhizobium IRBG74	0.07	0.12	0.02	0.70
Rhodobacter SW2	0.17	0.05	0.05	0.05
Rhodobacter sphaeroides ATCC 17029	0.47	0.13	0.14	2.85
Rhodobacteraceae bacterium KLH11	1.39	3.41	3.28	0.42
Rhodococcus rhodnii LMG 5362	0.24	1.34	0.07	0.22
Rhodococcus 29MFTsu3.1	0.06	0.02	0.02	0.02
Rhodococcus P27	0.22	0.06	0.07	0.07
Rhodopirellula baltica SWK14	0.55	0.15	0.17	0.17
Rhodopseudomonas palustris BisB5	0.28	0.08	0.08	0.08
Rhodospirillum rubrum ATCC 11170	0.02	0.01	0.01	0.05
Rickettsia helvetica C9P9	0.19	0.05	0.06	0.06
Rickettsia rickettsii str. `Sheila Smith'	0.49	0.32	1.97	1.27
Riemerella anatipestifer RA-YM	0.08	0.02	0.03	0.65
Rudanella lutea DSM 19387	0.72	0.20	0.22	0.22
Ruminiclostridium thermocellum ATCC 27405	0.42	0.11	0.13	0.13
Ruminiclostridium thermocellum YS	0.55	0.15	1.98	0.17
Ruminococcus CAG:382	0.10	0.03	0.03	0.03
Ruminococcus CAG:579	0.88	2.32	0.27	0.27
Saccharomonospora cyanea NA-134	0.13	0.04	0.04	0.04
Salinispora arenicola CNT849	0.86	0.23	0.26	0.26
Salinispora arenicola CNY234	0.61	0.17	0.19	0.19
Salinispora pacifica CNY330	0.52	0.72	0.16	0.16
Salmonella enterica SA-2	0.05	0.01	0.02	0.02
Salmonella enterica CFSAN001588	0.13	0.04	0.89	0.04
Selenomonas noxia ATCC 43541	0.06	0.02	0.02	0.02
Shewanella frigidimarina NCIMB 400	0.04	0.04	0.04	0.01
Shigella boydii 965-58	0.18	0.05	0.58	0.05
Shigella dysenteriae CDC 74-1112	0.20	1.73	0.50	0.45
Shigella flexneri 1485-80	0.11	0.03	0.03	0.03
Shigella flexneri 2930-71	0.11	0.03	0.03	0.03
Simonsiella muelleri ATCC 29453	0.31	0.08	0.09	0.09
Sphingomonas melonis DAPP-PG 224	0.48	0.13	0.15	0.15
Sphingopyxis MC1	0.07	0.07	0.02	0.08
Staphylococcus hominis SK119	0.09	0.03	0.03	0.08
Streptococcus agalactiae GB00264	0.09	0.05	0.03	0.03

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Streptococcus agalactiae MRI Z1-022	0.14	0.04	0.04	0.04
Streptococcus agalactiae MRI Z1-202	0.38	0.90	0.12	0.12
Streptococcus anginosus F0211	0.10	0.03	0.37	0.03
Streptococcus equi	0.37	0.10	1.54	0.11
Streptococcus equi SzS31A1	0.31	0.08	0.09	0.27
Streptococcus ferus DSM 20646	0.15	0.04	0.05	0.05
Streptococcus gordonii CH1	0.26	0.07	1.11	0.08
Streptococcus iniae 9117	0.01	0.00	0.01	0.09
Streptococcus intermedius ATCC 27335	3.19	0.87	0.98	0.97
Streptococcus mutans KK23	0.62	0.17	0.19	0.19
Streptococcus mutans SM6	0.10	0.03	0.03	0.03
Streptococcus pseudoporcinus LQ 940-04	0.03	0.01	0.01	0.01
Streptococcus salivarius 57.I	0.17	0.14	0.05	0.05
Streptococcus sanguinis SK340	0.02	0.06	0.01	0.03
Streptococcus sobrinus DSM 20742/ATCC 33478	0.42	0.11	0.13	0.29
Streptococcus sobrinus TCI-367	1.65	0.45	0.50	0.50
Streptococcus sobrinus TCI-98	0.23	0.06	0.34	0.07
Streptococcus I-P16	0.05	0.01	0.01	0.01
Streptococcus SK140	0.43	0.12	0.13	0.13
Streptococcus suis YB51	0.42	0.11	0.13	0.13
Streptomyces acidiscabies 84-104	0.22	0.26	0.07	0.07
Streptomyces albulus CCRC 11814	0.40	0.11	0.92	0.12
Streptomyces pristinaespiralis ATCC 25486	0.11	0.27	0.03	0.09
Streptomyces CNQ766	0.17	0.05	0.15	0.05
Streptomyces sulphureus DSM 40104	0.13	0.11	0.45	0.04
Streptomyces violaceusniger Tu 4113	0.27	0.08	0.08	0.08
Succinatimonas hippei YIT 12066	0.10	0.03	0.03	0.08
Sulfolobus islandicus REY15A	0.10	0.03	0.44	0.03
Synechococcus PCC 7336	0.42	0.12	0.13	0.13
Synechocystis PCC 6803	0.04	0.01	0.30	0.01
Synechocystis PCC 7509	0.04	0.05	0.01	0.01
Thauera linaloolentis 47Lol/DSM 12138	0.28	0.08	0.08	0.08
Thermococcus onnurineus NA1	0.15	0.18	0.09	0.04
Thermoplasmatales archaeon I-plasma	0.13	0.04	0.04	0.04
Thermosphaera aggregans DSM 11486	0.69	0.74	0.21	1.64
Thermotoga elfii NBRC 107921	0.16	0.04	0.24	0.91
Thermotoga EMP	0.51	0.14	0.16	3.02
Thermus CCB_US3_UF1	0.17	0.05	0.05	0.05
Thioalkalivibrio AKL6	0.36	0.10	0.11	0.11
Thioalkalivibrio ALE20	0.38	0.61	0.12	0.11
Thioalkalivibrio ALJ10	0.60	2.66	0.19	0.18
Thioalkalivibrio ALJ12	0.81	0.22	0.65	0.25
Thioalkalivibrio ALJ24	0.48	3.07	0.15	0.15
Thioalkalivibrio ALJ5	0.10	0.03	0.28	0.47
Thioalkalivibrio ALJ9	0.10	0.03	0.03	0.03
Tyzzerella nexilis DSM 1787	0.16	0.04	0.05	0.05
uncultured archaeon A07HR60	0.67	0.18	2.13	5.67
Ureaplasma urealyticum ATCC 27814	0.32	0.09	1.24	0.10

