ABSTRACT

Title of dissertation: COMPUTATIONAL METAGENOMICS: NETWORK, CLASSIFICATION AND ASSEMBLY

Bo Liu, Doctor of Philosophy, 2012

Dissertation directed by: Professor Mihai Pop
Department of Computer Science

Due to the rapid advance of DNA sequencing technologies in recent 10 years, large amounts of short DNA reads can be obtained quickly and cheaply. For example, a single Illumina HiSeq machine can produce several terabytes of data sets within a week. Metagenomics is a new scientific field that involves the analysis of genomic DNA sequences obtained directly from the environment, enabling studies of novel microbial systems. Metagenomics was made possible from high-throughput sequencing technologies. The analysis of the resulting data requires sophisticated computational analyses and data mining. In clinical settings, a fundamental goal of metagenomics is to help people diagnose and cure disease in clinical settings. One major bottleneck so far is how to analyze the huge noisy data sets quickly and precisely. My PhD research focuses on developing algorithms and tools to tackle these challenging and interesting computational problems.

From the functional perspective, a metagenomic sample can be represented as a weighted metabolic network, in which the nodes are molecules, edges are enzymes encoded by genes, and the weights can be considered as the number of organisms

providing the functions. One goal of functional comparison between metagenomic samples is to find differentially abundant metabolic subnetworks between two groups under comparison. We have developed a statistical network analysis tool - MetaPath, which uses a greedy search algorithm to find maximum weight subnetwork and a nonparametric permutation test to measure the statistical significance. Unlike previous approaches, MetaPath explicitly searches for significant subnetwork in the global network, enabling us to detect signatures at a finer level. In addition, we developed statistical methods that take into account the topology of the network when testing the significance of the subnetworks.

Another computational problem involves classifying anonymous DNA sequences obtained from metagenomic samples. There are several challenges here: (1) The classification labels follow a hierarchical tree structure, in which the leaves are most specific, and the internal nodes are more general. How can we classify novel sequences that do not belong to leaf categories (species) but belong to internal groups (e.g., phylum)? (2) For each classification how can we compute a confidence score, such that the users have a tradeoff between sensitivity and specificity? (3) How can we analyze billions of data items quickly? We have developed a novel hierarchical classifier (MetaPhyler) for the classification of anonymous DNA reads. Through simulation, MetaPhyler models the distribution of pairwise similarities within different hierarchical groups with nonparametric density estimation. The confidence score is computed by the ratio of likelihood function. For a query DNA sequence with arbitrary length, its similarity can be calculated through linear approximation. Through benchmark comparison, we have shown that MetaPhyler is significantly

faster and more accurate than previous tools.

DNA sequencing machines can only produce very short strings (e.g., 100bp) relative to the size of a genome (e.g., a typical bacterial genome is 5Mbp). One of the most challenging computational tasks is the assembly of millions of short reads into longer contigs, which are used as the basis of subsequent computational analyses. In this project, we have developed a comparative metagenomic assembler (MetaCompass), which utilizes the genomes that have already been sequenced previously, and produces long contigs through read mapping (alignment) and assembly. Given the availability of thousands of existing bacteria genomes, for a particular sample, MetaCompass first chooses a best subset as reference based on the taxonomic composition. Then, the reads are aligned against these genomes using MUMmer-map or Bowtie2. Afterwards, we use a greedy algorithm of the minimum set-covering problem to build long contigs, and the consensus sequences are computed by the majority rule. We also propose an iterative approach to improve the performance. Finally, MetaCompass has been successfully evaluated and tested on over 20 terabytes of metagenomic data sets generated from the Human Microbiome Project.

In addition, to facilitate the identification and characterization of antibiotic resistance genes, we have created Antibiotic Resistance Genes Database (ARDB), which provides a centralized compendium of information on antibiotic resistance. Furthermore, we have applied our tools to the analysis of a novel oral microbiome data set, and have discovered interesting functional mechanisms and ecological changes underlying the transition from health to periodontal disease of human mouth at a system level.

COMPUTATIONAL METAGENOMICS: NETWORK, CLASSIFICATION AND ASSEMBLY

by

Bo Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Professor Mihai Pop, Chair/Advisor
Professor Hector Corrado Bravo
Professor Charles Delwiche
Professor Sridhar Hannenhalli
Professor Carl Kingsford
Professor O Colin Stine

Preface

The algorithms, tools and results in this dissertation have either been published in peer-reviewed journals or are currently under preparation for submission. At the time of this writing, Chapters 2, 3, 4 and 5 have already been published. Chapter 6 and an improved version of Chapter 2 are under preparation for submission. I am indebted to my co-authors on these projects - their dedication and knowledge improved my research significantly from both computer science and biology perspectives. This thesis reflects not only my individual research, but also the contributions from my co-authors and collaborators.

In the following I list the papers that constitute my dissertation, and some other papers I have contributed to.

Chapter 2:

- Bo Liu, Theodore Gibbons, Mohammadreza Ghodsi, Mihai Pop. MetaPhyler: Taxonomic profiling for metagenomic sequences. Proceedings of 2010 IEEE Bioinformatics and Biomedicine:95-100.

- Bo Liu, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, Mihai Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 2011, 12(Suppl 2):S4

- Bo Liu, Mihai Pop. Training and classification of metagenomic sequences with confidence score. In preparation.

Chapter 3:

- Bo Liu, Mihai Pop. Identifying Differentially Abundant Metabolic Pathways in

Metagenomic Datasets. Lect Notes Comput Sci 2010, 6053: 101-112

- Bo Liu, Mihai Pop. MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proceedings 2011, 5(Suppl 2):S9

Chapter 4:

- Bo Liu, Mihai Pop. ARDB - Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D443-7.

Chapter 5:

- Bo Liu*, Lina Faller*, Niels Klitgord*, Varun Mazumdar*, Mohammad Ghodsi, Daniel D. Sommer, Theodore R. Gibbons, Todd J. Treangen, Yi-Chien Chang, Shan Li, O. Colin Stine, Hatice Hasturk, Simon Kasif, Daniel Segre, Mihai Pop, Salomon Amar. PLoS One 2012, 7(6):e37919.

Chapter 6:

- Bo Liu, Mihai Pop. MetaCompass: comparative assembly of metagenomic sequences. In preparation.

A list of papers I have contributed to (being a co-author):

- Mohammadreza Ghodsi, **Bo Liu**, Mihai Pop. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics 2011,12:271
- David R. Kelley, **Bo Liu**, Arthur L. Delcher, Mihai Pop, Steven L. Salzberg. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Research, 2011, doi: 10.1093/nar/gkr1067
- Treangen T*, Koren S*, Sommer D, Darling A, Astrovskaya I, **Liu B** and Pop M. metAMOS: A modular and open source metagenomic assembly and analysis

pipeline. In submission.

- The FlowCAP Consortium. Critical assessment of computational flow cytometry analysis techniques: results of FlowCAP. In submission.

- The HMP Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012, 486(7402):207-214.

- The HMP Consortium. A framework for human microbiome research. Nature 2012, 486(7402):215-221.

* Co-first authors

Acknowledgments

I am indebted to all the people, without whose help I could not have written this proposal, and because of whom my graduate research and life have been one that I will cherish forever.

First and foremost, I thank my advisor Mihai Pop for his guidance, unwavering support, and insightful collaboration on all the projects described in this dissertation. At every difficult moment during research in my past five years, he has always been supportive. I would also like to thank all my mentors and coworkers at the Center for Bioinformatics and Computational Biology, one of the greatest bioinformatics research center, and Department of Computer Science. Particular thanks to Steven Salzberg, director of CBCB, and who taught me metagenomics course and has inspired me about research in many ways. Many thanks to Carl Kingsford who introduced me to network biology, and MetaPath is the result from his course. Thanks to Lise Getoor who enlightened me about statistical relational learning. I would like to thank Hector Bravo, who has inspired me on statistics and drawing beautiful scientific figures. I am also very grateful to Colin Stine, as I have learned lot of Microbiology and Epidemiology from him.

At the Center for Bioinformatics and Computational Biology, I am very lucky to have shared an office, and worked in a same lab with some wonderful people: Mohammad Ghodsi, Ted Gibbons, Chris Hill, Chengxi Ye, Joseph Paulson, Lee Mendelowitz, Irina Astrovskaya, Henry Lin, Sergey Koren, Todd Treangen, Niranjan Nagarajan and James White. They helped me a lot when I have difficulties during

my research projects, and made my graduate life incredibly enjoyable. In addition, I gratefully acknowledge the collaborators of the oral microbiome project: Salomon Amar, Daniel Segre, Simon Kasif and Colin Stine.

Importantly, I would like to thank my family and girlfriend; they have been encouraging and supportive about every major decision during my life. Lastly, many thanks to my entire dissertation committee, Mihai Pop, Hector Bravo, Charles Delwiche, Sridhar Hannehalli, Carl Kingsford and Colin Stine, for their support and help of this dissertation. It is impossible to remember all, I apologize to anyone who may have been omitted from the acknowledgements.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Metagenomics

Microorganisms comprise the majority of Earth's biological diversity, and they play essential functional roles in virtually all ecosystems. Although a tiny fraction of them can induce diseases, the vast majority of microbes - life forms too tiny to see - are actually essential to keeping us alive. For example, microbes help us digest food, and they make Earth livable by maintaining the atmospheric conditions. In particular, human-associated microbial communities play a fundamentally important role in health and disease [53], and they actually do a lot of good for the health of humans. In many natural environments, however, more than 99% of the microorganisms cannot be cultured by standard laboratory techniques [104]. Metagenomics is a new scientific field that involves the analysis of organismal DNA sequences obtained directly from an environmental sample, enabling studies of microorganisms that are not easily cultured in a laboratory. It offers a powerful lens for examining the microbial world that has the potential to revolutionize our understanding of the entire living world.

Metagenomic studies, pioneered in the early 2000s, have recently increased in number and scope due to the rapid advances of high-throughput sequencing technologies, e.g., GS FLX system from 454 Life Sciences and Genome Analyzer from

Illumina company, which permit large amounts of DNA to be sequenced quickly and cheaply (Figure 1.1 shows the sequencing cost per Mb of DNA sequences in recent years [110]). For example the MetaHit consortium has generated about 500 billion base pairs raw sequences from 124 human gut samples in its initial analysis [101], the Human Microbiome Project has produced more than 7 trillion base pair DNA reads [20, 21], and the newly initiated Earth Microbiome Project is planning to sequence 200,000 samples with 6 billion base pairs per sample, totaling 1200 trillion bp DNA (`http://www.earthmicrobiome.org/`). The growth of publicly available DNA sequence data over the last two decades has been exponential, with a doubling time of about 14 months. And the doubling time appears to be dramatically shortened due to next-generation sequencing and the addition of metagenomics [135]. There is a pressing need of efficient and powerful computational algorithms and tools that can deal with the huge amount of complex and noisy metagenomic sequences (Figure 1.2 shows the cost compositions of a sequencing project [110]). Motivated by these computational challenges, the goal of this dissertation is to develop accurate and fast computational algorithms and tools to extract and interpret meaningful biological information from the deluge of metagenomic sequences.

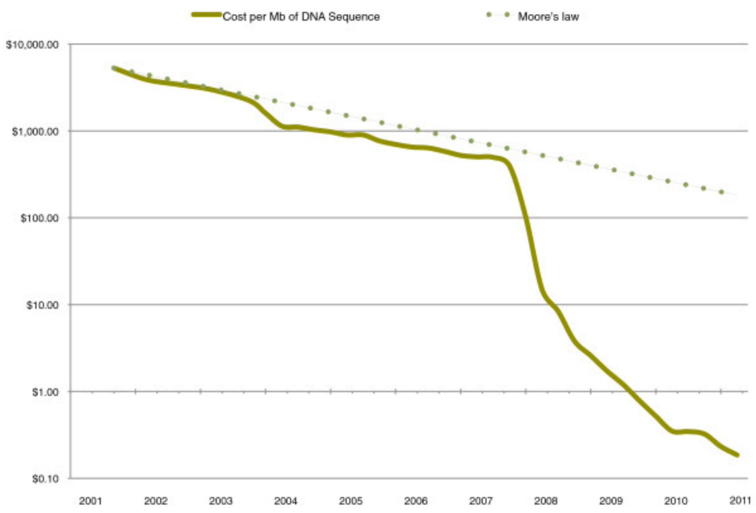


Figure 1.1: Cost of 1 Mb of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [110].

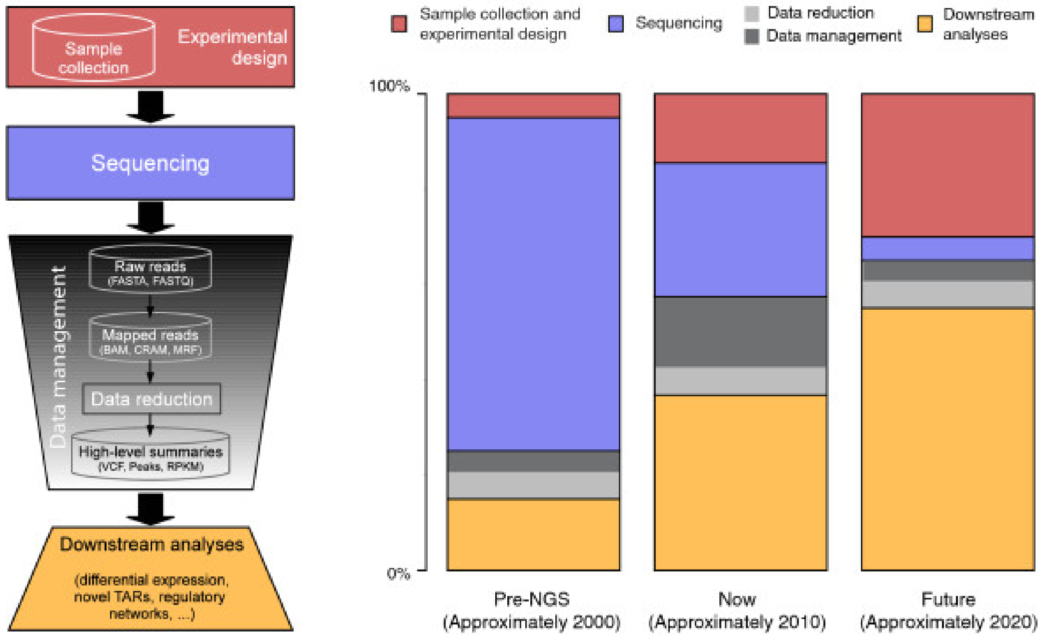


Figure 1.2: Contribution of different factors to the overall cost of a sequencing project across time. Adapted from [110].

1.2 Computational Problems in Metagenomics

Metagenomics introduces many interesting computational problems, e.g., gene prediction, sequence classification and clustering, genome assembly, statistical comparison, functional annotation, microbial interactions modeling, etc. Several review papers have been published summarizing and describing these challenges in detail [15, 45, 68, 102, 135]. In this section, I briefly describe some computational problems that are relevant to my research.

1.2.1 Genome assembly

The major challenge in metagenomic assembly arises from the heterogeneous nature of metagenomic data. Most samples contain an uneven representation of the member species. Furthermore, the genomes in a sample frequently belong to clusters of closely related strains whose genomes are mostly similar but differ due to small insertion, deletion and mutations. These characteristics of metagenomic data make it essentially impossible to construct a single genome for each species in a sample. In contrary, a lot of genomes will be under-represented and assembled into a highly fragmented form; while a group of closely related genomes will be assembled together into a single genome with polymorphisms in some locations [99].

In early metagenomic projects, the assembly was usually performed with conventional whole genome assemblers (e.g., Celera in [44]), which are designed for Sanger sequencing and are based on overlap-layout-consensus algorithms. Later on, the development of genome assembly algorithms has been boosted because of next-generation sequencing technologies, which can produce massive amount of short DNA reads. Several algorithms and tools have been developed specifically for this type of data using De Bruijn graph representation (e.g., Velvet [139], SOAPdenovo [79], ABySS [119], ALLPATHS [14]). We note that all of these assemblers are designed for assembling single genome but not metagenomes. These assemblers can be confused by two main issue: (i) uneven abundance of the genomes within a sample and (ii) polymorphisms between closely related genomes. For example, many assemblers use depth of coverage statistics to identify DNA repeats. In metagenomic data, however, due to uneven coverage, these assemblers would incorrectly label them as repeats, and avoid assembling these regions to prevent mistakes [99]. Usually, metagenomes are still assembled using these tools but with tuned parameters [101].

Recently, several metagenome assemblers have been developed (e.g., Genovo [74], Meta-IDBA [96], Bambus 2 [65], MAP [71], MetaVelvet [90]). Their performance, however, is mainly unknown in practice, since most of them have not been tested on large-scale metagenomic data sets.