Name	S1 (%)	S2 (%)	S3 (%)	S4 (%)
Variovorax paradoxus S110	0.08	0.02	0.11	0.07
Verrucomicrobium 3C	1.48	7.16	2.75	0.45
Vibrio cholerae HC-50A2	0.16	0.04	0.05	0.05
Vibrio cholerae HE39	0.08	0.02	0.02	0.02
Vibrio cholerae O1 str. 2009V-1085	0.03	0.01	0.08	0.01
Vibrio crassostreae 9ZC88	0.18	0.05	0.05	0.05
Vibrio gazogenes ATCC 43941	0.96	2.34	0.29	0.29
Vibrio nigripulchritudo ENn2	0.71	0.20	0.22	0.22
Vibrio nigripulchritudo SFn135	0.22	0.06	1.80	0.07
Vibrio nigripulchritudo SOn1	0.46	0.13	0.14	0.14
Weissella koreensis KACC 15510	0.25	0.38	0.25	0.08
Wolbachia endosymbiont JHB	0.39	0.11	0.12	0.12
Xanthomonas axonopodis IBSBF 614	0.06	0.02	0.10	0.02
Xanthomonas axonopodis UA306	0.15	0.04	0.05	0.05
Xanthomonas campestris NCPPB 2005	0.17	0.05	0.05	0.05
Xanthomonas oryzae BLS256	0.18	0.05	0.06	0.06
Xanthomonas SHU166	0.13	0.04	0.04	0.04
Xylella fastidiosa 32	0.18	0.23	0.05	0.05
Yersinia frederiksenii ATCC 33641	0.22	1.08	0.07	1.20
Yersinia pseudotuberculosis B-6863	0.22	0.06	0.07	0.07
Yersinia pseudotuberculosis B-6864	0.12	1.02	0.13	0.19



Supplementary Figure 1: Genome enrichment for 400 genomes in the three-fold cross-validation. For each genome, we measured the sensitivity, the percentage of each genome in the enriched sample, after filtering by a p-value cutoff and summing over the three data partitions. The solid lines shows the resulting average sensitivity over all 400 genomes. The variability between genomes is shown as quantiles in red.

Appendix C

Taxonomic Binning of Metagenome Samples Generated by Next-generation Sequencing Technologies

J. Dröge[1,2] and **A. C. McHardy**[1,2*]

[1]Max Planck Research Group for Computational Genomics & Epidemiology, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany

[2]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Institute for Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany

---

This is a pre-copyedited, author-produced version of an article accepted for publication in *Briefings in Bioinformatics* following peer review. This article version has been adapted to the thesis layout. The original article is available online by DOI 10.1093/bib/bbs031.

C.1 Abstract

Metagenome research uses random shotgun sequencing of microbial community DNA to study the genetic sequences of its members without cultivation. This development has been strongly supported by improvements in sequencing technologies, which have rendered sequencing cheaper than before. As a consequence, downstream computational analysis of metagenome sequence samples is now faced with large amounts of complex data. One of the essential steps in metagenome analysis is reconstruction of draft genomes for populations of a community or of draft 'pan-genomes' for higher-level clades. 'Taxonomic binning' corresponds to the process of assigning a taxonomic identifier to sequence fragments, based on information such as sequence similarity, sequence composition or read coverage. This is used for draft genome reconstruction, if sequencing coverage is insufficient for reconstruction based on assembly information alone. Subsequent functional and metabolic annotation of draft genomes allows a genome-level analysis of novel uncultured microbial species and even inference of their cultivation requirements.

C.2 Introduction

The application of genome sequencing technologies to the study of an entire community of microbial organisms, as opposed to a clonal culture of an individual isolate strain, is known as metagenomics (Kunin et al., 2008; Simon & Daniel, 2011). Such analysis allows one to determine genome sequence information for a vast portion of the microbial world for which cultivation conditions are unknown or difficult to reproduce under laboratory conditions (Amann, Ludwig & Schleifer, 1995; Hugenholtz, 2002). Even the first metagenome studies, investigating the Sargasso Sea (Venter et al., 2004) and Minnesota farm soil (Tringe et al., 2005), were able to demonstrate the enormous potential of the microbial world to serve as a treasure trove of genes with novel functionalities, as these studies resulted in the discovery of many thousands of new gene sequences that were only remotely similar to genes of known function. They also revealed the unexpected complexity of microbial communities in terms of the number of taxa contained therein.

Since then, much research has explored microbial ecosystems, soil, aquatic and host-associated, in more detail (Woyke et al., 2006; Warnecke et al., 2007; Turnbaugh et al., 2010; Suen et al., 2010; Mackelprang et al., 2011), and has revealed a great wealth of novel genetic information from microbial species that are only distantly related to well studied model organisms.

Both amplicon sequencing and random shotgun sequencing of microbial communities are sometimes referred to as metagenomics. Amplicon sequencing, or environmental tag sequencing, is used to determine the taxonomic composition and phylogenetic structure of a microbial community. In amplicon sequencing, informative marker regions of the genomes from a microbial community are amplified by polymerase chain reaction, and used as a proxy to determine which phylotypes or operational taxonomic units (OTUs) are present in a microbial community, and their relative abundance. Commonly used markers regions are the ribosomal genes (Huse et al., 2008) and the ITS (internal transcribed spacer) region (Jeewon & Hyde, 2007), which is positioned between ribosomal genes. In terms of numbers and the evolutionary closeness of the distinct species present, microbial community profiles can be correlated across environments and communities, linked to environmental parameters. They can be indicative of the presence of genes that are relevant for particular metabolic functionalities (Fuhrman, 2009), given that the respective genes are already known. However, the gene inventory and the encoded functionality of most microbial species are largely unknown and may also vary considerably between strains.

Shotgun sequencing can be used to study the genetic information of microbial communities by sequencing DNA that has been extracted and randomly sheared into smaller fragments. Even though subject to different technology-dependent biases, this procedure allows functional and process-level characterization of microbial communities as a whole and the reconstruction of draft genome sequences for individual community members.

C.3 Next-generation sequencing technologies

DNA sequencing technologies have rapidly advanced over the last five years and these developments have substantially shaped the way metagenome research is performed. Post-Sanger sequencing technologies are commonly referred to as next-generation sequencing (NGS) (Mardis, 2008; Metzker, 2009). In comparison to Sanger sequencing, NGS methods can sequence DNA more quickly and at lower cost through massive parallelization. This is generally achieved by amplification and fixation of millions of individual template molecules or their enzyme counterparts on a solid phase prior to sequencing. While Sanger sequencing results in read lengths of around 800 bp, the commercially available NGS technologies (tbl. C.1) currently generate reads of approximately 50–75 bp (Applied Biosciences/Life Technologies – SOLiD), 75–150 bp (Solexa/Illumina – Sequencing by Synthesis), 100–200 bp (IonTorrent/Life Technologies – Semiconductor Chip Sequencing) and 550–1000 bp (454/Roche – Pyrosequencing). The upcoming generation (Schadt, Turner & Kasarskis, 2010; Thompson & Milos, 2011) of sequencers using single molecule sequencing produces read lengths of over 1 kb (PacBio, SMRT, 15–20% assumed error rate (Schadt, Turner & Kasarskis, 2010)) and of 5–10 kb (Oxford Nanopore technology, 5% assumed error rate). Besides different read lengths and amounts of sequence data produced, each technology has a characteristic profile of sequencing errors, resulting from the technology-specific preparation and detection procedures. The choice of an appropriate sequencing technology depends on the scientific questions asked. For instance, while an 80 bp read is sufficient to cover a hyper-variable region in the 16S gene (Huse et al., 2008) for analysis of microbial community composition, de novo recovery of draft microbial genome sequences by taxonomic binning from a complex organismal mixture requires substantially longer reads or higher sequencing depth and sequencing of short paired reads (Turnbaugh et al., 2010; Hess et al., 2011; Mackelprang et al., 2011; Iverson et al., 2012).