1.2.2 Sequence classification

One of the primary goals of metagenomic studies is to characterize the taxonomic composition of a sample, by classifying a set of anonymous DNA reads. Several computational tools have been developed for this purpose, e.g., PhymmBL [11] uses a hybrid approach combining an Interpolated Markov Model framework with similarity based BLAST search, MEGAN [55] uses the lowest common ancestor of the top three nearest neighbors based on similarity search, AMPHORA [136] infers a phylogenetic tree based on pair-wise similarity matrix, etc. These tools indeed represent significant improvements for classifying DNA reads, but the problem is still far from solved. For example, most of these tools are not good at handling novel taxonomic groups, require huge amount of computational power, or are not accurate enough.

1.2.3 Statistical comparison

After computational analysis of raw metagenomic sequences (e.g., gene prediction, functional annotation, sequences classification, etc), each sample can be represented as an abundance matrix. The columns are samples, the rows are different features and the cells represent the corresponding abundance value. Typically, the samples come from two distinct groups in a designed study (e.g., disease and heathy), and the goal is to find differential features that are statistically significant. The challenge is to find the relevant features that may contribute to the disease. Some recent studies have more complicated experimental design, for example, within the disease and healthy samples, we may have different geographical partitions and age stratifications. In addition, each sample may have several sets of heterogenous features, resulting in multiple abundance matrices [137]. Several statistical tools have been developed for comparing metagenomic samples: Metastat developed a nonparametric t-test without the assumption that the underlying distribution is normal, and for low abundance features, it utilizes Fisher's exact test [133]. STAMP [95] argues that knowledge of p-value from a statistical hypothesis test is insufficient to make inferences about biological relevance, and it provides several strategies to explore the sources of potential error. LEFSe [112] proposes a linear discriminant analysis effect size method to determine the features that are most likely to explain differences between classes by coupling standard tests for statistical significance with additional tests encoding biological consistency and effect relevance.

1.3 Contributions of This Dissertation

Figure 1.3 shows an overview of the projects we have finished during my PhD research. Most of these projects are centered around the computational analysis of high-throughput metagenomic sequence data, but the algorithms and tools we have developed can be potentially adapted for other relevant computational problems.

1.3.1 Chapter 2 - Hierarchical Classification of Biological Sequences

MetaPhyler is a hierarchical classifier for biological sequences, and can be used to characterize the phylogenetic diversity (or taxonomic composition) of a microbial community. Given a training data set (including biological sequences and their taxonomic labels), MetaPhyler automatically learns the similarity distribution within each group induced by the given hierarchical taxonomy using non-parametric kernel density estimation. During classification, for a query sequence, MetaPhyler first identifies its nearest neighbor, then tries to classify it based on the best neighbor incrementally starting from the leaf to the root of the hierarchy. A confidence score is computed for the classification of each group, representing the likelihood that

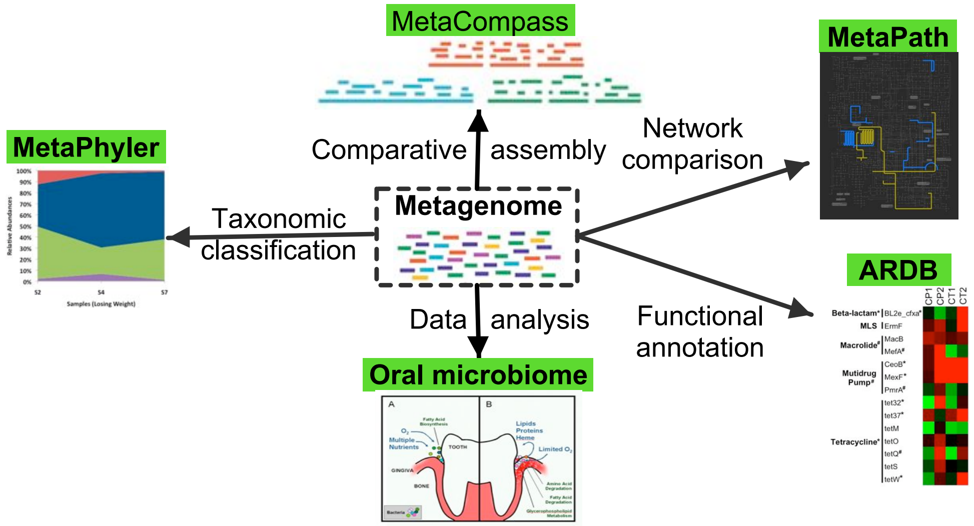


Figure 1.3: Overview of the projects that I have finished during my dissertation. All of these projects are centered around the computational analysis of high-throughput metagenomic sequence data. MetaPhyler [83] is used for taxonomic classification of large-scale whole-metagenome shotgun sequences. MetaPath [84] uses a statistical framework to compare the abundances of weighted metabolic networks and find functional subnetwork signatures. Antibiotic resistance genes database (ARDB) [81] provides a centralized compendium of information on antibiotic resistance, and facilitates the identification and characterization of new resistance genes. MetaCompass is a comparative assembler for metagenomic shotgun sequences. We have applied our tools to a real human oral microbiome dataset to investigate the functional mechanisms at a system level and the ecological changes underlying the transition from health to disease.

this query comes from the same class as its best neighbor. MetaPhyler has been pre-trained for a set of phylogenetic marker genes, and is directly applicable to the classification of metagenomic sequences. Compared with other classifiers, which usually build models for a set of flat classes, MetaPhyler models the hierarchical groups directly, and can easily identify novel taxonomic group from the query, which belongs to an internal class, but not the most specific leaf class.

In addition to taxonomic classification, MetaPhyler can also be adapted to functional annotation of biological sequences. Currently, many gene annotation tasks are performed or supervised manually by some experts, and this process usually is not very efficient and also introduces lots of human biases. Instead, MetaPhyler can be trained to each functional categories or groups, the experts only need to tune various parameters to build optimal classifiers, and the annotation process will be performed in a completely automatic fashion.

1.3.2 Chapter 3 - Statistical comparison of weighted metabolic networks

Analyzing the phylogenetic diversity and comparing the taxonomic composition of metagenomic samples are only the first steps. Another challenge is to measure and compare the biological functions performed by the environmental microbes, as they play pivotal roles in natural environmental processes. A previous study has suggested that the system is better characterized by its gene complement than by its taxonomic composition, given that similar biological functions can be performed by microbes of distinct taxonomic origins [130]. Here our goal is to find differentially abundant metabolic subnetworks that are selected for by their local environments. MetaPath [84] is a statistical tool that uses a greedy search algorithm to find a maximum weight subnetwork, and a nonparametric permutation test to gauge the statistical significance. Previous tools usually partition the whole metabolic network into a set of connected components, which are then considered as a set of discrete features, and traditional statistical comparisons are then performed on entire components at a time. Unlike previous approaches, MetaPath explicitly searches for significant subnetwork in the global network (the set of all known metabolic processes), enabling us to detect signatures at a finer level. In addition, we developed a statistical method that takes into account the topology of the network when testing the significance of the subnetworks.

1.3.3 Chapter 4 - Antibiotic resistance genes database

Antibiotic resistance genes database (ARDB) is an online database, and information retrieval and analysis tool to facilitate the identification and characterization of resistance genes. It unifies most of the publicly available information on antibiotic resistance from various other database sources, e.g., NCBI Clusters of Orthologous Groups, NCBI Conserved Domains Database, KEGG DRUG database, the Chemical Entities of Biological Interest ontology, etc. It allows the users to automatically identify and annotate antibiotic resistance genes for new genome projects, to mine previously identified genes and to compare the resistance profiles across samples. In addition, the information analysis and retrieval infrastructure, and data model can be generalized to building databases for other biological genes. Lot of biomedical information is stored in semi-structured data, and is distributed across different database sources. ARDB also provides a framework for compiling and building such specialized gene information databases.

1.3.4 Chapter 5 - Mining oral microbiome from a system perspective

While much is known about individual species associated with pathogenesis, the system-level mechanisms underlying the transition from health to disease are still poorly understood. Through the sequencing of the 16S rRNA gene and of whole metagenome shotgun DNA, we provide a glimpse at the global genetic, metabolic, and ecological changes associated with periodontitis in 15 subgingival plaque samples, four from each of two periodontitis patients, and the remaining samples from three healthy individuals. We also demonstrate the power of whole-metagenome sequencing approaches in characterizing the genomes of key players in the oral microbiome, including an unculturable TM7 organism.

This project is accomplished through extensive collaborations with several external research groups, including Dr. Salomon Amar, Dr. Daniel Segre, Dr. Simon Kasif from Boston University and Dr. Colin Stine from University of Maryland, School of Medicine. My specific contributions to this project are mainly data analyses about taxonomic diversity, statistical comparison of metabolic networks, genome assembly and genomic variation. For example, taxonomic diversities of the samples are estimated from both 16S rDNA sequences (using DNACLUST [42] and RDP-Classifier [132]) and whole-metagenome shotgun sequences (using MetaPhyler [80]). Further statistical comparison and principal component analysis revealed that diseased samples share a common structure that was not found in completely healthy samples, suggesting that the disease state may occupy a narrow region within the space of possible configurations of the oral microbiome. The computational methodologies developed in this paper can be adapted to other metagenomic analysis, e.g., the estimation of single nucleotide polymorphism rates and genetic diversity based on multinomial distribution.

1.3.5 Chapter 6 - Comparative assembly of metagenomic sequences

So far, one of the biggest challenges in analyzing next-generation sequencing is the short read length. Although many sophisticated computational tools have been designed for short reads, the performance is still not very satisfactory in many applications. All these limitations can be alleviated through metagenomic assembly. Genome assembly is the process of piecing together short DNA fragments that are randomly extracted from a sample, to form a set of longer contiguous stretches of DNA strings called contigs. Assembly, however, is one of the most computationally challenging tasks in metagenomics, and the contig sizes and accuracy generated from current tools are far from satisfactory.

We have developed MetaCompass - a comparative metagenomic assembler, which utilizes the genomes that have already been sequenced, and produces long contigs through read mapping and assembly. Given the availability of thousands of existing bacterial genomes, MetaCompass uses taxonomic information to identify a subset that can be used as appropriate reference set for the data being assembled. Then, the reads are aligned against these genomes through MUMmer-map (alignment algorithm developed by me for this specific task) or Bowtie 2. Afterwards, we use a greedy algorithm for the minimum set-cover problem to build long contigs. The consensus sequences are computed by majority rule. We also propose an iterative assembly approach to improve the accuracy and contiguity of the resulting assembly. MetaCompass is the first available comparative assembler for metagenomic sequences. It has been evaluated on over 7 Tbp metagenomic sequences, from which we show that, when combined with previous approaches, MetaCompass significantly improves the overall assembly quality by about 40%.

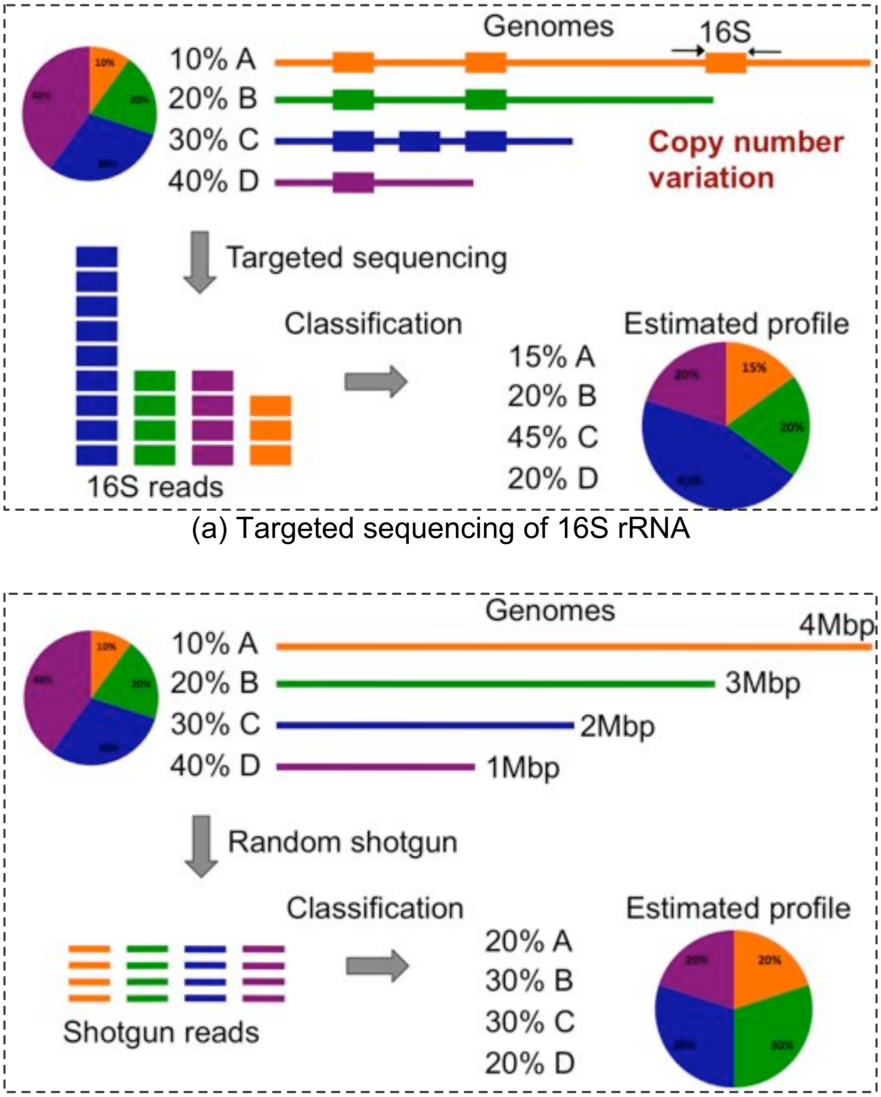
Chapter 2

MetaPhyler: Hierarchical Classification of Metagenomic Sequences

MetaPhyler has been published in [80], and an improved version is under preparation. The methods, algorithms and experiments in this study originated from discussions between Dr. Mihai Pop and me. I developed the program and performed the experiments. Dr. Mihai Pop and I write the paper together.

2.1 Introduction - Taxonomic Profiling for Metagenomics

One fundamental goal in metagenomics is to characterize the taxonomic diversity of a microbial community - taxonomic profiling. This is usually achieved by the targeted sequencing of the 16S rRNA gene, either as a whole, or focused on a hypervariable region within the gene [125]. Then the sequences are classified based on similarity against a curated reference 16S rRNA database [132]. This approach has been a powerful research tool allowing biologists to explore the majority of previously unknown microorganisms populating our world. Approaches based on 16S rRNA sequencing, however, provide a biased estimate of microbial diversity due to the wide variability in copy number of the 16S gene even within closely related organisms (Figure 2.1a), and due to amplification biases inherent in PCR.



(a) Targeted sequencing of 16S rRNA

(b) Metagenome shotgun sequencing

Figure 2.1: Estimating taxonomic profiles using 16S rRNA targeted sequencing or metagenome shotgun sequencing. Panel (a) shows that the taxonomic profile estimated from 16S rRNA targeted sequencing is biased because of copy number variation. Panel (b) shows that classification of whole-metagenome shotgun sequences may produce biased estimation because of the variations in genome size.

A more direct approach for taxonomic profiling is to classify metagenomic reads through similarity search against a reference genes database. MEGAN [55] maps query sequences to the NCBI *nr* database using BLAST, and assigns them taxonomic labels according to the lowest common ancestor of the top database hits.

CARMA [66] first searches for conserved Pfam domains and protein families within the unassembled reads of a sample, then constructs a phylogenetic tree of each matching Pfam family and the corresponding query reads, and finally the reads are classified into a higher-order taxonomy depending on their phylogenetic relationships with respect to the database sequences that have known taxonomic origins. In contrast to similarity-based approaches, machine learning and statistical methods [11, 86] have been used to classify DNA sequences based on DNA base composition signatures (usually k-mer frequencies). Further, a hybrid approach PhymmBL [11] has demonstrated that the combination of machine learning (Phymm) and similarity information (BLAST) produces higher accuracy than either method alone. Despite the difficulties in accurately classifying whole-metagenome shotgun sequences, the estimated taxonomic profiles may be biased because of variations in genome size (Figure 2.1b).