Table C.1: Throughput and read lengths of different sequencing technologies. *: Normalized throughput is scaled to a one-hour period and rounded. **: The throughput scale is compared to Life Technologies 3730 Sanger chemistry based sequencer and shows the ratio of throughput values in terms of order of magnitude. ***: Numbers are based on vendor information: Illumina Inc. (www.illumina.com), Life Technologies (www.lifetechnologies.com), Roche/454 (www.454.com). Due to lack of information on sequencing statistics or commercial availability, Pacific Biosciences (www.pacificbiosciences.com), Oxford Nanopore Technologies (www.nanoporetech.com) and Helicos Biosciences (www.helicosbio.com) are excluded.

Manufacturer & technology	Length (bp)	Throughput***	Normalized throughput* (Mb/h)	Throughput scale**	Time per run
Solexa/ Illumina Sequencing by Synthesis	100 – 150	300 Gb/8.5 d – 600 Gb/11 d	1,500 – 2,300	10^4	8.5 d – 11 d
Life Technologies/ Applied Biosystems SOLiD	50 – 75	7 Gb/d – 20 Gb/d	300 – 800	10^3 - 10^4	2 d – 7 d
Life Technologies/ Ion Torrent	100 – 200	10 Mb/2 h – 1 Gb/2 h	5 – 500	10^1 - 10^3	2 h
Roche/ 454 Pyrosequencing	550 – 1000	450 Mb/10 h – 700 Mb/23 h	30 – 45	10^2	10 h – 23 h
Life Technologies Capillary Sanger sequencing	600 – 900	690 Kb/d – 2,100 Kb/d	0.029 – 0.088	10^0	~ 7 h [15]

C.4 Bioinformatic analysis of metagenome samples

NGS produces large volumes of sequence data (tbl. C.1). Currently, a single run of an Illumina HiSeq machine generates up to 600 Gb per run (www.illumina.com), which is of the order of 104 times the amount of data produced in a similar timeframe by a Sanger sequencing chemistry based sequencer (tbl. C.1). This, in turn, results in drastically increased runtimes for all the bioinformatics procedures applied in metagenomics (Wilkening et al., 2009), such as assembly of sequence fragments, taxonomic binning, prediction of protein encoding genes, as well as functional and process-level gene annotation. Together, taxonomic binning and assembly allow draft genome reconstructions for community members for which sequencing has recovered substantial amounts of sequence. Assembly corresponds to the computational process of placing individual reads into longer pieces of contiguous sequences, known as contigs, based on sequence overlaps and paired read information. Taxonomic binning sorts the contigs of a metagenome sample into 'bins' that represent the populations or higher-level clades of community members. Though both tasks are performed independently and evaluate different types of information, the problem of metagenome sequence assembly is closely related to taxonomic binning, as both allow the reconstruction of draft genome sequences. The terms "taxonomic" and "phylogenetic" binning are both used in the literature, as modern taxonomies such as the NCBI taxonomy (Sayers et al., 2009) or the ribosomal gene based RDP-II (Cole et al., 2009), GreenGenes (DeSantis et al., 2006) and ARB-SILVA (Pruesse et al., 2007) taxonomies are built upon phylogenetic principles. Even though it is less consistent, taxonomic binning software for shotgun metagenomics most frequently relies on the NCBI taxonomy, probably due to its widespread use in annotation of public sequence data.

Similar to the assembly of individual isolated genomes (Miller, Koren & Sutton, 2010), assembly in metagenomics aims to recover long contiguous pieces of sequence from the sequence collection of reads that represent parts of the genomes of individual community members. Massively increased amounts of data, vary-

ing organism abundances within a sampled community, differing complexities in terms of the overall number of organisms contained and the presence of multiple closely related organisms all challenge the sequence assemblers that were originally designed for isolated genomes. To address these challenges, methods designed for assembly of microbial community NGS data (Laserson, Jojic & Koller, 2011; Peng et al., 2011; Koren, Treangen & Pop, 2011; Pell et al., 2012) are being developed. Paired-end or mate-pair protocols, which add distance information between two individual reads, can greatly aid in the assembly process. Assembly information such as the ordering of contigs within a scaffold can also be used to check binning quality, and binning has been used to refine assembly in a feedback process. In recent studies, the joint analysis of assembly information and sequence composition allowed the reconstruction of several partial genomes by taxonomic binning (Hess et al., 2011; Iverson et al., 2012). Thus, a closer integration of the two approaches appears promising for draft genome reconstruction from NGS metagenome data.

Following assembly and binning, further bioinformatic analyses include the prediction of genes, as well as functional annotation and reconstruction of potential pathways. For these steps, dedicated web servers exist, such as MG-RAST (Meyer et al., 2008), IMG/M (Markowitz et al., 2012) and CAMERA (Sun et al., 2011). Analysis of the gene content of individual bins allows inference of the functional and metabolic capabilities of individual community members, and allows a metagenome sample to be studied in its entirety. If read lengths or sequencing depth are insufficient for assembly, the functional analysis of a metagenome sample is restricted to what can be inferred without partial genome reconstructions for individual community members.

C.5 Binning strategies

The term binning was originally coined for the problem of separating the sequence fragments of a metagenome according to the microbial populations they originate from (Tyson et al., 2004; Woyke et al., 2006). The definition has been extended to include bins that represent all fragments that originate from a com-

mon higher-level clade, in cases where resolution down to individual populations is not possible. For placement of sequence fragments into taxonomic bins, attributes which are indicative of the taxonomic origin of a fragment are evaluated. Different types of information can be used for this purpose: (a) local sequence similarity to sequences of known taxa (used in similarity-based taxonomic assignment), (b) similarity in sequence composition to sequences of a given taxon (used in composition-based taxonomic assignment) or to other sequences in the sample (used in composition-based clustering), or (c) similarity in read coverage and linkage information from assembly for contigs within a metagenome sample. The underlying rationale of using read coverage is that similar coverage of two contigs in the sample indicates similar abundance and therefore potentially the same underlying source population in the community.

How accurately fragments can be assigned to taxonomic bins depends on several factors. The first is fragment length. Shorter, noisier fragments cannot be assigned as accurately as longer fragments of 2 kb or more (Patil et al., 2011). In particular, assignment of individual reads or of fragments below 1 kb in length poses significant challenges. Reported assignment accuracies for 100 bp fragments to a clade at the genus level are 60% under somewhat idealized conditions, with only reference data from the same species being removed. This, however, means that 40% of fragments are misassigned (Brady & Salzberg, 2009). Furthermore, accuracy drops to less than 30% if the reference data is depleted of sequences from the same genus, meaning 70% of 100 bp fragments are misassigned at the family level.

Another influential factor for binning accuracy is the community's complexity in terms of the number of distinct phylotypes it comprises. Metagenome sequencing of complex communities, such as those found in soil (Mackelprang et al., 2011), results in lower sequencing coverage of most populations and therefore shorter contigs in assembly. This amounts to many short fragments, or even predominantly unassembled samples, which have to be separated into a multitude of taxonomic bins. The larger the number of bins, the harder the problem becomes, as the chances of randomly assigning a fragment correctly decrease with increasing numbers of bins. Finally, for taxonomic assignment, the availability of

reference data from taxa that are closely related to the microbes of the sequenced community is important for accurate assignment. Similarity-based assignment of metagenome shotgun sequence data requires homologous reference sequences from related taxa to be available for a fragment to be assigned; ideally, entire sequenced genomes should be available. The sequencing of many isolate genomes of the human microbiome in the Human Microbiome Project has immensely helped similarity-based taxonomic assignment of human gut metagenome samples (Qin et al., 2010; Nelson et al., 2010). A 'shallow' (i.e. to high-ranking clades only) taxonomic assignment of a sample based on sequence similarities indicates the presence of many taxa that are only distantly related to isolated sequenced genomes. If no sequenced genomes from related taxa are available, composition-based assignment can be used for higher resolution taxonomic binning. Clustering of metagenome fragments based on sequence composition does not require reference sequences and comparably small amounts of non-homologous reference sequences are required for composition-based taxonomic classification.

Table C.2: Overview of existing web applications for taxonomic assignment and phylotyping of metagenome sequence samples. Phylotyping methods assign only a subset of contigs based on taxonomic marker genes.

Name	Phylo-typing	Tax. assign-ment	Funct. anno-tation	Techniques & web link
CAMERA v.2 (2011)	X	—	X	Reverse Psi-BLAST http://camera.calit2.net
MetaABC (2011)	—	X	—	BLAST, PhymmBL, MEGAN, Sort-ITEMS http://bits2.iis.sinica.edu.tw/MetaABC/
MG-RAST v.3.1.2 (2008)	X	—	X	BLAST/BLAT http://metagenomics.anl.gov
MLTreeMap v.2.06.1 (2010)	X	—	X	BLAST, HMMER, RaxML http://mltreemap.org

Name	Phylo-typing	Tax. assign-ment	Funct. anno-tation	Techniques & web link
NBC v.1.1 CLI (2011)	—	X	—	Naïve Bayesian Classifier http://nbc.ece.drexel.edu
PhyloPythia (2007), PhyloPythiaS (2011)	—	X	—	(Structured) Support Vector Machine http://binning.bioinf.mpi-inf.mpg.de
TaxSOM (2011)	—	X	—	Self-Organizing Maps http://soma.arb-silva.de
WebCARMA v.3.0 (2011)	X	X	X	BLAST, HMM search versus Pfam http://webcarma.cebitec.uni-bielefeld.de

C.6 Taxonomic binning based on sequence similarities

Similarity-based taxonomic assignment utilizes the local similarity of a query sequence to sequences of known taxonomic origin. Taxonomic identifiers are commonly assigned either by identifying the lowest common ancestor (LCA) from the taxonomy for the taxa of the most similar sequences found (Patil et al., 2011) or by using phylogenetic placement methods. Phylogenetic placement methods, such as pplacer (Matsen, Kodner & Armbrust, 2010), EPA/RaxML (Berger, Krompass & Stamatakis, 2011) and SEPP (Mirarab, Nguyen & Warnow, 2012) place the query sequence within a fixed reference tree. The taxonomic label assigned then corresponds to the LCA of the taxa associated with the first ancestral node's children. Both methods are related to 'nearest neighbor' classification. In both cases, there has to be a search phase in which such similarities are identified. Typically, local similarities to sequence database entries are searched for with alignment

programs such as BLAST (Camacho et al., 2009). Searches for gene family or protein domain motifs in the query sequence can be performed with a reference collection of profile Hidden Markov Models (HMMs). HMMER 3.0, released in 2010, has a 100-fold increase in speed compared to prior versions, with runtimes being competitive to blastp (Finn, Clements & Eddy, 2011). Screening a large metagenome sample with a collection of profile HMMs for marker genes is computationally much less demanding than a full search for similar regions in large sequence collections (Finn, Clements & Eddy, 2011). This is because the number of entries to be searched against is typically several orders of magnitude lower. HMMs are popular in combination with phylogenetic placement approaches, as the required multiple alignment of a query sequence to the homologs can be directly deduced from the state path of the sequence through the HMM and the multiple alignment used in its construction. However, known marker genes or protein families from reference collections such as PFAM only cover a small part of the genes found across diverse environments. Therefore, most HMM-based approaches (Stark et al., 2010; Gerlach & Stoye, 2011; Wu & Scott, 2012) may be seen as phylotypers of metagenome samples, rather than binning methods, as they indicate the taxonomic composition of the sample based on placement of a fraction of the fragments, rather than assigning the entire sample. Searching for similar sequences in large sequence collections results in a higher fragment coverage with hits than when profile HMMs are used. Analysis of a metagenome sequence sample therefore comes with high computational costs, beyond what a typical desktop computer is capable of. When using a similarity search, one is therefore confronted with the question of which reference sequences to compare to. The choice depends on the available time and computational resources. Databases that are often searched are NCBI RefSeq, a non-redundant nucleotide and protein collection for medical, functional and diversity studies; NCBI whole genomes; NCBI nt, a large nucleotide collection; and NCBI nr, a large non-redundant protein collection (Sayers et al., 2009). Software such as MEGAN (Huson et al., 2011) allows the output of BLAST to be interpreted for the taxonomic and functional characterization of metagenome samples based on sequence similarity. If sequenced genomes of related species to the sampled taxa exist, recruitment analysis has been used (Qin et al., 2010). Here, each read is compared to a set of

genome sequences and 'recruited' to the most similar genome, allowing the identification of reads of the prevalent species that are closely related to a sequenced reference collection, if performed with stringent alignment cut-offs (Xie et al., 2010).

C.6.1 Case study 1: Recruitment analysis

In (Xie et al., 2010), Illumina and Roche/454 sequencing were jointly used to generate 860 Mb of non-human sequence data from a microbial community of human dental plaque. All obtained reads were aligned against 50 available reference genomes for human oral microbes from the Human Microbiome Project using Mummer, resulting in recruitment of 4% of all reads with more than 97% sequence identity to one of the reference genomes. This indicates that most of the sampled microbes originate from species that are too distantly related to the sequenced reference collection for similarity-based recruitment.