In this paper, we present a novel taxonomic profiling tool (MetaPhyler) [83] for metagenomic sequences, which relies on 31 phylogenetic marker genes [136] as a taxonomic reference. We extend the database described by Wu and Eisen [136] by including marker genes from all complete genomes, the NCBI *nr* protein database and 60 draft genomes. One major limitation of prior methods used in this context is the use of a universal classification threshold for all genes at all taxonomic levels (e.g., BLASTP E-value=0.1 used by AMPHORA [136]). However, individual bacterial genomes and proteins can have different evolutionary rates, and metagenomic reads contain gene fragments of different lengths. We propose that better classification results can be obtained by tuning the taxonomic classifier to the length of each HSP (high-scoring segment pairs in BLAST), to the reference gene, and to the taxonomic level. Our classifier, based on BLAST, uses different thresholds for each of these parameters, which are automatically learned from the structure of the reference database. A side-effect, and an important feature of our tool, is the ability to identify novel organisms or taxa. Results on simulated metagenomic datasets demonstrate that MetaPhyler outperforms previous tools used in this context (CARMA, Megan and PhymmBL). Further, MetaPhyler is much faster than previous tools for two reasons: (1) the size of the reference database is much smaller than the NCBI *nr* database; and (2) our classifier based on BLAST bit scores involves much less computation than some previous approaches which build phylogenetic trees [66, 131, 136]. Finally, we present several interesting results obtained by applying MetaPhyler to the gut microbiomes of obese and lean twins [130].

2.2 MetaPhyler Classifiers

2.2.1 Building a reliable phylogenetic marker genes database

To use metagenomic sequences for taxonomic profiling, we analyzed 31 protein coding marker genes previously shown to provide sufficient information for phylogenetic analysis [136]. These phylogenetic marker genes are universal, present only once in most genomes, and are rarely subject to horizontal gene transfer. Hence, they provide a more accurate estimation of the microbial composition than methods relying on 16S rRNA alone. In order to create an accurate and comprehensive reference dataset, we used the manually curated marker genes from AMPHORA as a seed dataset, and extended them by including marker genes from all complete genomes, the NCBI *nr* protein database and 60 draft genomes. Specifically, we first build MetaPhyler classifiers (see below) on the seed dataset, and then use them to classify potential marker genes. In addition, we have also included phylogenetic marker genes from Archaea, whose information is not available in the seed dataset from AMPHORA. As a result, our final marker genes dataset covers 581 genera, 214 families, 99 orders, 46 classes and 27 phyla.

2.2.2 Building MetaPhyler classifiers

Many previous metagenomic studies employ similarity-based classification methods, and apply a universal threshold for all genes. The taxonomic label of the best similarity hit is then transferred to the query sequence. An improved variant of this approach involves combining the top hits instead of only using the best one [55]. We propose that better classification results can be obtained by tuning the taxonomic classifier to each BLAST HSP length, reference gene, and taxonomic rank. Specifically, by learning parameters from the reference database, we build a taxonomic classifier for a particular reference gene G as follows (Figure 2.2):

1. Simulate 60bp metagenomic reads from all reference marker genes that were curated as described in the previous section and, as a negative set, from genomic sequences that do not contain marker genes.

2. Map these simulated reads against reference gene G using BLASTX.

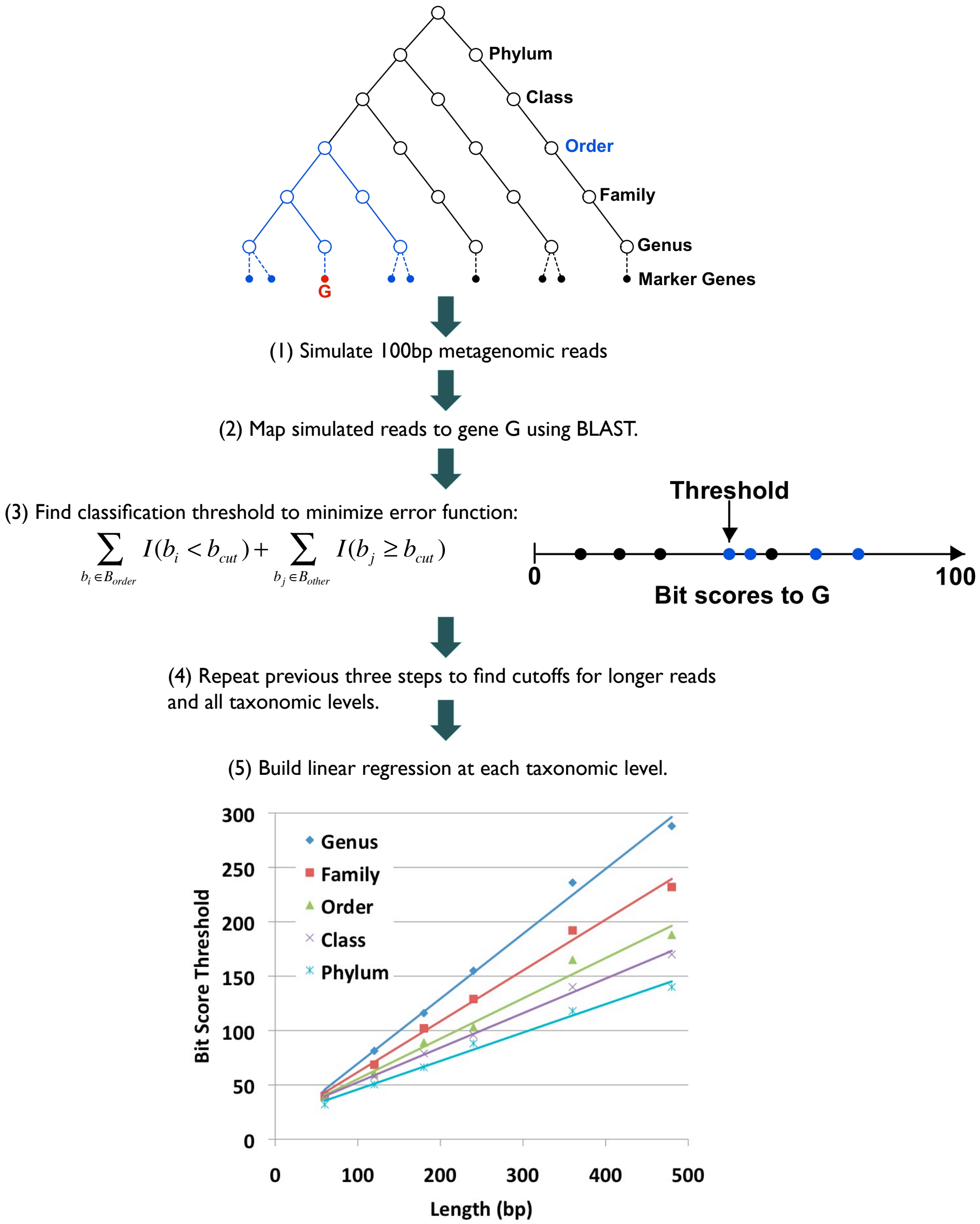

Phylum
Class
Order
Family
Genus
Marker Genes
G
(1) Simulate 100bp metagenomic reads
(2) Map simulated reads to gene G using BLAST.
(3) Find classification threshold to minimize error function:
$\sum_{b_i \in B_{order}} I(b_i < b_{cut}) + \sum_{b_j \in B_{other}} I(b_j \geq b_{cut})$
Threshold
0
100
Bit scores to G
(4) Repeat previous three steps to find cutoffs for longer reads and all taxonomic levels.
(5) Build linear regression at each taxonomic level.
◆ Genus
■ Family
▲ Order
× Class
✳ Phylum
Bit Score Threshold
300
250
200
150
100
50
0
0
100
200
300
400
500
Length (bp)

Figure 2.2: Building MetaPhyler classifier for 60bp long reads of gene G. We first simulate metagenomic reads from all reference marker genes, and as a negative set, from genomic sequences that do not contain marker genes. We then map these simulated reads against reference gene G using BLASTX. To build a classifier for gene G at a specific taxonomic level, say order, in vector B_{order} we store BLASTX bit scores between gene G and the simulated reads that are from the same order; in vector B_{else} we store bit scores for aligning all other reads against G. We then find the bit score cutoff b_{cut} that minimizes Equation 2.1. Finally, we repeat the previous steps to find bit score cutoffs for simulated reads of other lengths and for other genes.

3. To build a classifier for gene G at a specific taxonomic level, say order, in vector B_{order} we store BLASTX bit scores between gene G and the simulated reads that are from the same order; in vector B_{else} we store bit scores for alignments of all other reads against G. Then, we find the bit score cutoff b_{cut} that minimizes the following error function:

   $$\sum_{b_i \in B_{order}} I(b_i < b_{cut}) + \sum_{b_j \in B_{else}} I(b_j > b_{cut}) \qquad (2.1)$$

   where I is an indicator function, which equals 1 when the condition is met, and 0 otherwise. The taxonomic tree used in our analysis is downloaded from the NCBI taxonomy database, however our analysis can be redone with a different taxonomic tree.

4. Repeat the previous three steps to find bit score cutoffs for simulated reads of lengths 120bp, 180bp and up to the length of gene G in 60bp increments.

5. To find cutoffs for sequences of arbitrary matching lengths, we build a linear regression: $b_{cut}^{L} = a + bL$ (see below for why we choose linear regression), where L is the sequence length, b_{cut}^{L} is the bit score cutoff for length L, and a and b are parameters estimated from the data.

6. Repeat steps (3), (4) and (5) to build bit score cutoff regressions for other taxonomic levels (genus, family, class and phylum) for gene G.

We, then, repeat the above procedures to build classifiers for all reference marker genes in our database.

In step (3), we assume that bit scores from close phylogenetic neighbors are higher than distant neighbors. This is generally true because marker genes, which are more closely related phylogenetically, tend to have more similar sequences. However the phylogenetic relationships of the marker genes are not fully consistent with the corresponding taxonomic tree, which is downloaded from the NCBI taxonomy database. Ideally we would expect to see the cutoff b_{cut} to be lower than all the scores in B_{order}, but higher than scores in B_{else}. The error metric (Equation 2.1) we used is a count of the number of misclassified points, which is similar to the 2-norm distance used by SVM classifiers.

Next, we show that in step (5) linear regression is a reasonable approximation of bit scores based on the matching HSP length. As described in [4], the bit score is

$$S_{bit} = (\lambda S - \ln K)/\ln 2 \tag{2.2}$$

where S is the raw score of the BLAST alignment, and λ and K are parameters depending on the database. In addition, the raw score S equals the sum of the scores of matching amino acids [4]

$$S = \sum S_{ij} = \log(q_{ij}/p_i p_j) \tag{2.3}$$

which is the log-odds ratio of the observed and expected frequencies. For gene G of length L, we can rewrite Equation 3 as $S = L \sum S_{ij}/L$. For metagenomic read G' of length $L'(L' \leq L)$, which only contains a subsequence of the full-length gene G, the raw score $S' = L' \sum S_{mn}/L'$. Further, if we assume that the evolutionary mutations and amino acid compositions are randomly distributed across gene G, then

$$\sum S_{mn}/L' \approx \sum S_{ij}/L = \overline{S}_{ij} \tag{2.4}$$

which indicates that $S' = L'\overline{S}_{ij}$. Hence, we can rewrite equation 2.2 for a gene fragment as

$$S'_{bit} = (\lambda L'\overline{S}_{ij} - \ln K)/\ln 2 \tag{2.5}$$

where $\overline{S}_{ij}$ is a constant for a particular gene G. As a result, the bit score (S'_{bit}) of a subsequence of gene G is linearly correlated with the HSP length (L'), and we can estimate this relationship with a linear regression as in step (5).

2.2.3 Classifying metagenomic sequences

The query metagenomic sequences are initially mapped to the reference marker genes using BLASTX. MetaPhyler classifies each sequence individually based on its best reference hit. For example, assume that a query sequence Q has gene G as its best hit, the BLAST bit score is b and the HSP length is L. First we try to classify Q at the genus level by calculating the bit score cutoff b_{cut} of gene G using the pre-computed linear regression function. If the bit score is higher than the cutoff ($b \geq b_{cut}$), then we transfer the genus label of reference G to query Q. Otherwise, we try to classify Q at higher taxonomic levels (family, order, class and phylum) using level-specific classifiers built for gene G, until either the classification is successful at one of the taxonomic levels or the query can not be classified.

A side-effect of this algorithm, specifically the stringent classification strategy that can avoid assigning an organism to a lower-level taxonomic group if the evidence does not support this assignment, is the ability to identify novel organisms or taxa. The presence of novel organisms leads to a detectable discrepancy between the number of sequences assigned to a lower taxonomic level, and the number of sequences assigned to a higher (less specific) taxonomic level. For example, if a set of query sequences are classified into a particular order, but cannot be classified into any existing families under this order, then this indicates that these reads come from novel family-level clades. These sequences can be further analyzed using a *de novo* approach, e.g., using Minimus [120], which will potentially recover the full-length gene and, thus, help characterize the novel bacterium. In order to help the users easily identity novel bacteria from MetaPhyler output, we used the following naming rule for example: if a sequence is classified at the family level as Enterobacteriaceae, but can not be classified to any genera under it, then we name this sequence as Enterobacteriaceae{family} at the genus level.

2.2.4 Estimating bacterial composition

After taxonomic classification of phylogenetic marker genes from metagenomic sequences in the previous step, for each taxonomic unit, we have a set of reads assigned to each phylogenetic marker gene. The depth of coverage of this taxonomic unit is calculated as the median of that of the 31 phylogenetic marker genes. Then the relative abundances of all taxonomic units are computed using the depth of coverage instead of the number of reads classified. Table 2.1 shows an example of MetaPhyler output at the genus level for the simulated metagenomic sample in Table 2.

Genus	Coverage	Abundance	# Reads Mapped
Bifidobacterium bifidum PRL2010	24.98	49.97%	3765
Bacteroides fragilis NCTC 9343	10.19	20.37%	1806
Staphylococcus aureus USA300	5.12	10.24%	879
Enterococcus faecalis V583	5.03	10.06%	823
Clostridium difficile 630	4.68	9.36%	748

Table 2.1: An example of MetaPhyler output.

2.2.5 Customized training and classification confidence score

The previous version of MetaPhyler (described above) is specifically designed for taxonomic classification of phylogenetic marker genes. The hierarchical classification algorithm, however, is applicable to all homologous biological sequences or genes. In this section, we describe the second version of MetaPhyler with the following two new features: (1) allow users to train their own MetaPhyler classifiers, given a set of training data set which includes homologous sequences and their hierarchical labels; (2) instead of calculating a classification threshold during training, we compute a confidence score for during hierarchical classification.

The training step is performed in about the same way as in Figure 2.2. The difference is step (3). Instead of finding a classification cutoff that minimizes the training error rate, we model the distribution of the BLAST similarity scores within different hierarchical groups using a nonparametric density estimation approach. For example, in terms of NCBI taxonomy, MetaPhyler will estimate the distribution of similarity scores at the genus, family, order, class, and phylum taxonomic groups. Figure 2.3 shows the distribution of BLAST bit score with different taxonomic groups for rplB gene. As expected, the scores for lower taxonomic groups are higher than that for upper ones.

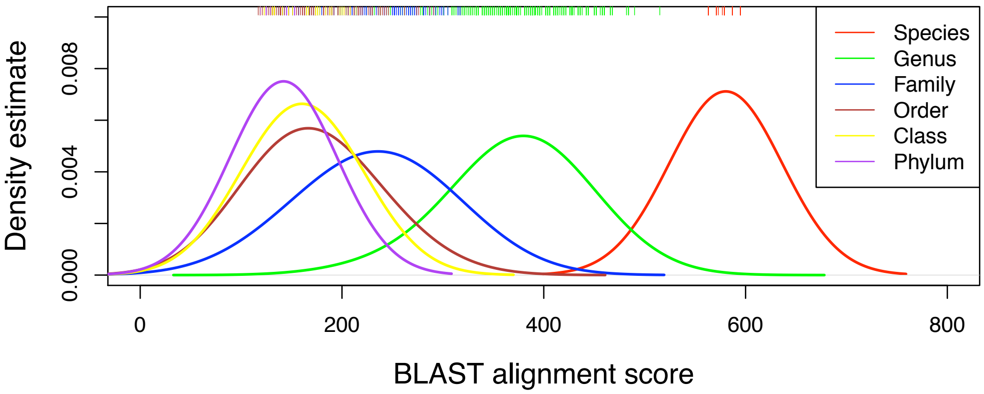


Figure 2.3: Distribution of BLAST scores within different taxonomic clusters for rplB genes with 300bp length. 300bp DNA sequences are simulated from available rplB genes. Density distribution is estimated using Gaussian kernel. On average, similarity scores from lower hierarchical/taxonomic clusters are higher than that from upper clusters.

Once we have built the classifiers as described above, to classify a query reads, we first find its best neighbor through BLAST search. And use the BLAST bit score and the density distribution we already have for the best neighbor to compute the confidence score for each hierarchical classification. Specifically, the confidence score is defined as follows: Given an anonymous query sequence Q, its best neighbor R is determined by a BLAST search, and the bit score is S. Then we will use the classifier built for R to classify query Q. The confidence score at a taxonomic level k is calculated as

$$conf_k = \frac{\sum_{i=1}^{k} P_i(X \leq S)}{\sum_{i=1}^{k} P_i(X \leq S) + \sum_{j=k+1}^{n} P_j(X \geq S)} \tag{2.6}$$

where $P_i(X \leq S)$ represents the probability of observing a score that is less than or equal to S in the estimated density distribution at taxonomic level i. n is the total number of taxonomic levels for R. The taxonomic levels are ranked as 1 to k starting from leaves to root. Figure 2.4 shows an example about how to compute the confidence score.