C.7 Taxonomic binning based on sequence composition

The composition-based approach to taxonomic binning is to utilize the taxonomic signal contained in fragment-wide GC content, codon usage or the use of short oligomers (kmers), typically 4–6 bp long. The observation that such properties tend to vary more across the genomes of different species than within a given one gave rise to the term genome signatures (Karlin & Burge, 1995; Deschavanne et al., 1999). Such signatures can also be inferred for higher-level clades, allowing their use for taxonomic fragment assignment across various ranks (McHardy et al., 2007).

Taxonomic binning based on sequence composition can be performed with supervised or unsupervised methods. The choice of which to use depends on the availability of suitable reference data. Unsupervised methods group fragments with similar composition profiles into clusters, corresponding to individual taxo-

nomic bins. Inference of the taxonomic label for a bin can be performed based on taxonomic assignment of marker genes found in the fragments of a bin. To infer the clustering of fragments, existing methods use, for example, a graph-cut algorithm or variations of a self-organizing map algorithm (Chatterji et al., 2008; Weber et al., 2011). A sample can also be binned with supervised methods, which assign fragments to clades using a model trained with available reference sequences. Supervised methods tend to have higher accuracy than unsupervised methods for taxonomic assignment and are more easily applied to complex microbial mixtures with skewed organism abundances. However, they require sufficient amounts of reference sequences to be identified for the sample populations or higher-level clades which are to be included in the model. In practice, therefore, each approach has its own appeal and both are being applied. Methods used for supervised classification are, for example, (structural) Support Vector Machines (SVMs) (McHardy et al., 2007), the naïve Bayes classifier (Rosen, Reichenberger & Rosenfeld, 2011), a k-nearest neighbor classifier (Diaz et al., 2009) and Interpolated Markov Models (Brady & Salzberg, 2009). As composition-based signatures are a global attribute of sequences, no entire reference genomes are required, but only sufficient amounts of sequences for inference of a composition-based signature. For SVM-based classification, this has been found to be around 100 kb per clade (Patil et al., 2011). Reference sequences can be identified among publicly available genomes or by taxonomic assignment of conserved marker-genes of the sample contigs, which allows the respective contigs to be used as training material. If necessary, fosmids carrying marker genes can be sequenced to generate training material for interesting sample populations or higher-level clades (Warnecke et al., 2007; Pope et al., 2010, 2011).

C.7.1 Case study 2: Taxonomic binning by composition-based taxonomic assignment

In (Pope et al., 2010), a microbial gut community from the Australian Tammar wallaby was studied by Sanger and 454 sequencing of metagenome plasmid and fosmid libraries. This microbial community is involved in the breakdown of plant

biomass consumed by the host animal. Using 16S rRNA analysis, 236 distinct phylotypes were observed. Of the 16S rRNA sequences, 9% originated from a novel species, Wallaby group 1 (WG-1), in the family of Succinivibrionaceae. PhyloPythia, a composition-based taxonomic classifier, was used to train a model including the WG-1 and other relevant clades for species present in the community. Composition-based taxonomic assignment of the metagenome sample recovered a 2 Mb draft genome for WG-1. Metabolic reconstruction based on the draft genome allowed the cultivation requirements for WG-1 to be deduced, leading to isolation, characterization and a draft genome sequence for the previously unknown species. It also resulted in the finding that WG-1 contributes to the low-methane emission phenotype of plant biomass degradation in the Tammar wallaby. The draft genome sequences from the isolate culture showed 98.9% sequence identity to the WG-1 metagenome bin, and 90% of shared reads and assemblies, indicating accurate reconstruction of the draft genome from the metagenome sample by composition-based taxonomic binning.

C.8 Hybrid methods

Several methods combine different types of information to improve predictive accuracy (Brady & Salzberg, 2009; Hess et al., 2011; Huson et al., 2011; Iverson et al., 2012). For instance, read coverage is combined with an analysis of kmer frequencies in clustering of fragments (Tyson et al., 2004; Hess et al., 2011). Searches for similar sequences and analysis of linkage information from an assembly are also combined with composition-based taxonomic assignment, if the computational burden can be borne. This has particular advantages for short fragment analysis. Kmer signatures for fragments below 1 kb in length, particularly those of individual reads, are noisy, even more so than taxonomic conservation of sequence similarities (Patil et al., 2011).

C.8.1 Case study 3: Taxonomic binning based on clustering by sequence composition and read coverage

In one of the most in-depth metagenome studies of a particular environment undertaken so far, 286 Gb of paired-end Illumina sequence reads were generated from a sample of the plant-fiber adherent microbiome from a cow rumen (Hess et al., 2011). Rarefaction analysis of 16S rRNA indicated the presence of ~1000 distinct OTUs. Clustering of assembled contigs by agglomerative hierarchical clustering, based on tetramer frequencies and read coverage, resulted in the formation of 466 taxonomic bins. Fifteen of these were estimated to represent largely complete genomes (between 60% and 92%), based on their association with fully sequenced genomes from their respective clades. This estimate was based on the presence of a minimal set of core genes found in all sequenced genomes from the respective phylogenetic order.

C.8.2 Case study 4: Taxonomic binning based on assembly information and sequence composition (Iverson et al., 2012).

SOLID sequencing of two marine samples generated 58.5 Gb of mate-paired reads of 50 bps in length. The number of phylotypes observed with16S rRNA analysis was not specified in detail; however, family-level taxonomic groups were observed with abundances of less than 10%. From the metagenome data, 300 Mb of contigs were assembled. Scaffolds – linked sets of contigs assumed to originate from one genome – were generated by splitting the assembly graph, which links contigs based on mate-pair information, according to mate-pair linkage scores, read coverage and tetranucleotide usage. Scaffold clustering by tetranucleotide usage generated 14 partial genome reconstructions from the two samples, for populations ranging in abundance from 4% to 10 % each in one of the samples. Reassembly of 11 mate-pair connected scaffolds that are binned together based on similar tetranucleotide statistics and manual gap closure allowed the recovery of a closed circular 2 Mb

genome from an uncultured group, the marine group II Euryarchaeota.

C.9 Advantages and disadvantages of different binning approaches

Which binning methodology to use depends on multiple factors, such as the complexity of the analyzed microbial community, available reference sequences and computing ressources. For taxonomic assignment of arbitrary sequence fragments to a particular species based on sequence similarity, completely sequenced reference genomes of closely related taxa are ideally required, which are often not available. If no reference data exists for the species of the metagenome sample, homology-based taxonomic assignment to higher-level clades is more accurate than composition-based taxonomic assignment for short fragments of 1 kb or less (Patil et al., 2011). This length corresponds to individual reads with most sequencing technologies. The assignment of individual reads in general is, however, notably less accurate than assignment of longer fragments.

The runtime of sequence similarity searches increases proportional to the product of the metagenome sample size (number and length of contigs) and the size of the reference sequence collection. This makes it a computationally very demanding task for next-generation sequencing data sets. The required computing resources are not available in many experimental laboratories. If researchers are willing to submit their data to external facilities, data can be processed by web servers such as MG-RAST, IMG-M or CAMERA, which offer their computational resources to the community.

The choice of whether to cluster or classify based on sequence composition depends on availability of some reference data to train a composition-based classifier. Classification is likely to be more accurate than clustering in taxonomic assignment. However, if no reference data is avaible, clustering will allow resolution of taxonomic bins which otherweise would go undetected. If multiple types of information are included into the binning process, like it is done in hybrid approaches, this is likely to increase the overall amount and accuracy of assignments.

Composition-based taxonomic assignment requires less reference sequences than homology-based assignment. This is because sequence composition is a globally conserved property, while sequence similarity depends on local sequence conservation between a query and target. Training times of a composition-based taxonomic classifier depend on the method used, but it requires typically considerably less time than searching a reference sequence collection. Once a composition-based model for taxonomic classification has been trained, execution times for classification again typically scale linearly with the metagenome sample size and are independent of a reference sequence collection. For composition-based clustering, no training phase is needed. The runtime of clustering typically scales at least quadratically with the sample size, as it often involves pairwise comparisons.

C.10 Future directions

The recent developments in sequencing technologies have considerably pushed the boundaries in terms of what can be learned from metagenome sequence samples. The high sequencing depth of microbial communities, in combination with the application of sophisticated algorithms, has allowed the retrieval of near-complete draft genomes from the metagenomes of many microbial communities, including highly complex ones, such as those found in soil (Mackelprang et al., 2011). However, the size and heterogeneity of the different data types produced by the various novel techniques have created new challenges, which remain to be addressed. A prominent one is how to further reduce the computational requirements of searching for local similarities between giga- or even terabase-sized sequence samples and equivalently large reference sequence collections. Secondly, it remains to be explored how taxonomic assignment accuracy can be further improved for the vast majority of microbial community members that are only distantly related to sequenced isolate genomes. Due to the value of available sequences from related taxa for the taxonomic binning of a particular sample, efforts such as GEBA might help in this regard (Wu et al., 2009). The GEBA project aims to construct a "Genomic Encyclopedia for Bacteria and Archaea" by strategic sequencing of microbial genomes from all major and minor taxonomic groups. As the cost of

sequencing has decreased, partial genome reconstruction by single-cell genome sequencing is an attractive option for obtaining reference sequences for taxonomic binning and draft genome reconstruction (Woyke et al., 2010) from metagenomes. Here, an individual cell from a microbial population within a community is isolated using techniques such as optical tweezers, fluorescence assisted cell sorting and others, and is then lysed and its genome sequence amplified with multiple displacement amplification prior to random shotgun sequencing.

Advances in single-molecule sequencing technologies now allow longer reads to be generated than what was possible using traditional Sanger sequencing. Even though this promises to resolve several issues associated with short read analysis, such as high error rates in binning, assembly and functional annotation, the larger sequencing error of some of these technologies, currently estimated to be around 15%, presents a different substantial hurdle. Therefore, assessing technology-specific errors and developing technology-specific denoising procedures, such as have been developed for 454 amplicon data (Quince et al., 2009), will be prerequisite to leveraging the value of these techniques for metagenome research. An interesting research direction is to investigate whether composition-based binning is applicable for the analysis of samples with both microbial and viral content. Composition-based taxonomic binning has been successfully applied for the analysis of viral metagenome samples, however, bacteriophage codon usage to some extent reflects properties of the host (Pride & Schoenfeld, 2008; Lucks et al., 2008). Therefore, classification accuracy and level of taxonomic resolution attainable for viral taxa will have to be investigated in more detail.

C.11 Summary of key points

NGS technologies generate massive amounts of sequencing data allowing the in-depth analysis of microbial communities. Taxonomic binning has allowed draft genomes of microbial species from many environments to be reconstructed, and the cultivation requirements of a novel uncultured species to be deduced. To further advance draft genome reconstruction from metagenome samples, the existing techniques could be further refined by integrating multiple sources of infor-

mation and by appropriately denoising the data under consideration to remove technology-specific sequencing errors.

C.12 Acknowledgements

We thank Chris Quince and Alex Scyrba for providing comments.

C.13 Funding

This work was supported by the German Max Planck society and Heinrich-Heine University Düsseldorf.

C.14 References

Amann RI., Ludwig W., Schleifer KH. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews* 59:143–69.

Berger SA., Krompass D., Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic biology* 60:291–302. DOI: 10.1093/sysbio/syr010.

Brady A., Salzberg SL. 2009. Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature methods* 6:673–6. DOI: 10.1038/nmeth.1358.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009. BLAST+: Architecture and applications. *BMC bioinformatics* 10:421. DOI: 10.1186/1471-2105-10-421.

Chatterji S., Yamazaki I., Bai Z., Eisen JA. 2008. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In:

Annual International Conference on Research in Computational Molecular Biology. Springer, 17–28.

Cole JR., Wang Q., Cardenas E., Fish J., Chai B., Farris RJ., Kulam-Syed-Mohideen AS., McGarrell DM., Marsh T., Garrity GM., Tiedje JM. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* 37:D141–D145. DOI: 10.1093/nar/gkn879.

DeSantis TZ., Hugenholtz P., Larsen N., Rojas M., Brodie EL., Keller K., Huber T., Dalevi D., Hu P., Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* 72:5069–72. DOI: 10.1128/AEM.03006-05.

Deschavanne PJ., Giron a., Vilain J., Fagot G., Fertil B. 1999. Genomic signature: Characterization and classification of species assessed by chaos game representation of sequences. *Molecular biology and evolution* 16:1391–9.

Diaz NN., Krause L., Goesmann A., Niehaus K., Nattkemper TW. 2009. TACOA: Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC bioinformatics* 10:56. DOI: 10.1186/1471-2105-10-56.

Finn RD., Clements J., Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic acids research* 39 Suppl 2:W29–37. DOI: 10.1093/nar/gkr367.

Fuhrman J a. 2009. Microbial community structure and its functional implications. *Nature* 459:193–9. DOI: 10.1038/nature08058.

Gerlach W., Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic acids research*:1–11. DOI: 10.1093/nar/gkr225.

Hess M., Sczyrba A., Egan R., Kim T-W., Chokhawala H., Schroth G., Luo S., Clark DS., Chen F., Zhang T., Mackie RI., Pennacchio L a., Tringe SG., Visel A., Woyke T., Wang Z., Rubin EM. 2011.

Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, N.Y.)* 331:463–7. DOI: 10.1126/science.1200387.

Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome biology* 3:REVIEWS0003.

Huse SM., Dethlefsen L., Huber JA., Welch DM., Relman DA., Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS genetics* 4:e1000255. DOI: 10.1371/journal.pgen.1000255.

Huson DH., Mitra S., Ruscheweyh H-J., Weber N., Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome research* 21:1552–60. DOI: 10.1101/gr.120618.111.