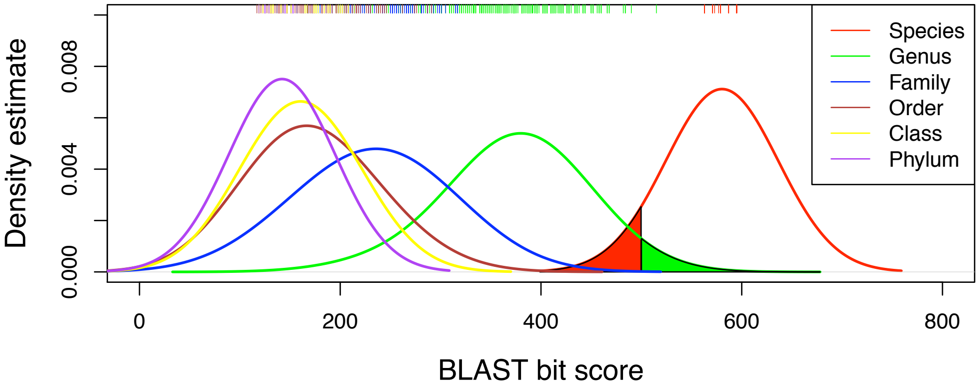


Figure 2.4: The figure configuration is the same as that in Figure 2.3. Suppose the alignment score for a query is 500, then the confidence score is computed as the ratio of red region/(red region + green region).

2.3 Results

2.3.1 Performance evaluation using simulated datasets

We carried out a simulated metagenomic study by comparing MetaPhyler with three other widely used tools: WebCarma [66], MEGAN [55] and PhymmBL [11]. We have randomly simulated around 300K 60bp and 70K 300bp DNA sequences from 31 phylogenetic marker genes. Figure 2.5 compares the sensitivity (*number of correct predictions / number of simulated reads*) and precision (*number of correct* *predictions / number of predictions*) of the phylogenetic assignments at five taxonomic levels. The query sequence itself was removed from the reference dataset when running MetaPhyler, MEGAN and PhymmBL. We can see that MetaPhyler, MEGAN and PhymmBL have comparable precisions in almost all cases, and MetaPhyler is a little bit better than others at the genus level. However, the sensitivity of MetaPhyler is significantly better than other tools in all situations, perhaps due to the fact that the classifiers are explicitly trained at each taxonomic level. Figure 2.6 and Figure 2.7 show the classification precision and sensitivity for six marker genes separately. We can see that the variations in performance on different marker genes are not significant.

One of the major challenges of metagenomic analysis is the presence of novel DNA sequences which do not match well any data in current databases. One major goal of metagenomic analysis is to discover and classify such novel sequences. For example, we asked the following question: given a read from an organism whose genome has not been sequenced before, and also no sequences from the same genus are available, can we classify this sequence correctly at the family level provided that we have sequences from other organisms within the same family? We further examined the performance of MetaPhyler using progressively less data from organisms related to those from which the query sequences were simulated. Table 2.2 summarizes the sensitivity and precision performance evaluated on 60bp and 300bp

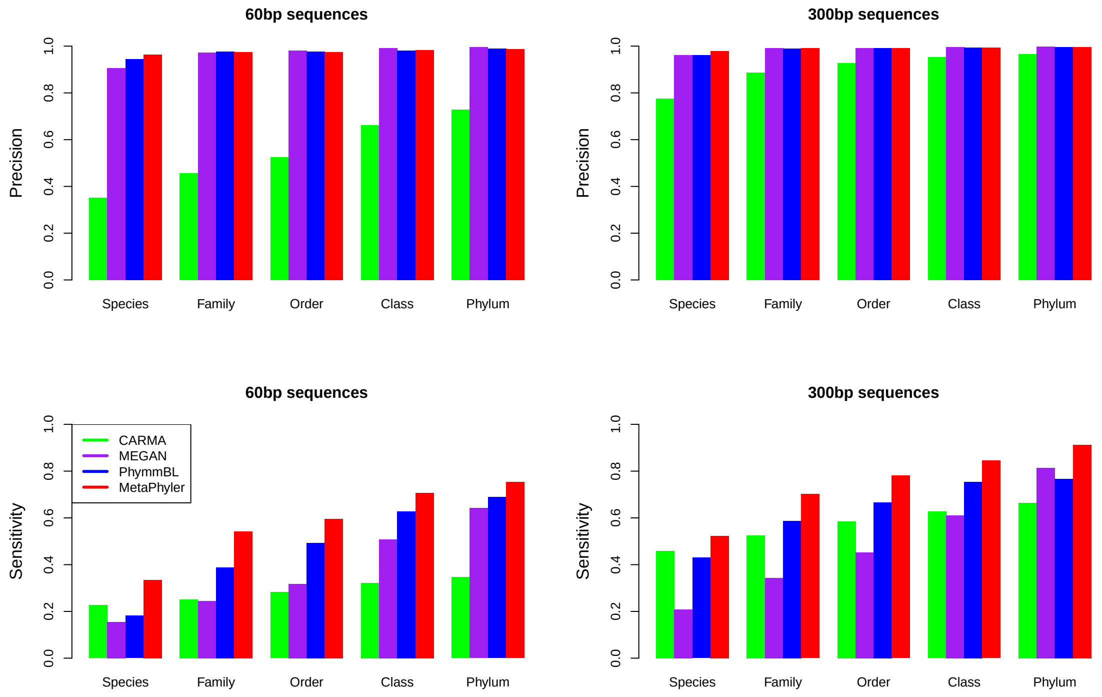


Figure 2.5: Comparison of phylogenetic classification performance of MetaPhyler, MEGAN, CARMA and PhymmBL. The sensitivity and precision are calculated across five taxonomic levels using 60bp and 300bp simulated metagenomic reads. During the classification with MetaPhyler, MEGAN, and PhymmBL, reference sequences that are from the same genome as the query reads are excluded. CARMA results are from the classifications based on WebCARMA server. This figure shows that the sensitivity of MetaPhyler significantly outperforms the other three methods, and that the precision is also slightly better at the genus level.

simulated metagenomic reads. Overall the classification precision is still very high when fewer reference marker genes are available. This is especially true for the 300bp reads: even if no sequences in the database originate from the same genus as the query reads, the precision is still higher than 92% when classifying at higher taxonomic levels.

As we have discussed in the Introduction section, estimating the abundance of taxonomic groups in a sample through the classification of phylogenetic marker genes

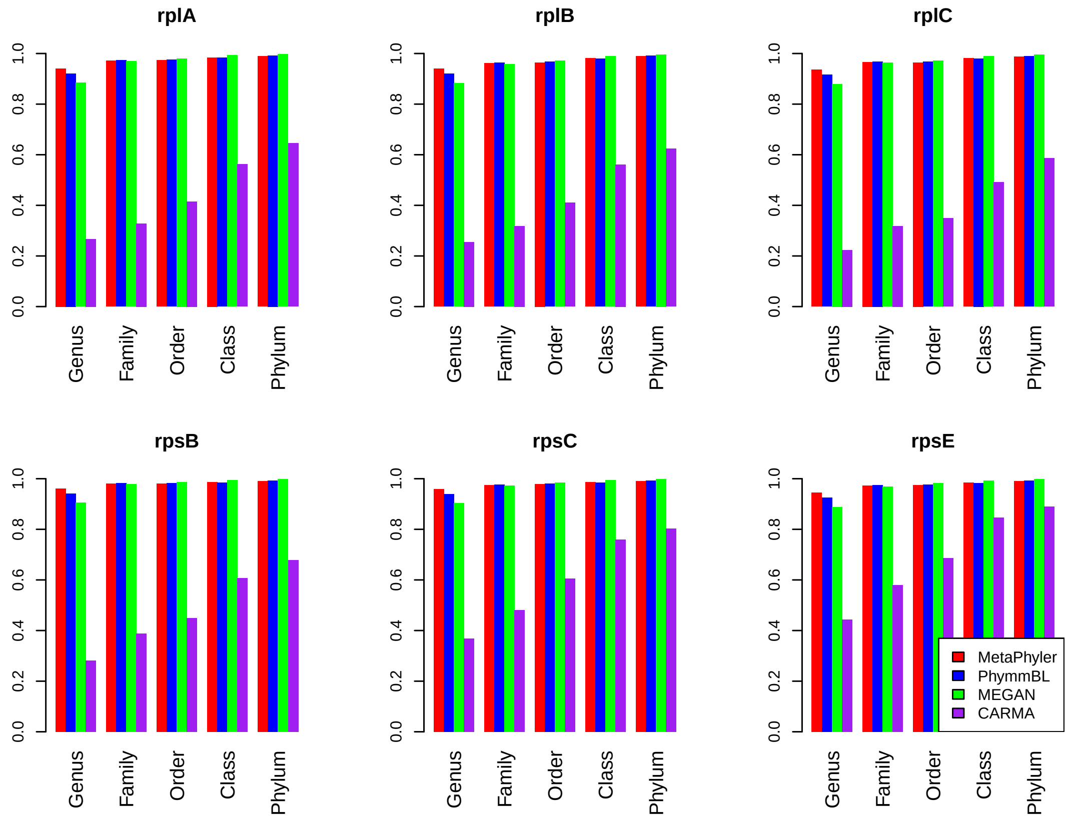


Figure 2.6: Classification precision for six marker genes (rplA, rplB, rplC, rpsB, rpsC and rpsE) evaluated on four tools (MetaPhyler, PhymmBL, MEGAN and CARMA). Red bar is MetaPhyler (1st column); blue bar is PhymmBL (2nd column); green bar is MEGAN (3rd column); purple bar is CARMA (4th column).

is more accurate than that obtained through 16S rRNA analysis or classification of all of the metagenomic shotgun sequences. In order to validate our hypothesis, we have created a simple simulated metagenomic sample comprising 5 genomes (Table 2.3). We compared the accuracy of the taxonomic profiles estimated by different approaches (Figure 3). The genomes, which are present in the simulated sample, are eliminated from MetaPhyler reference database. MetaPhyler outperforms other approaches dramatically, and is very close to the true taxonomic profile. While for approaches based on classifying 16S rRNA and all the shotgun sequences, even if we

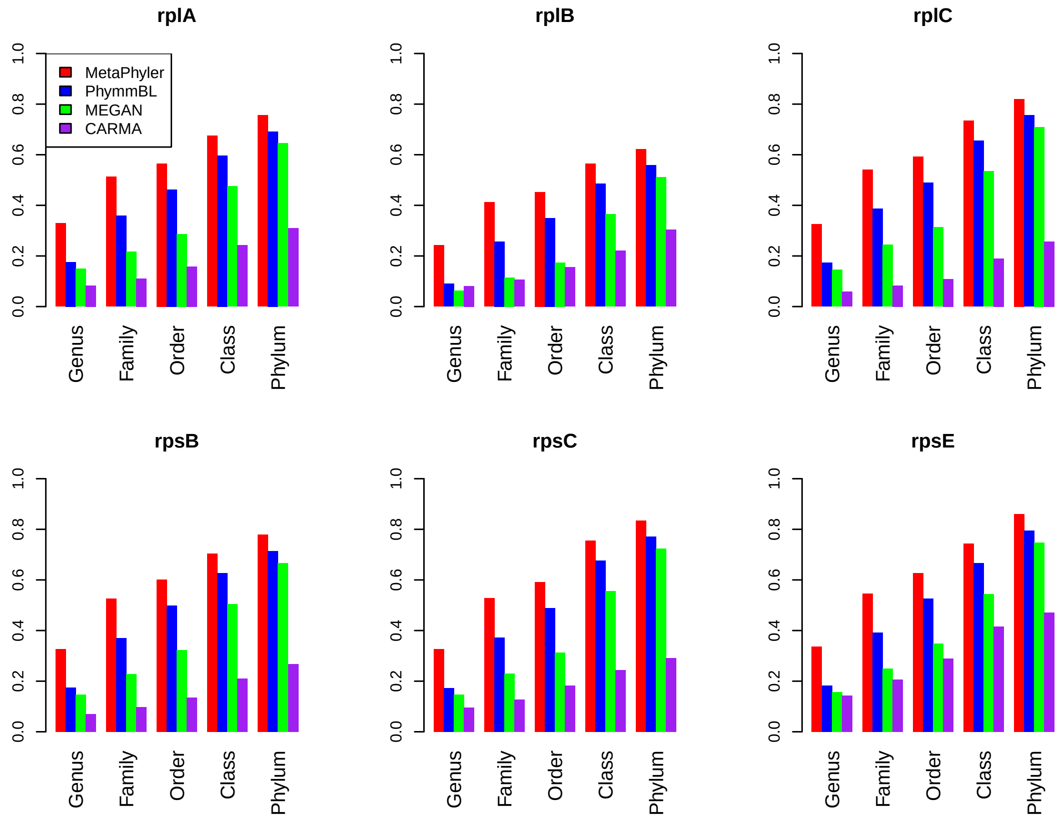


Figure 2.7: Classification sensitivity for six marker genes (rplA, rplB, rplC, rpsB, rpsC and rpsE) evaluated on four tools (MetaPhyler, PhymmBL, MEGAN and CARMA). Red bar is MetaPhyler (1st column); blue bar is PhymmBL (2nd column); green bar is MEGAN (3rd column); purple bar is CARMA (4th column).

assume that the classification is perfect ("16S Ideal" and "Shotgun Ideal" in Figure 2.8), the resulting taxonomic profile is still highly biased.

2.3.2 Detecting novel organisms

As mentioned in the Introduction (see Methods for details), MetaPhyler can help to identify novel bacteria from metagenomic sequences. Here we show a concrete example based on sample F10T1Ob1 from the above-mentioned human gut

Exclude	**60bp**					**300bp**				
Training	Genus	Family	Order	Class	Phylum	Genus	Family	Order	Class	Phylum
Genome	90.72 33.45	97.18 54.22	98.10 59.59	99.11 70.72	99.56 75.30	97.90 52.39	99.14 70.17	99.15 78.09	99.34 84.52	99.64 91.18
Genus		77.15 16.47	86.32 23.16	94.92 34.60	96.72 43.48		92.55 31.06	95.71 48.63	98.23 64.22	98.84 77.35
Family			63.62 13.19	90.31 24.64	94.65 34.99			85.25 26.65	96.78 53.15	97.66 69.42
Order				80.04 17.73	90.29 27.80				93.69 39.97	96.26 58.86
Class					78.16 16.59					90.94 42.62

Table 2.2: MetaPhyler performance using fewer and fewer training dataset. MetaPhyler phylogenetic classification performance on 60bp and 300bp simulated metagenomic reads. For each prediction, the top and bottom numbers are precision and sensitivity in percentage, respectively. Different taxonomic levels are excluded when evaluating the classification, e.g., ‘Genus’ means genes that have the same genus label as the query read are excluded from the reference training dataset.