Iverson V., Morris RM., Frazar CD., Berthiaume CT., Morales RL., Armbrust EV. 2012. Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science* 335:587–590. DOI: 10.1126/science.1212665.

Jeewon R., Hyde KD. 2007. Detection and diversity of fungi from environmental samples: Traditional versus molecular approaches. *Advanced techniques in soil microbiology* 11:1–15.

Karlin S., Burge C. 1995. Dinucleotide relative abundance extremes: A genomic signature. *Trends in genetics* 11:283–90.

Koren S., Treangen TJ., Pop M. 2011. Bambus 2: Scaffolding metagenomes. *Bioinformatics* 27:2964–2971. DOI: 10.1093/bioinformatics/btr520.

Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P. 2008. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews* 72:557–578. DOI: 10.1128/MMBR.00009-08.

Laserson J., Jojic V., Koller D. 2011. Genovo: De novo assembly for metagenomes. *Journal of computational biology* 18:429–43. DOI: 10.1089/cmb.2010.0244.

Lucks JB., Nelson DR., Kudla GR., Plotkin JB. 2008. Genome landscapes and bacteriophage codon usage. *PLoS computational biology* 4:e1000001. DOI:

10.1371/journal.pcbi.1000001.

Mackelprang R., Waldrop MP., DeAngelis KM., David MM., Chavarria KL., Blazewicz SJ., Rubin EM., Jansson JK. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480:368–371. DOI: 10.1038/nature10576.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* 24:133–41. DOI: 10.1016/j.tig.2007.12.007.

Markowitz VM., Chen I-MA., Chu K., Szeto E., Palaniappan K., Grechkin Y., Ratner A., Jacob B., Pati A., Huntemann M., Liolios K., Pagani I., Anderson I., Mavromatis K., Ivanova NN., Kyrpides NC. 2012. IMG/M: The integrated metagenome data management and comparative analysis system. *Nucleic Acids Research* 40:D123–D129. DOI: 10.1093/nar/gkr975.

Matsen FA., Kodner RB., Armbrust EV. 2010. Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics* 11:538. DOI: 10.1186/1471-2105-11-538.

McHardy AC., Martín HG., Tsirigos A., Hugenholtz P., Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nature methods* 4:63–72. DOI: 10.1038/nmeth976.

Metzker ML. 2009. Sequencing technologies — the next generation. *Nature reviews genetics* 11:31–46. DOI: 10.1038/nrg2626.

Meyer F., Paarmann D., D'Souza M., Olson R., Glass EM., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R a. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 9:386. DOI: 10.1186/1471-2105-9-386.

Miller JR., Koren S., Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–27. DOI:

10.1016/j.ygeno.2010.03.001.

Mirarab S., Nguyen N., Warnow T. 2012. SEPP: SATé-Enabled Phylogenetic Placement. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*:247–58.

Nelson KE., Weinstock GM., Highlander SK., Worley KC., Creasy HH., Wortman JR., Rusch DB., Mitreva M., Sodergren E., Chinwalla AT., Feldgarden M., Gevers D., Haas BJ., Madupu R., Ward DV., Birren BW., Gibbs R a., Methe B., Petrosino JF., Strausberg RL., Sutton GG., White OR., Wilson RK., Durkin S., Giglio MG., Gujja S., Howarth C., Kodira CD., Kyrpides N., Mehta T., Muzny DM., Pearson M., Pepin K., Pati A., Qin X., Yandava C., Zeng Q., Zhang L., Berlin AM., Chen L., Hepburn T a., Johnson J., McCorrison J., Miller J., Minx P., Nusbaum C., Russ C., Sykes SM., Tomlinson CM., Young S., Warren WC., Badger J., Crabtree J., Markowitz VM., Orvis J., Cree A., Ferriera S., Fulton LL., Fulton RS., Gillis M., Hemphill LD., Joshi V., Kovar C., Torralba M., Wetterstrand K a., Abouellleil A., Wollam AM., Buhay CJ., Ding Y., Dugan S., FitzGerald MG., Holder M., Hostetler J., Clifton SW., Allen-Vercoe E., Earl AM., Farmer CN., Liolios K., Surette MG., Xu Q., Pohl C., Wilczek-Boney K., Zhu D. 2010. A catalog of reference genomes from the human microbiome. *Science (New York, N.Y.)* 328:994–9. DOI: 10.1126/science.1183605.

Patil KR., Haider P., Pope PB., Turnbaugh PJ., Morrison M., Scheffer T., McHardy AC. 2011. Taxonomic metagenome sequence assignment with structured output models. *Nature Methods* 8:191–192. DOI: 10.1038/nmeth0311-191.

Pell J., Hintze A., Canino-Koning R., Howe A., Tiedje J., Brown C. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Arxiv preprint arXiv:1112.4193* I:1–11. DOI:

10.1073/pnas.1121464109.

Peng Y., Leung HCM., Yiu SM., Chin FYL. 2011. Meta-IDBA: A de novo assembler for metagenomic data. *Bioinformatics (Oxford, England)* 27:i94–i101. DOI: 10.1093/bioinformatics/btr216.

Pope PB., Denman SE., Jones M., Tringe SG., Barry K., Malfatti S a., McHardy a C., Cheng J-F., Hugenholtz P., McSweeney CS., Morrison M. 2010. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proceedings of the National Academy of Sciences of the United States of America* 107:14793–8. DOI: 10.1073/pnas.1005297107.

Pope PB., Smith W., Denman SE., Tringe SG., Barry K., Hugenholtz P., McSweeney CS., McHardy a C., Morrison M. 2011. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science (New York, N.Y.)* 333:646–8. DOI: 10.1126/science.1205760.

Pride DT., Schoenfeld T. 2008. Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC genomics* 9:420. DOI: 10.1186/1471-2164-9-420.

Pruesse E., Quast C., Knittel K., Fuchs BM., Ludwig W., Peplies J., Glöckner FO. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research* 35:7188–96. DOI: 10.1093/nar/gkm864.

Qin J., Li R., Raes J., Arumugam M., Burgdorf KS., Manichanh C., Nielsen T., Pons N., Levenez F., Yamada T., Mende DR., Li J., Xu J., Li S., Li D., Cao J., Wang B., Liang H., Zheng H., Xie Y., Tap J., Lepage P., Bertalan M., Batto J-M., Hansen T., Le Paslier D., Linneberg A., Nielsen HB., Pelletier E., Renault P., Sicheritz-Ponten T., Turner K., Zhu H., Yu C., Li S., Jian M., Zhou Y., Li Y., Zhang X., Li S., Qin N., Yang H., Wang J., Brunak S., Doré J., Guarner F., Kristiansen K., Pedersen O., Parkhill J., Weissenbach J., Bork P., Ehrlich SD., Wang J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59–65. DOI:

10.1038/nature08821.

Quince C., Lanzén A., Curtis TP., Davenport RJ., Hall N., Head IM., Read LF., Sloan WT. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods* 6:639–41. DOI: 10.1038/nmeth.1361.