Species	**Coverage**	**Abundance**	**Genome Size**	**# 16S rRNA**
Bifidobacterium bifidum PRL2010	25	50%	2.2Mbp	3 copies
Bacteroides fragilis NCTC 9343	10	20%	5.1Mbp	6 copies
Staphylococcus aureus USA300	5	10%	2.8Mbp	5 copies
Enterococcus faecalis V583	5	10%	3.2Mbp	4 copies
Clostridium difficile 630	5	10%	4.2Mbp	11 copies

Table 2.3: Simulated metagenomic sample. To evaluate the performance of different approaches in estimating the bacterial composition, we have created a simulated metagenomic sample consisting of 5 species with 100bp reads. ”Coverage” indicates the depth of the coverage of the simulated reads in the simulated sample for the genomes.

metagenome dataset. We have identified a set of reads belonging to the order Clostridiales, but novel at the family level. We then used Minimus [120] to assemble 9 reads that are mapped to the *rplB* gene. One of the resulting contigs (comprising 5 reads) contained the full-length *rplB* gene. We searched the contig against the NCBI *nr* database and identified as the best hit the *rplB* gene from species *Ruminococcus*

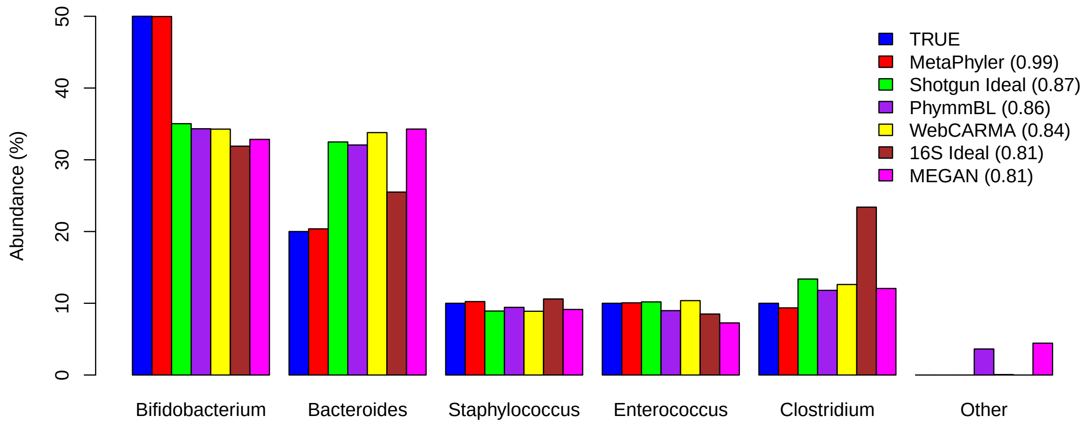


Figure 2.8: Comparison of bacterial compositions estimated from different approaches. We have created a simulated metagenomic sample (Table 2.3) with 100bp reads to evaluate the performance of different approaches in estimating the bacterial compositions. "16S Ideal" and "Shotgun Ideal" represent results obtained by analyzing 16S rRNA genes and whole genome shotgun sequences assuming the classification accuracy is perfect. Genus "Other" indicates that sequences have been classified into genera other than that in the simulated sample. Different approaches are ranked by their correlation coefficients (shown in legend) between the estimated and true taxonomic profile. When running MetaPhyler, the genomes from which the reads were simulated are removed from the reference database.

sp. SR1/5 with 94% and 86% similarity at the amino acid and nucleotide levels, respectively. In addition, our assembly of another contig containing a fragment of *rplB* gene had 93% and 82% similarity with *Blautia hansenii DSM 20583* at the amino acid and nucleotide levels, respectively. Given the low level of similarity at the nucleotide level between the genes extracted from the dataset and all previously characterized genes, we can be fairly confident that the *rplB* genes we identified are novel and likely belong to previously unsequenced members of the Clostridiales order. It is important to note that this discovery was made possible by the stringent strategy we employ which avoids assigning an organism to a lower-level taxonomic group if the evidence does not support this assignment, a feature not available in other taxonomic profiling tools.

2.3.3 Comparison of running times

We compared the running time of MetaPhyler with three other tools (PhymmBL, MEGAN and WebCarma) on 70K 300bp simulated phylogenetic marker gene fragments (Table 4). On a single 2.4GHz processor, the running times (including BLAST search) of MetaPhyler, PhymmBL and MEGAN for analyzing the simulated dataset are 8 hours, 4 days, and 34 days, respectively. On the same dataset WebCarma [41] took 24 hours. MetaPhyler is much faster than other tools in estimating the taxonomic compositions from metagenome shotgun sequences.

Dataset	**Wallclock hours**			
	MetaPhyler	**PhymmBL**	**MEGAN**	**WebCarma**
70K reads	6 hrs	96 hrs	816 hrs	24 hrs

Table 2.4: Comparison of the estimations of bacterial compositions from different approaches. On a single 2.4GHz processor, the computation time (Wallclock hours) used by MetaPhyler, PhymmBL and MEGAN for analyzing 70K 300bp simulated sequences. CPU hours for WebCarma are calculated using its web server.

2.3.4 Conclusion

We have introduced a novel taxonomic classification method for analyzing the microbial diversity of metagenomic sequences. Compared with previous approaches, MetaPhyler provides significantly higher sensitivity when classifying 60bp and 300bp simulated reads; MetaPhyler has slightly higher classification precision at the genus level, and comparable precision at higher taxonomic levels. More importantly, the taxonomic profiles estimated by MetaPhyler are much more accurate than those estimated by other tools. In addition, MetaPhyler is much faster than other tools for taxonomic profiling because (1) the reference marker genes database is much smaller than a general reference genes database (e.g., the NCBI *nr* database), (2) and also our classifier based on BLAST statistics involves much less computation than building phylogenetic trees (another approach used for taxonomic profiling). The high performance of MetaPhyler makes it suitable for large scale metagenomic studies, e.g., the Human Microbiome Project. Furthermore, analysis of publicly available metagenomic data agrees with previous observations, and also provides new insights into the microbial diversity of the human gut ecosystem. Finally, we have demonstrated that MetaPhyler can be used to guide the discovery of novel organisms from metagenomic sequences.

The novel classification algorithm for short DNA reads we have introduced in this paper can also be applied to other conserved genes. We are planning to release a general gene fragment classifier, which can learn classification thresholds automatically from a user provided dataset. In addition, instead of providing a binary result for each classification, we will also explore techniques for generating ”fuzzy” classifications based on confidence scores. The software described in this paper is freely available under an open-source license from `http://metaphyler.cbcb.umd.edu`.

Chapter 3

Comparative Analysis of Metabolic Networks

MetaPath has been published in [84]. The methods, algorithms and experiments in this study originated from discussions between Dr. Mihai Pop and me. I developed the program and performed the experiments. Dr. Mihai Pop and I write the paper together.

3.1 Introduction - Functional Signatures in Metagenomics

Metagenomics is a new scientific field that involves the analysis of organismal DNA sequences obtained directly from an environmental sample, enabling studies of microorganisms that are not easily cultured in a laboratory [104]. Metagenomic studies, pioneered in the early 2000s [9], have recently increased in number and scope due to the emergence of next generation sequencing technologies.

Due to the difficulty of assembling entire organisms from a metagenomic dataset, most comparative metagenomic analyses take a gene-centric view, treating the community as an aggregate and ignoring the exact assignment of genes to individual organisms. In fact, it can be argued that the environment is better characterized by its gene complement than by its taxonomic composition, given that similar biological functions can be performed by microbes of distinct taxonomic origins [130]. The functional profile for a sample can be recovered by mapping sequences to gene families [123], subsystems [87] or metabolic pathways [61]. The relative abundance of each functional category can be estimated by counting how many sequences are assigned to each category, and this information is the basis for detailed comparisons of the functional potential of different functions. In a typical comparative metagenomics experiment, sequences are generated from a collection of samples belonging to two groups, for example, obese or lean twins [130], and healthy infants or adults [69]. An important biological problem is to find differentially abundant functional signatures (e.g., genes or metabolic pathways) that are selected for by their local environments.

Traditional analysis approaches compare the relative abundances of the categories one-at-a-time between different phenotypes, and compute the significance using one of several statistical approaches [43,107,133]. When comparing communities at the gene family level, many functional categories are commonly found to be differentially abundant, even after correcting for multiple hypothesis testing [69,130]. The interpretation of these data can be daunting. An alternative approach focuses on functional subsystems and metabolic pathway comparisons [126], the number of which is much smaller than gene families. Results at these levels are easier to interpret and can provide a stronger evidence of distinct functional capacities than at the level of individual gene families. Such analyses, however, can be unnecessarily coarse. For example, the use of KEGG pathways as a basis for analysis is complicated by the following issues: (1) the definitions of pathways in KEGG are coercive, and the interactions between these pathways are ignored; (2) the genes in a pathway may not be fully covered by the identified genes in a metagenomic sample; (3) significant differences in the abundance of certain genes may be masked once the abundance of all genes in a pathway is aggregated.

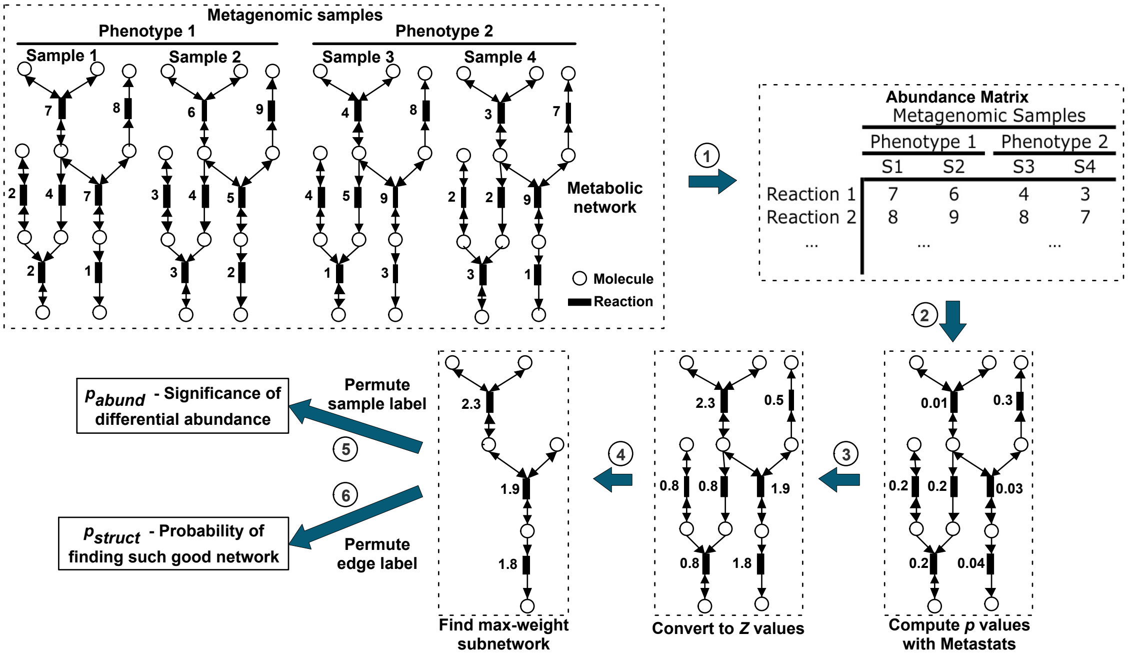
To address these problems, we introduce a general method (MetaPath) for searching the global metabolic network to find differentially abundant finer-level subnetworks. For the purposes of this paper we define a subnetwork to be a connected set of genes that is statistically enriched or depleted in one group of samples. Underlying our approach is a statistical scoring system that captures the differential abundance for a given subnetwork, combined with a greedy search algorithm for a maximum weighted subgraph, to indentify the highest scoring subnetworks. Unlike previous approaches, MetaPath explicitly searches significant subnetwork in the global metabolic network (rather than the KEGG defined pathways), enabling us to detect subnetworks spanning predefined “containers”. In addition, we developed rigorous statistical methods that take into account the topology of the network when testing the significance of the subnetworks.

Using simulated datasets, we demonstrate that Metapath outperforms previously described approaches for comparing biological networks based on abundance data. We show that our findings are more robust to noisy data than the results of single gene comparisons, and that MetaPath can find finer-level subnetwork than can be found by comparing predefined KEGG pathways. We also discuss the biological significance of the results derived from the application of MetaPath to actual metagenomic datasets, demonstrating that the output from MetaPath is easy to interpret and provides valuable biological insights. The software is freely available at http://cbcb.umd.edu/ boliu/metapath/.

3.2 Methods and Materials

3.2.1 Datasets

We tested our methods on two previously published metagenomic datasets, which were downloaded from the NCBI Trace Archive or Short Read Archive databases: (1) gut microbiomes from obese and lean twins [130]; (2) metagenomes from adult- and infant-type gut microbiomes [69]. Each dataset is divided into two populations of distinct phenotypes. The metabolic pathway data were downloaded from the KEGG pathways database [61]. The metabolic network is represented as a graph where nodes are metabolic substrates, and edges are molecular reactions (Figure 3.1). The edges could be unidirectional or bidirectional depending on whether the corresponding reaction is reversible (as specified in KEGG database). Multiple reactions that are related to a same biological process are aggregated by KEGG into a "pathway" (e.g., glycolysis pathway). In addition, we refer to the network comprising all metabolic pathways in KEGG as the "global metabolic network". Metagenomic sequences are annotated through BLASTX searches against KEGG genes database. The abundance of each molecular reaction is estimated as the number of metagenomic sequences mapped to it. Note that more accurate abundance estimates can be obtained by taking into account the length of individual genes [117] and we plan to explore the use of such estimates (and the associated statistics) in future versions of our software.



	Metagenomic Samples			
	Phenotype 1		Phenotype 2	
	S1	S2	S3	S4
Reaction 1	7	6	4	3
Reaction 2	8	9	8	7
...		...		...

Figure 3.1: Schematic diagram of MetaPath methods. Sequences are annotated against KEGG database and are mapped to metabolic network leading to an abundance matrix where the rows are different reactions and columns are samples. Then p values are computed for all reactions using Metastats [133], and are converted into Z values, then greedy search is performed to find subnetworks with high weights. Finally, we estimate the null distribution of the subnetwork score by randomly permuting the sample labels, and compute the p values.

3.2.2 Scoring Metabolic Subnetworks

To score the biological activity of a particular subnetwork, we first use Metastats [133] to calculate the significance of differential abundance for each reaction between two groups. Under the null hypothesis, the relative abundances are random and have the same distribution across phenotypes, thus the p values follow a uniform distribution from 0 to 1. Based on this assumption, p values can be converted to Z scores [56]. Because Metastats performs a two-tailed test for each reaction, the two-tailed p values can be converted back to the original Z values using the following equation:

$$Z_i = \begin{cases} CDF_{sn}^{-1}(1 - p_i/2) \times -1 & \text{, if mean(G1)} < \text{mean(G2)} \\ CDF_{sn}^{-1}(1 - p_i/2) & \text{, if mean(G1)} > \text{mean(G2)} \end{cases} \tag{3.1}$$

$CDF_{sn}^{-1}$is the inverse cumulative density function of standard normal distribution; G1 and G2 represent populations 1 and 2. Using this formula, if a reaction is more abundant in population G1, then its Z score will be positive and vice versa. We are specifically interested in finding a network whose reactions are either enriched or depleted as a whole, as apposed to previous approaches [28, 56] that identify active or perturbed subnetworks, which may contain a mixture of enriched and depleted components. We define the aggregate score for a particular subnetwork to be the sum of the Z scores over all reactions contained within it: $Z = \frac{1}{\sqrt{k}} \sum_{1,k} Z_i$.

3.2.3 Identifying High-Scoring Subnetworks

We attempt to find networks that maximize the cumulative Z-score defined above. Unfortunately, this problem is NP-hard, equivalent to finding a maximum-weight subgraph [56]. Several approaches to solving this problem have been previously proposed: [56] used simulated annealing, but this heuristic is slow; [28] used integer linear programming that can find provably optimal subnetworks quickly, but it requires the commercial software CPLEX which is not available to the general public (re-coding this algorithm using other freely available ILP solvers is beyond the scope of this paper). Here we rely on a greedy heuristic that is fast, and, while not guaranteed to find maximally scoring networks, performs well in practice (Algorithm 1).

Algorithm 1 Searching Max-Weight Subnetwork with a Greedy Heuristic

Input: A global metabolic network $G = (V, E)$, where V and E are metabolites (vertices) and reactions (edges); a set of weight values Z associated with each edge in graph G.
Output: A max-weight subnetwork G_{max} of G and its score W_{max}.

```
 1: Initialize $W_{max}$ to 0;
 2: for all edge $e_i$ in E do
 3:    Initialize $G_{now}$ by including $e_i$;
 4:    Initialize $W_{now}$ to be the weight of $e_i$;
 5:    Initialize $W_{pre}$ to be 0;
 6:    while $W_{now} \geq W_{pre}$ do
 7:       $W_{now} = W_{pre}$;
 8:       Pick an edge $e_k$ which has the highest weight among all edges adjacent to
          $G_{now}$;
 9:       Include $e_k$ into $G_{now}$;
10:       Calculate the score $W_{now}$ of $G_{now}$;
11:    end while
12:    $W_{now} = W_{pre}$;
13:    $G_{now} = G_{now} - e_j$;
14:    if  $W_{now} > W_{max}$ then
15:       $W_{max} = W_{now}$;
16:       $G_{max} = G_{now}$;
17:    end if
18: end for
19: Output $G_{max}$ and its score $W_{max}$
```

This algorithm tries to find a connected metabolic subnetwork, which can have any arbitrary structure, with maximum weight. However, it is believed that in metabolic networks, chains are especially more biologically meaningful and interesting, because they attempt to capture the structure of a series of reactions that are successively connected. To allow this idea, we modify line 8 of the above algorithm to Pick an edge e_k which has the highest weight of the edges that are adjacent to and have the same direction with e_{k-1}. Both searching algorithms are implemented in our program and can be selected through command-line parameters. To find all significant subnetworks (computing significance is discussed below), we iteratively remove the edges in the global network that are contained in previously found significant subnetworks, and rerun our greedy search on the rest of the network until we can no longer find any additional significant subnetworks.

This algorithm tries to find a connected subnetwork with k edges, which can have any arbitrary structure. However, it is believed that in metabolic network, chains are especially more biologically meaningful and interesting, because they attempt to capture the structure of a series of reactions that are successively connected. To allow this idea, we modify line 5 of the above algorithm to Pick an edge e_j which has the highest weight of the edges that are adjacent to and have the same direction with e_{j-1}. Both searching algorithms are implemented in our program. In addition, we also compute the top m high-scoring subnetworks by iteratively removing the edges in the graph associated with subnetworks already considered by our algorithm.

3.2.4 Computing the p Values of Significance

The null score distribution for a specific subnetwork can be estimated by permuting the sample labels (columns of the abundance matrix) of the reactions and computing the subnetwork scores from the permuted abundance matrix. The significance p value is estimated as the number of random permutations that produce higher scores than the original subnetwork. The p value computed through this approach (termed p_{abund}), however, ignores the topology of the underlying global metabolic network, and potentially leads to incorrect conclusions. For example, assume we have a densely connected metabolic network, in which every edge is connected with all other edges. Then, the best subnetwork is simply composed of the top differentially abundant metabolic reactions. This indicates that whenever there are significant reactions, which may simply come from random noise given the large number of edges, they will form a significant subnetwork because of the biases from the network topology (Figure 3.2). To address this problem, we compute another p value (termed p_{struct}), relying on a topological definition of the null distribution of subnetwork scores. Specifically, instead of treating each subnetwork as a bag of genes, we estimate the distribution of scores for actual subnetworks identified within the underlying global metabolic network. Since this null-distribution depends on the size (number of edges) of the subnetwork, let k be the size of a subnetwork generated by the greedy search algorithm described above, and Z be the corresponding Z-score. The p_{struct} value for this subnetwork can be calculated as follows:

1. Permute the edge weights (row labels of the abundance matrix) of the global metabolic network.

2. Perform greedy search to find a maximal weighted subnetwork of size k.

3. Repeat step 1 and step 2 for 1000 times, and generate 1000 weights of the max-weight subnetwork (null distribution).

4. The p_{struct} value is the proportion of the 1000 permutations in step 3 that we see scores higher than our original observation Z.

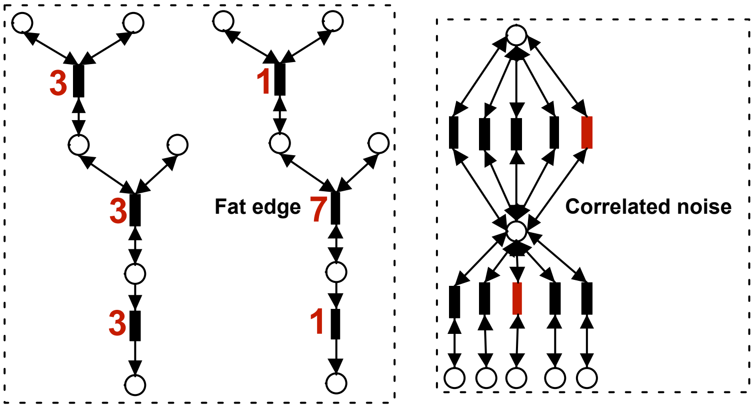


Figure 3.2: Significant subnetworks caused by structural biases. On the left side, the two pathways have equal weight, indicating equal statistical significance. The high weight of the second pathway, however, mainly comes from the middle fat edge. On the right side, in a densely connected network, any random high-weight edges will form a subnetwork with high weight (correlated noise).

3.2.5 MetaPath Methods Summary

To summarize the methods described above, the MetaPath algorithm proceeds as follows:

1. Differential abundance is assessed on an edge-by-edge basis (reaction-byreaction) using Metastats.

2. The significance estimates (p values) from Metastats are fed into a greedy search algorithm to determine all maximally weighted subnetworks(in terms of statistical Z scores) in the global metabolic network.

3. The significance of each subnetwork detected by the greedy search algorithm is assessed using both a topology-independent bootstrapping approach (p_{abund}), and a topology-dependent bootstrapping approach (p_{struct}).

4. The subnetworks determined to be significant ($p_{abund} \leq 0.05$ and $p_{struct} \leq 0.05$)

are reported to the user (Note: the threshold for significance can be adjusted through command-line parameters). The pathways are ranked by p_{abund} values.

3.3 Results

3.3.1 Performance Evaluation Using Simulated Datasets

In order to validate our methods, we have designed a simulated metagenomic study and compared the results with three previous approaches: (i) identifying active subnetworks using simulated annealing and greedy search [56]; (ii) discovering significant individual reactions using Metastats [133]; and (iii) finding differentially abundant KEGG defined pathways, an approach widely used in metagenomic functional comparison [43, 69, 130]. We choose these tools because they are addressing similar biological problems. However they do not exactly solve the problem in this paper, which is finding differentially abundant pathways. Here the goal of this simulated study is to show that our problem can not be solved by directly applying methods previously developed in a related context.

We designed a simulated metagenomic study in which five subjects are created for each of the two groups with distinct phenotypes. To generate the artificial reaction abundance matrix (where rows represent reactions and columns represent subjects), for each reaction a normal distribution is created, whose mean is randomly chosen from real metagenomic datasets (obese and lean twins in our study). The variance is calculated by setting the relative standard deviation (standard deviation divided by the mean) to 0.2. If we define a reaction to be equally abundant between two populations, then a random abundance value is generated from the same normal distribution for each subject. Otherwise, if a reaction is defined to be significantly enriched in one population, then another normal distribution is created for this reaction by increasing the mean such that the p value of the t-test for the two distributions is less than a predefined value (0.05 and 0.01 were used in our study). In this study, we have chosen a series of reactions (length 5 or 10) to be enriched in one population. The goal is to compare different methods in recovering these significant reactions based on the simulated abundance matrix. Biologically, the enriched pathways indicate functional enrichment of certain biological processes in a microbial community.

The receiver operating characteristic (ROC) curve is plotted for each method (Fig. 3.3). MetaPath outperforms all other methods dramatically showing the advantage in finding small significant subpathways. The most commonly used approach – comparing KEGG defined pathways – performs the worst in our simulation study (Fig. 3.3).

3.3.2 Comparing Obese and Lean Gut Metagenomes

We used MetaPath to compare the abundances of the metabolic networks of the gut microbiome in lean and obese subjects, relying on data from [130]. This metagenomic dataset comprises 6 samples from obese subjects and 6 samples from lean objects. The sequences are annotated and mapped to KEGG reactions using BLASTX (E value $<$ 10-5, bitscore $>$ 50, and %identity $>$ 50), resulting in total

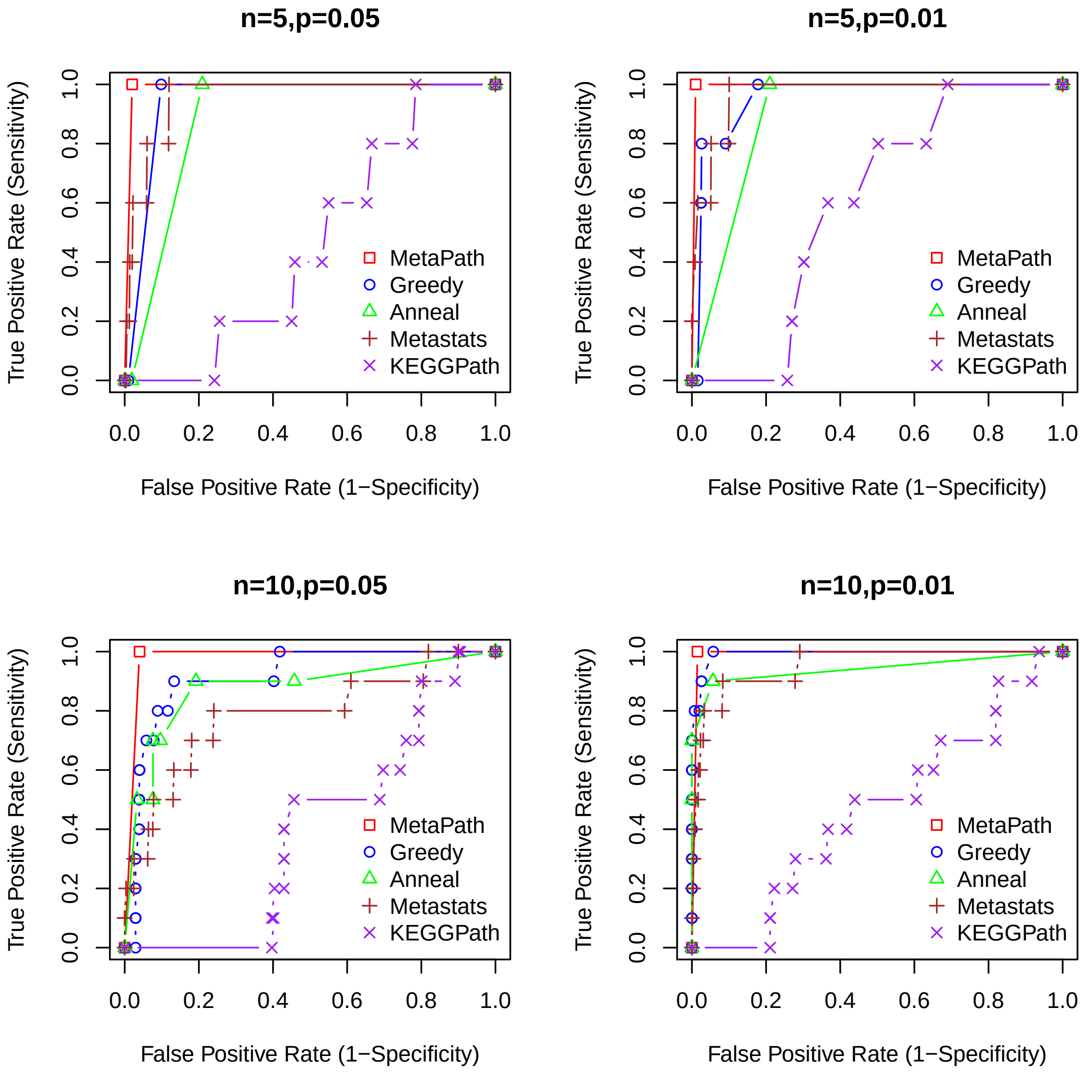


Figure 3.3: Comparison of statistical methods of discovering significant reactions in simulated datasets. Four methods are evaluated: discovering active subnetworks using simulated annealing (Anneal) and greedy search (Greedy) [56], discovering significant individual reactions using Metastats [133], finding differentially abundant KEGG-defined pathways (KEGGPath), and MetaPath. Four datasets are created by varying the number of significant reactions n and their significance value p.

1832 unique reactions within the 12 metagenomic samples. First, we computed p values and q values using Metastats to find differentially abundant reactions. Using a p value cutoff of 0.05, 92.7±9.1 (meanstandard deviation) reactions are significant including 37.1±6.6 and 55.6±3.1 enriched reactions in obese and lean groups, respectively, based on 10 runs of Metastats. The high variance of the number of significant genes can be primarily explained by two reasons: (1) some reactions are slightly below or above significance (0.05), thus p values computed through bootstrapping will jump between being considered significant and nonsignificant (Figure 3.4); (2) there are large variances of the abundance values within individuals in a same phenotypic group. In addition to p values, Metastats also provides an estimate of the False Discovery Rate (q value), information that is not used by MetaPath.The q values for all reactions are 1 (except R01676 where q=0.73), i.e. a literal interpretation of Metastats results would indicate no pathways are significantly different between the two populations. This result can be explained by the flat distribution of the p values (Figure 3.4), from which the q values are estimated. This observation highlights the limitation of relying on the false discovery rate, which requires the estimation of the proportion of features that are truly null [122], approach that does not perform well when only few features are truly significant.

We, then, applied MetaPath to this dataset, and have found 9 differentially abundant subnetwork (Figure 3.5) using 0.05 cutoff value for both p_{abund} and p_{struct}. All these subnetworks are enriched in obese subjects; none was found to be enriched in lean subjects. These 9 significant subnetworks contain 48 unique reactions, 22 of which are significant. It is worth pointing out that the number of significant reactions varies between different runs of statistical permutations (using Metastats) as shown above, but the significant pathways identified by Metapath stay the same (Figure 3.5). This observation confirms that the results from MetaPath are more

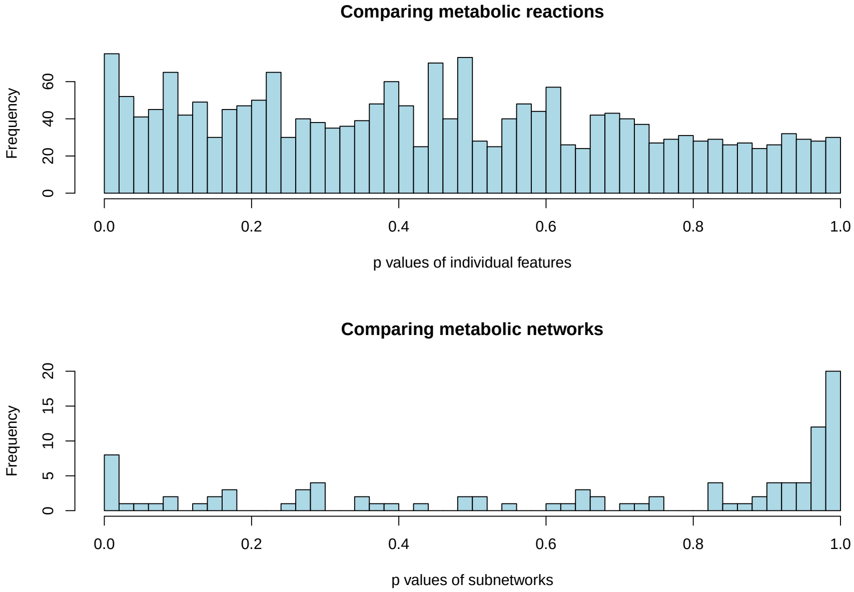


Figure 3.4: p values distributions from comparing individual metabolic reactions by Metastats and from comparing metabolic networks by MetaPath. The top histogram is the distribution of the p values of individual metabolic reactions calculated by Metastats. The Bottom histogram is the distribution of the p_{abund} values of the subnetworks calculated by MetaPath.

robust in the presence of noise in the data than the gene-by-gene approach. In the p values distribution of subnetworks (Figure 3.4), most of them are either very significant or insignificant and very few are around the p value cutoff, allowing the users to easily interpret the results.

Five subnetworks (Figure 3.5a-3.5e) are completely contained in the KEGG Fatty Acid Biosynthesis pathway, which consists of catabolic processes that can generate energy and primary metabolites from fatty acids. Our findings are consistent with previous observations and biochemical analysis in microbiota transplantation experiments in germ-free mice [129], where the concentrations of short-chain fatty acids in the caeca of obese mice are higher than lean mice, suggesting that the gut

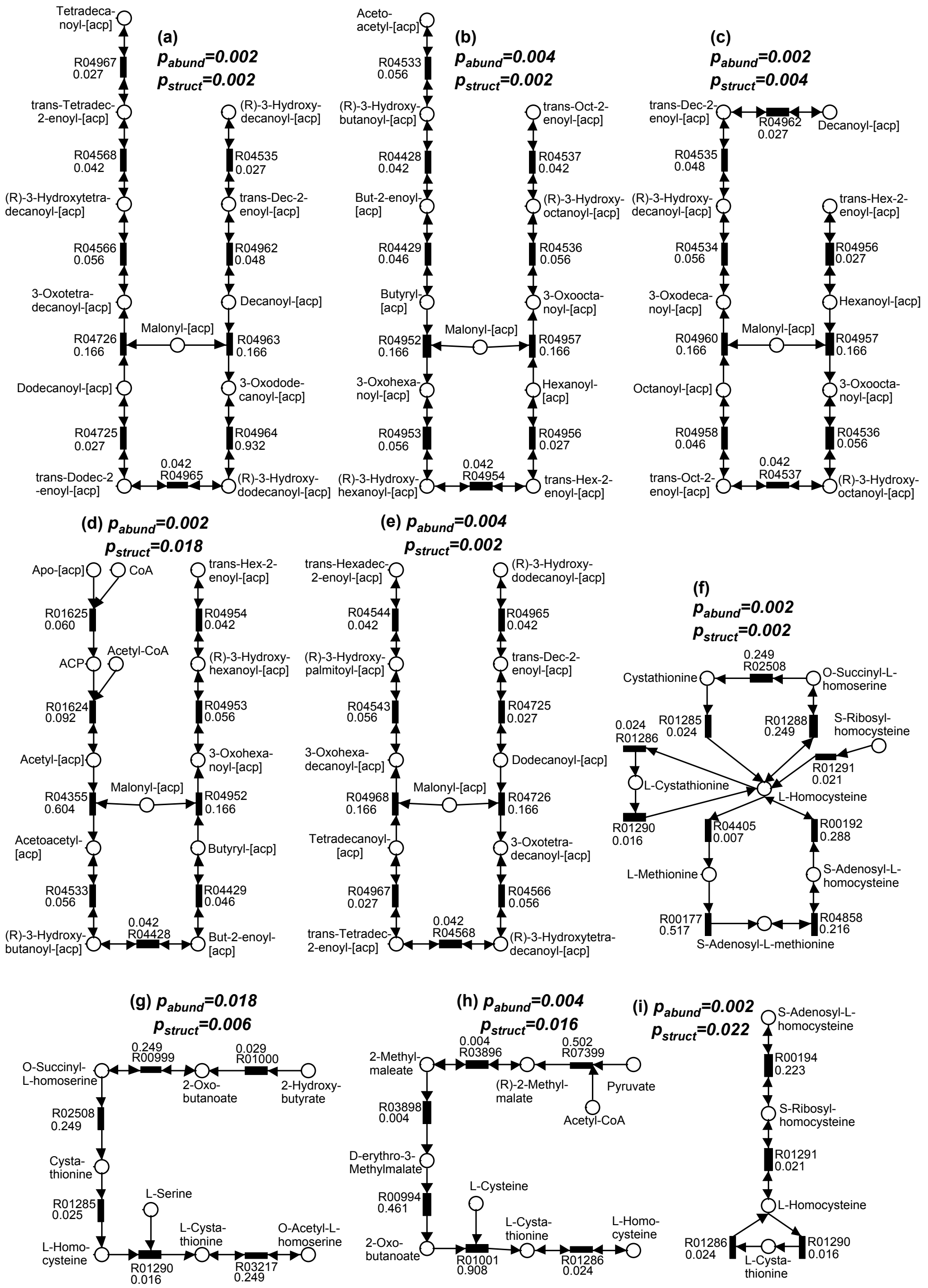


Figure 3.5: 9 statistically significant subnetworks are found in the comparison of the gut microbiome from the obese and lean subjects. All these subnetworks are enriched in the obese subjects. p_{abund} and p_{struct} significance values are shown above each subnetwork. p values for each reaction are shown with the KEGG reaction number. Five pathways (a)-(e) belong to the Fatty Acid Metabolism pathway in KEGG. Four pathways (f)-(i) contain the L-Homocysteine molecules.

microbiome in obese subjects has an increased capacity for dietary energy harvest.