Rosen GL., Reichenberger ER., Rosenfeld AM. 2011. NBC: The Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics (Oxford, England)* 27:127–9. DOI: 10.1093/bioinformatics/btq619.

Sayers EW., Barrett T., Benson D., Bryant SH., Canese K., Chetvernin V., Church DM., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer LY., Helmberg W., Kapustin Y., Landsman D., Lipman DJ., Madden TL., Maglott DR., Miller V., Mizrachi I., Ostell J., Pruitt KD., Schuler GD., Sequeira E., Sherry ST., Shumway M., Sirotkin K., Souvorov A., Starchenko G., Tatusova T a., Wagner L., Yaschenko E., Ye J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 37:D5–15. DOI: 10.1093/nar/gkn741.

Schadt EE., Turner S., Kasarskis A. 2010. A window into third-generation sequencing. *Human molecular genetics* 19:R227–40. DOI: 10.1093/hmg/ddq416.

Simon C., Daniel R. 2011. Metagenomic Analyses: Past and Future Trends. *Applied and Environmental Microbiology* 77:1153–1161. DOI: 10.1128/AEM.02345-10.

Stark M., Berger S., Stamatakis A., von Mering C. 2010. MLTreeMap - accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC genomics* 11:461. DOI: 10.1186/1471-2164-11-461.

Su C-H., Hsu M-T., Wang T-Y., Chiang S., Cheng J-H., Weng FC., Kao C-Y., Wang D., Tsai H-K. 2011. MetaABC–an integrated metagenomics platform for data adjustment, binning and clustering. *Bioinformatics*

(Oxford, England) 27:2298–9. DOI: 10.1093/bioinformatics/btr376.

Suen G., Scott JJ., Aylward FO., Adams SM., Tringe SG., Pinto-Tomás A a., Foster CE., Pauly M., Weimer PJ., Barry KW., Goodwin L a., Bouffard P., Li L., Osterberger J., Harkins TT., Slater SC., Donohue TJ., Currie CR. 2010. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS genetics* 6:e1001129. DOI: 10.1371/journal.pgen.1001129.

Sun S., Chen J., Li W., Altintas I., Lin A., Peltier S., Stocks K., Allen EE., Ellisman M., Grethe J., Wooley J. 2011. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: The CAMERA resource. *Nucleic Acids Research* 39:D546–D551. DOI: 10.1093/nar/gkq1102.

Thompson JF., Milos PM. 2011. The properties and applications of single-molecule DNA sequencing. *Genome biology* 12:217. DOI: 10.1186/gb-2011-12-2-217.

Tringe SG., von Mering C., Kobayashi A., a Salamov A., Chen K., Chang HW., Podar M., Short JM., Mathur EJ., Detter JC., Bork P., Hugenholtz P., Rubin EM. 2005. Comparative metagenomics of microbial communities. *Science (New York, N.Y.)* 308:554–7. DOI: 10.1126/science.1107851.

Turnbaugh PJ., Quince C., Faith JJ., McHardy AC., Yatsunenko T., Niazi F., Affourtit J., Egholm M., Henrissat B., Knight R., Gordon JI. 2010. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences of the United States of America* 107:7503–8. DOI: 10.1073/pnas.1002355107.

Tyson GW., Chapman J., Hugenholtz P., Allen EE., Ram RJ., Richardson PM., Solovyev VV., Rubin EM., Rokhsar DS., Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial

genomes from the environment. *Nature* 428:37–43. DOI: 10.1038/nature02340.

Venter JC., Remington K., Heidelberg JF., Halpern AL., Rusch D., Eisen J a., Wu D., Paulsen I., Nelson KE., Nelson W., Fouts DE., Levy S., Knap AH., Lomas MW., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y-H., Smith HO. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)* 304:66–74. DOI: 10.1126/science.1093857.

Warnecke F., Luginbühl P., Ivanova N., Ghassemian M., Richardson TH., Stege JT., Cayouette M., McHardy AC., Djordjevic G., Aboushadi N., Sorek R., Tringe SG., Podar M., Martin HG., Kunin V., Dalevi D., Madejska J., Kirton E., Platt D., Szeto E., Salamov A., Barry K., Mikhailova N., Kyrpides NC., Matson EG., Ottesen E a., Zhang X., Hernández M., Murillo C., Acosta LG., Rigoutsos I., Tamayo G., Green BD., Chang C., Rubin EM., Mathur EJ., Robertson DE., Hugenholtz P., Leadbetter JR. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450:560–5. DOI: 10.1038/nature06269.

Weber M., Teeling H., Huang S., Waldmann J., Kassabgy M., Fuchs BM., Klindworth A., Klockow C., Wichels A., Gerdts G., Amann R., Glöckner FO. 2011. Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *The ISME journal* 5:918–28. DOI: 10.1038/ismej.2010.180.

Wilkening J., Wilke A., Desai N., Meyer F. 2009. Using clouds for metagenomics: A case study. In: *2009 IEEE International Conference on Cluster Computing and Workshops*. IEEE, 1–6. DOI: 10.1109/CLUSTR.2009.5289187.

Woyke T., Teeling H., Ivanova NN., Huntemann M., Richter M., Gloeckner FO., Boffelli D., Anderson IJ., Barry KW., Shapiro HJ., Szeto E., Kyrpides NC., Mussmann M., Amann R., Bergin C., Ruehland C., Rubin EM., Dubilier N. 2006. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443:950–5. DOI:

10.1038/nature05192.

Woyke T., Tighe D., Mavromatis K., Clum A., Copeland A., Schackwitz W., Lapidus A., Wu D., McCutcheon JP., McDonald BR., Moran N a., Bristow J., Cheng J-F. 2010. One bacterial cell, one complete genome. *PloS one* 5:e10314. DOI: 10.1371/journal.pone.0010314.

Wu M., Scott AJ. 2012. Phylogenomic Analysis of Bacterial and Archaeal Sequences with AMPHORA2. *Bioinformatics (Oxford, England)*:1–2. DOI: 10.1093/bioinformatics/bts079.

Wu D., Hugenholtz P., Mavromatis K., Pukall R., Dalin E., Ivanova NN., Kunin V., Goodwin L., Wu M., Tindall BJ., Hooper SD., Pati A., Lykidis A., Spring S., Anderson IJ., D'haeseleer P., Zemla A., Singer M., Lapidus A., Nolan M., Copeland A., Han C., Chen F., Cheng J-F., Lucas S., Kerfeld C., Lang E., Gronow S., Chain P., Bruce D., Rubin EM., Kyrpides NC., Klenk H-P., Eisen J a. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462:1056–60. DOI: 10.1038/nature08656.

Xie G., Chain PSG., Lo C-C., Liu K-L., Gans J., Merritt J., Qi F. 2010. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Molecular oral microbiology* 25:391–405. DOI: 10.1111/j.2041-1014.2010.00587.x.