Another interesting significant subnetwork consists of 10 reactions (Figure 3.5f), of which 8 belong to Cysteine and Methionine Metabolism and 2 belong to Sulfur Metabolism. Many reactions in this subnetwork are connected by the L-Homocysteine molecule. In addition, three other subnetworks (Figure 3.5g-Figure 3.5i) we discovered further confirm its potential involvement in obesity, because all these three pathways contain Lhomocysteine as metabolite. It is well-known that a high level of blood serum homocysteine is a risk factor for cardiovascular disease [40], and obesity - an increasingly prevalent metabolic disorder - is closely associated with heart disease [30]. Significant correlations between plasma homocysteine concentrations and obesity have been previously reported [51, 88]. The finding of increased potential for homocysteine metabolism within the obese gut microbiome provides an interesting hypothesis for future studies that, the gut microbiome may either have a direct role in the elevation of homocysteine levels in plasma, or may indirectly affect the hepatic biosynthesis of this amino-acid in the human body.

3.3.3 Comparing Infant and Adult Gut Metagenomes

A second data-set comprises gut microbiome samples from 4 infants and 9 adults individuals which were sequenced by [69]. The sequences were annotated and mapped to the reactions of KEGG pathway using BLASTX (E value $<$ 10-8, hit length coverage $\geq 50\%$ of a query sequence), resulting in total 1781 unique reactions. Based on 10 runs of Metastats, 383.7±1.56 reactions are significant using p value cutoff of 0.05 and 167.2±2.7 reactions are significant using a q value cutoff of 0.05. including 268.7 ± 1.56 and 115 ± 0 reactions enriched in infant and adult subjects respectively. Using a q value cutoff of 0.05, 167.2 ± 2.7 reactions are significant, including 133.2 ± 2.7 and 34 ± 0 reactions enriched in infant and adult subjects respectively. Compared with the previous dataset (obese and lean twins samples), the predictions of significant reactions are much more consistent across different permutations.

Applying MetaPath to search for significant subpathways, we have found 6 subpathways (Fig. 3.6a-3.6f) enriched in infant subjects and 4 subpathways (Fig. 3.6g-3.6j) enriched in adult subjects. These 10 significant subpathways contain 55 unique reactions, including 38 significant reactions and 17 reactions not found significant by Metastats. Three subpathways (Fig. 3.6a,c,d) enriched in infant subjects involve the metabolite L-homocysteine, which is consistent with previous observation that breastfed babies have an higher plasma homocysteine level possibly caused by suboptimal availability of folate in breast milk [39]. The concentration of folate is negatively correlated with that of homocysteine, as folate is a necessary coenzyme for reactions that metabolize homocysteine. In addition, babies normally have high protein diet, which may also cause the concentration of homocysteine to increase. A second pathway in Fig. 3.6e involves substrates citrate and succinate, and is closely related with oxidative tricarboxylic acid (TCA) cycle. TCA cycle is part of carbohydrate metabolism and can convert carbohydrates into usable energy in aerobic organisms. Because the adult gut ecosystem is dominated by strict anaerobes, it is reasonable to find this subpathway enriched in infant individuals where the gut microbiota also includes aerobes. This finding is consistent with results obtained by comparing COG functional categories [69]. We also find a subpathway Fig. 3.6f belonging to atrazine metabolism to be enriched in infant subjects. Atrazine is one of the most widely used herbicides, and it contaminates water and soil throughout the world. Our finding possibly indicates a side-effect of this contamination.

The pathway in Fig. 3.6i (enriched in adult subjects) is part of the lipopolysaccharide biosynthesis. Lipopolysaccharides are a building block of the outer membrane of Gram-negative bacteria. The enrichment of pathway Fig. 3.6i in adult subject may be a result of the fact that Gram-negative bacteria are also enriched in adults. Specifically, Bacteroides, a genus of Gram-negative bacteria, are a major constituent of adult gut microbiome, but not highly prevalent in infants. Fig. 3.6h and Fig. 3.6j (enriched in adult) are pathways related with pyrimidine metabolism. The metabolites RNA, cytidine and uridine, which are contained in pyrimidine metabolism, are normally obtained from high RNA food such as organ meats, broccoli, and brewers yeast, which are not available to unweaned infants, as they are not present in high abundance in milk. The pathway in Fig. 3.6g (enriched in adult) is part of fructose and mannose metabolism a pathway related to carbohydrate metabolism. This is also consistent with COG-based analyses indicating that many mono- or disaccharides metabolism genes are enriched in adults [69], explained by the fact that colonic microbiota in adults uses indigestible polysaccharides as resources for energy production and biosynthesis of cellular components.

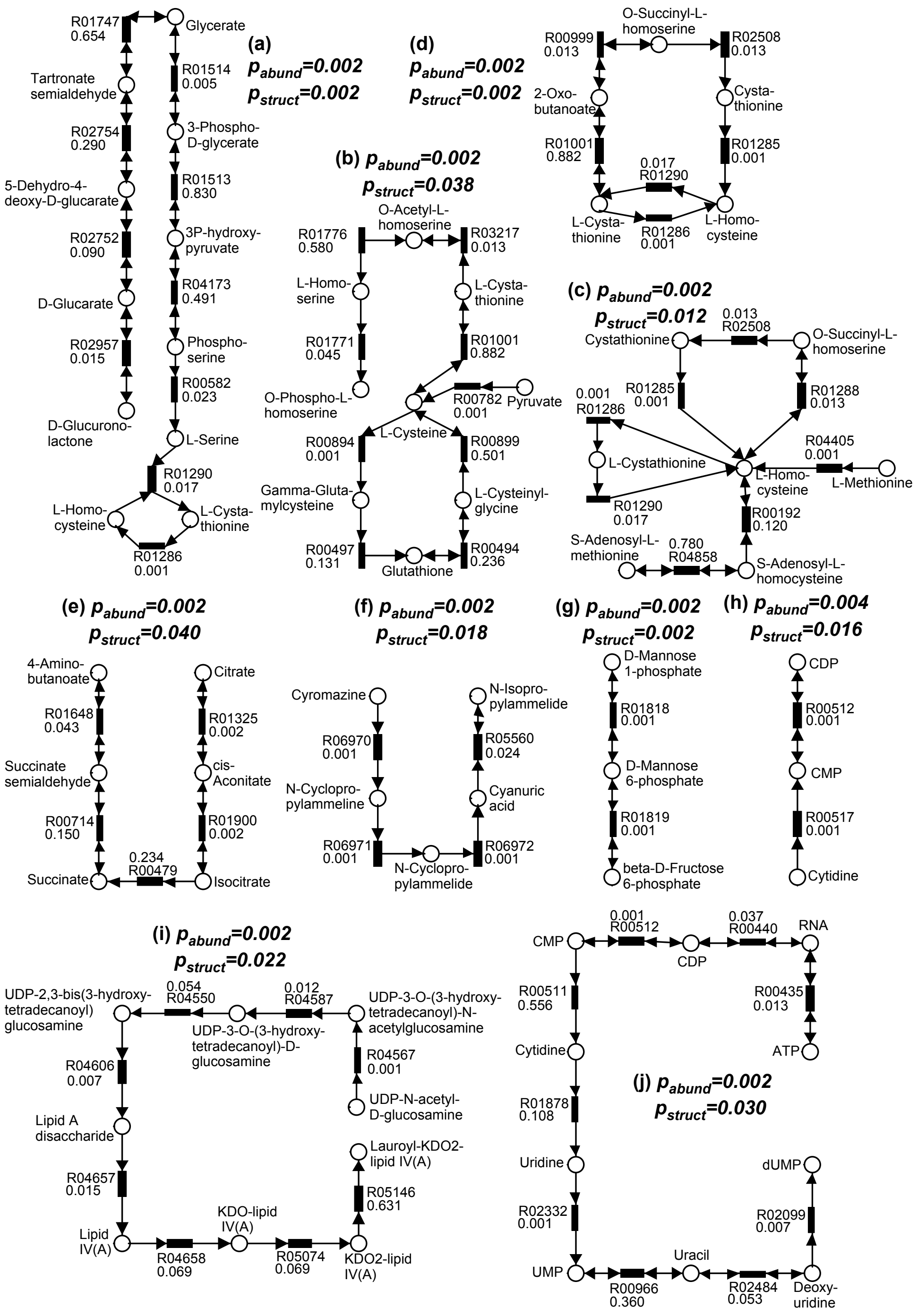


Figure 3.6: 10 statistically significant subpathways are found in the infant and adult individuals dataset. 6 subpathways are enriched in the infant subjects (a)-(f), and 4 subpathways are enriched in the adult subjects (g)-(j). p_{abund} and p_{struct} significance values are shown above each pathway. p values for each reaction are shown with the KEGG reaction number.

3.4 Conclusions

We have introduced a statistical method for finding significant metabolic subpathways from metagenomic datasets. Compared with previous methods, results from MetaPath are more robust to noise in the data, and have significantly higher sensitivity and specificity (when tested on simulated datasets). When applied to two publicly available metagenomic data-sets the output of MetaPath is consistent with previous observations and also provides several new insights into the metabolic activity of the gut microbiome. Finally, MetaPath is efficient: a typical metagenomic dataset and the corresponding metabolic network (about 2000 edges) can be analyzed in half an hour on a single processor.

While showing promising results, our methods have several limitations that we plan to address in the near future. First, and foremost, we restrict ourselves to pathways of a fixed length a restriction necessary for accurately computing the null distribution of pathway scores. This can severely affect our ability to discover long pathways whose abundance differs only slightly, but significantly, between samples. Second, we currently estimate gene abundances by simply counting the number of sequencing reads that map to a certain gene. Such an approach ignores differences in the length of genes, potentially leading to incorrect conclusions. We plan to address this issue by incorporating a recently-published [117] method that can accurately correct for genelength effects. The software described in this paper is freely-available under an opensource license from http://cbcb.umd.edu/ boliu/metapath/

Chapter 4

Antibiotic Resistance Genes Database

ARDB has been published in [81]. The methods, algorithms and experiments in this study originated from discussions between Dr. Mihai Pop and me. I developed the program and performed the experiments. Dr. Mihai Pop and I write the paper together.

4.1 Introduction

The discovery of penicillin in 1928 by Alexander Fleming has revolutionized the treatment of bacterial infections. The large-scale use of antibiotics, however, has also led to an increase in the number of microbes that can resist treatment. Drug resistant bacteria are an increasing threat to public health, as highlighted by a recent estimate that in the US methicillin-resistant Staphylococcus aureus (MRSA) may contribute to more deaths than HIV [5]. Methicillin-resistant strains of S. aureus were initially documented in the 1960s [6] and have been associated with higher mortality rates [26, 114] than their drug-sensitive counterparts. Similar challenges are posed by the emergence of multidrug- and extensively-drug resistant tuberculosis (MDR-TB and XDR-TB, respectively) [113]. Antibiotic resistance can result from large genomic changes, such as the acquisition of entire plasmids or mobile elements encoding resistance factors. Recent studies are, however, revealing the important role small mutations play in the evolution of resistance. For example, only 35 point-mutations distinguish a vancomycin-resistant strain of S. aureus from its sensitive counterpart, and these mutations evolved in just 3 months within an infected patient [89]. Furthermore, antibiotic resistance genes have the potential to be used for bioterrorism purposes through genetically modified organisms. These factors emphasize the urgent need for a better understanding of the mechanisms through which bacteria develop resistance, as well as for the development of new techniques for the rapid identification of resistance factors. The database presented in this article provides a first component of an informatics infrastructure aimed at enabling such studies.

Several mechanisms have been characterized through which bacteria become resistant to antibiotics [2]: (i) the production of enzymes that digest/metabolize the antibiotic; (ii) efflux pumps that eliminate the drug from the cell; (iii) modifications to the cellular target of the antibiotic that prevent binding; (iv) activation of an alternate pathway that bypasses drug action; and (v) particularly for gram-negative bacteria, down-regulation or elimination of transmembrane porins through which drugs enter the cell. The annotation information commonly associated with genes deposited in public databases is insufficiently detailed for representing this variety of resistance mechanisms and the additional meta-information relevant in this context. Specifically, each resistance gene is associated with a resistance profile (set of antibiotics or classes of antibiotics targeted by the gene), yet this information is usually not available. Second, resistance often requires the cooperation of multiple genes, usually within a same operon (e.g. vancomycin resistance VanA operon requires seven genes [23]), while most annotation information is targeted at individual genes. Finally, resistance frequently results from modifications to, or the disruption of an individual gene (e.g. modifications of the drug target), information incompatible with standard annotation procedures. Consequently, specialized resources are necessary for annotating and cataloging information related to antibiotic resistance.

Several recent efforts have been made to partially unify this information, such as Antibiotic Resistance Genes Online (ARGO) [111], MvirDB [140] and a compendium of TEM β-lactamase genes at the Lahey Clinic (www.lahey.org/Studies/). All, however, have limited functionality. ARGO only contains part of β-lactamase, vancomycin and tetracycline resistance genes. In addition, it does not include rich annotation information such as resistance profile, mechanism of action, operon information or gene sequence. Furthermore, many of the links between ARGO and GenBank target incorrect records (e.g. links to a genome instead of the relevant gene record). MvirDB is a broad repository of virulence-associated genes, including toxins, virulence factors and antibiotic resistance. The latter information is simply a replicate of the ARGO database. The Lahey Clinic website is a comprehensive collection of TEM type β-lactamases, which attempts to standardize the nomenclature for these genes. In addition to these specialized resources, antibiotic resistance information can be extracted in a restricted manner from GenBank and SwissProt, databases that lack many important types of information relevant in this domain.

4.2 Database contents and construction

To address the limitations of currently available public resources, and to facilitate the identification and characterization of antibiotic resistance genes, we have created a manually curated database (Antibiotic Resistance Genes Database (ARDB)) unifying most of the publicly available genes and related information. Our motivations in creating ARDB are (i) to provide a centralized compendium of information on antibiotic resistance; (ii) to facilitate the consistent annotation of resistance information in newly sequenced organisms; and (iii) to facilitate the identification and characterization of new genes. We believe this resource will be found useful by a broad range of scientists, including microbiologists, clinicians and the bio-defense research community.

The diversity of antibiotic resistance genes, types and mechanisms, combined with the fact that related information, such as resistance profile, is mostly 'paper-bound' made the construction of ARDB both difficult and time-consuming. To compile, confirm and validate this collection of data, several textbooks and several hundred journal articles were searched and summarized.

The majority of protein and nucleic acid sequences of known antibiotic resistance genes were retrieved from the NCBI nucleotide and protein databases and additional sequences were retrieved from the Swiss-Prot database. Genes were grouped into resistance types based on their protein sequence similarity using the following approach. First, the sequence of an experimentally confirmed representative was identified for every type of resistance, based on literature searches and meta-information provided by the NCBI protein database. These representative resistance genes were then used to fish out additional homologues using similarity searches against the NCBI nr database. The similarity cutoff was set at 80% unless a different value was recommended in the literature for a specific resistance type. Using this approach we identified 13 254 protein sequences putatively involved in antibiotic resistance. We filtered this set by removing vector sequences, synthetic constructs and redundant genes, resulting in a non-redundant set of 6206 proteins. This set was further refined by removing incomplete sequences, thereby yielding a core set of 4554 antibiotic resistance proteins. Each sequence was associated with corresponding CDD, COG, ontology and source organism information. Furthermore, the genes were grouped into resistance types, corresponding to clusters of genes with similar resistance profiles, operon membership and mechanism of action. In addition, basic information about known antibiotics was extracted from KEGG DRUG, PubChem, PubMed MeSH database and the Chemical Entities of Biological Interest (ChEBI) ontology. Although ARDB is mainly targeted at antibiotic resistance genes, 12 additional drug targets have also been included into ARDB with relevant information (16S rRNA (16), 23S rRNA, gyrA (17), gyrB, parC, parE, rpoB, katG, pncA, embB, folP, dfr), whose modification has been shown to confer resistance.

The data flow for the curation process is highlighted in Figure 4.1. ARDB is implemented as a MySQL relational database. Access to this database is provided through a CGI-based web interface.

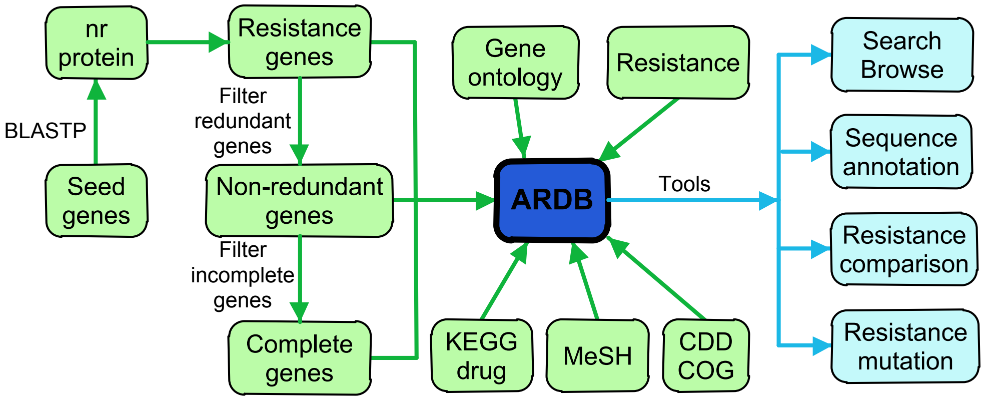


Figure 4.1: ARDB pipeline, including the construction process and the services provided. The initial seed genes are manually curated from published literature.

4.3 Ontology information

No comprehensive ontology is currently available for annotating antibiotic resistance information. To facilitate the computational analysis of antibiotic resistance information we have created a set of ontology terms aimed at characterizing both the resistance profile conferred by a specific gene and its specific mechanism of action. Specifically, for every antibiotic X, we have created a set of X resistance terms. Furthermore, we classify several mechanisms of action, including drug target modification, replacement or protection, drug enzymatic destruction and drug transport. Drug transport is further subclassified into ATP-binding cassette (ABC) drug efflux, major facilitator superfamily (MFS) drug efflux, small multidrug resistance (SMR) drug efflux and resistance-nodulation-cell division (RND) drug efflux, following the terminology used in [49]. These terms are defined within an Antibiotic Resistance (AR) ontology and are associated with each record present in our database. We are currently working with the broader ontology community to further refine this information and integrate it within existing ontology development efforts.

4.4 Data access and data mining

Users can access our database through a web interface at http://ardb.cbcb.umd.edu. This interface provides several modes of interaction as highlighted below.

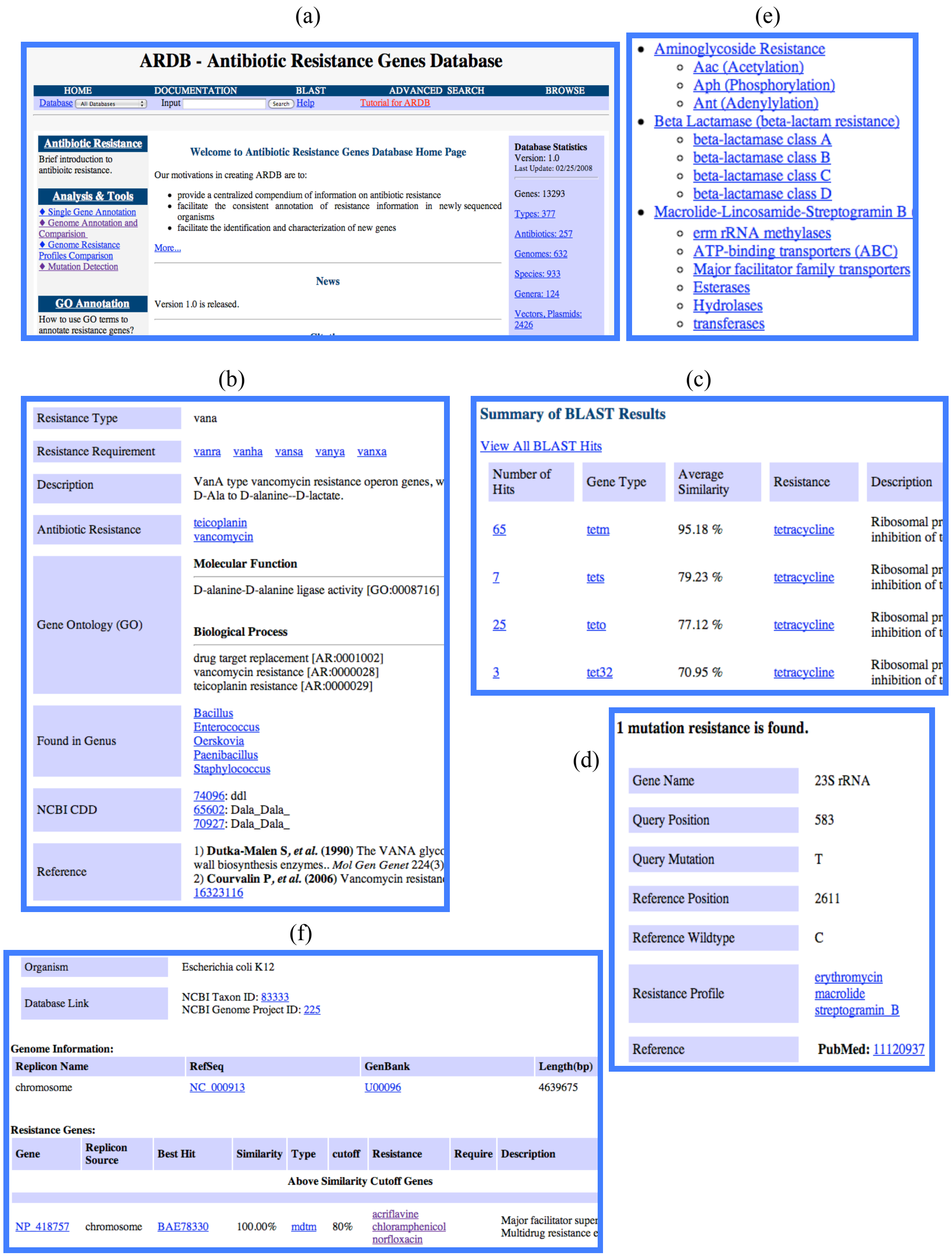
4.4.1 Keyword searches

Simple keyword search is available at the top of each page of ARDB website (Figure 4.2a), providing a quick means for searching a specific object in our database (gene, type, antibiotic, genome and genus) (Figure 4.2b to 4.2f). Users can search all of the data, or narrow down the search to a specific type of information. For example, users interested in the molecular mechanisms of resistance to tetracycline can search for the keyword 'tetracycline' within the 'Resistance Type' database. An advanced search function is also available, allowing users to select from among the available keywords associated with each database field.

4.4.2 Similarity searches

4.4.2.1 BLAST

To help identify and annotate antibiotic resistance genes, a BLAST interface is also provided. One or more sequences can be provided to this interface in a multi-FASTA file, corresponding to a set of gene sequences. Furthermore, both nucleotide and amino-acid sequences are accepted by our system. The results can be visualized as standard BLAST output, however additional displays are provided that are



(a)

ARDB - Antibiotic Resistance Genes Database

HOME DOCUMENTATION BLAST ADVANCED SEARCH BROWSE

Database All Databases Input Search Help Tutorial for ARDB

Antibiotic Resistance

Brief introduction to antibioitc resistance.

Analysis & Tools

- ♦ Single Gene Annotation
- ♦ Genome Annotation and Comparision
- ♦ Genome Resistance Profiles Comparison
- ♦ Mutation Detection

GO Annotation

How to use GO terms to annotate resistance genes?

Welcome to Antibiotic Resistance Genes Database Home Page

Our motivations in creating ARDB are to:

- provide a centralized compendium of information on antibiotic resistance
- facilitate the consistent annotation of resistance information in newly sequenced organisms
- facilitate the identification and characterization of new genes

More...

News

Version 1.0 is released.

Database Statistics
Version: 1.0
Last Update: 02/25/2008

Genes: 13293

Types: 377

Antibiotics: 257

Genomes: 632

Species: 933

Genera: 124

Vectors, Plasmids: 2426

(e)

- Aminoglycoside Resistance
  - Aac (Acetylation)
  - Aph (Phosphorylation)
  - Ant (Adenylylation)
- Beta Lactamase (beta-lactam resistance)
  - beta-lactamase class A
  - beta-lactamase class B
  - beta-lactamase class C
  - beta-lactamase class D
- Macrolide-Lincosamide-Streptogramin B
  - erm rRNA methylases
  - ATP-binding transporters (ABC)
  - Major facilitator family transporters
  - Esterases
  - Hydrolases
  - transferases

(b)

Resistance Type	vana
Resistance Requirement	vanra vanha vansa vanya vanxa
Description	VanA type vancomycin resistance operon genes, w D-Ala to D-alanine--D-lactate.
Antibiotic Resistance	teicoplanin vancomycin
Gene Ontology (GO)	**Molecular Function** D-alanine-D-alanine ligase activity [GO:0008716] **Biological Process** drug target replacement [AR:0001002] vancomycin resistance [AR:0000028] teicoplanin resistance [AR:0000029]
Found in Genus	Bacillus Enterococcus Oerskovia Paenibacillus Staphylococcus
NCBI CDD	74096: ddl 65602: Dala_Dala_ 70927: Dala_Dala_
Reference	1) **Dutka-Malen S, *et al.* (1990)** The VANA glyco wall biosynthesis enzymes.. *Mol Gen Genet* 224(3) 2) **Courvalin P, *et al.* (2006)** Vancomycin resistan 16323116

(c)

Summary of BLAST Results

View All BLAST Hits

Number of Hits	Gene Type	Average Similarity	Resistance	Description
65	tetm	95.18 %	tetracycline	Ribosomal pr inhibition of t
7	tets	79.23 %	tetracycline	Ribosomal pr inhibition of t
25	teto	77.12 %	tetracycline	Ribosomal pr inhibition of t
3	tet32	70.95 %	tetracycline	Ribosomal pr inhibition of t

(d)

1 mutation resistance is found.

Gene Name	23S rRNA
Query Position	583
Query Mutation	T
Reference Position	2611
Reference Wildtype	C
Resistance Profile	erythromycin macrolide streptogramin_B
Reference	**PubMed:** 11120937

(f)

Organism	Escherichia coli K12
Database Link	NCBI Taxon ID: 83333 NCBI Genome Project ID: 225

Genome Information:

Replicon Name	RefSeq	GenBank	Length(bp)
chromosome	NC_000913	U00096	4639675

Resistance Genes:

Gene	Replicon Source	Best Hit	Similarity	Type	cutoff	Resistance	Require	Description
				Above Similarity Cutoff Genes				
NP_418757	chromosome	BAE78330	100.00%	mdtm	80%	acriflavine chloramphenicol norfloxacin		Major facilitator super Multidrug resistance e

Figure 4.2: Sample web pages from ARDB. (a) Front page, (b) resistance type, (c) blast result, (d) mutation annotation, (e) browse and (f) genome information.

specific to antibiotic resistance information. Our 'ARDB annotation format' groups individual BLAST hits according to resistance type as inferred from the level of similarity to the genes within the database associated with a specific type of resistance (Figure 4.2c). A second view allows users to download a tab-delimited spreadsheet summary of the antibiotic resistance genes identified within the uploaded file.

4.4.2.2 RPS-BLAST

In addition to BLAST we also provide an RPSBLAST interface relying on Position Specific Scoring Matrix (PSSM) created from sequences associated with each resistance type, using an approach similar to the NCBI Conserved Domain Database. The output of this interface is similar to that provided by the BLAST interface mentioned above.

4.4.2.3 Polymorphism detection

Additionally, a mutation-specific search function is provided to identify polymorphisms previously characterized to confer resistance (Figure 4.2d). For example, a G-C mutation at position 1058 of the Escherichia coli 16S rRNA has been shown to confer resistance to tetracycline [108]. This information is extracted from the detailed BLAST alignment between the query sequence and a reference sequence in our database. Currently this function is available for 12 genes (16S rRNA, 23S rRNA, gyrA, gyrB, parC, parE, rpoB, katG, pncA, embB, folP, dfr), and we expect to extend it as more information becomes available in the literature.

4.4.3 Pre-annotated information

The antibiotic resistance profiles of 632 complete bacterial genomes have already been annotated and deposited in ARDB allowing quick search. This information can be conveniently extracted through keyword searches against the genome database, or through the Genome Resistance Profiles Comparison link from the front page. The latter approach allows users to summarize and compare the resistance profiles of multiple organisms present in our database.

4.4.4 Browse

A 'browse' function is available that allows the users to visualize several classes of antibiotic resistance genes, grouped by their resistance profile. This functionality is currently available for aminoglycoside, β-lactam, macrolidelincosamidestreptogramin B, multidrug transporter, tetracycline and vancomycin resistance (Figure 4.2e).

4.4.5 Submission

In order to facilitate community-driven refinement of our database we provide an interface through which users can submit information about novel resistance genes. This interface captures several types of information not commonly available in other databases [Minimum Inhibitory Concentration (MIC), resistance type, ontology, citation information, etc.]. Furthermore we provide a simple file format and upload functionality to facilitate the submission of information for multiple genes. The information received will be vetted and inserted into the database. We are also planning to develop an interface that allows community-deposited information to be directly added to the database as provisional records, pending additional manual curation.

4.4.6 Example: mining resistance genes with ARDB

Recent studies, through a functional metagenomic screening approach, have shown that many environmental microflora are potential antibiotic resistance reservoirs even without much exposure to antibiotics (e.g., human microbiome [121] and soil [3, 24]). We annotated four oral metagenomic samples including two controls and two periodontal diseases (Figure 4.3), and found that the most common resistance factors are for tetracycline, including genes previously encountered in the oral microbiome tetM, tetW, tetO, tetQ, and tetS [115], as well as gene tet37 that confers tetracycline resistance through unknown mechanisms and was originally cloned from oral samples [27]. It is important to note that antibiotic resistance genes are universally found in both cases and controls and cannot be exclusively linked to periodontal disease, although the relatively abundances could be different.

4.5 Conclusion

The database described in this article, ARDB, unifies most of the publicly available antibiotic resistance genes and provides a reliable annotation service to researchers investigating the molecular basis for resistance in bacteria. Because of the large diversity and the rapid identification of new resistance genes, the current version of ARDB is just a first catalog of currently available information, and will continue to be updated over the coming months and years. We plan to coordinate

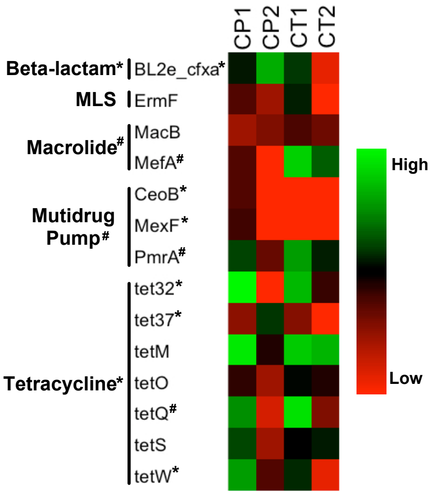


Figure 4.3: The abundances of antibiotic resistance genes identified in four oral microbiome samples, including two controls (CT1 and CT2) and two periodontal diseases (CP1 and CP2). The heatmap is colored according to the log_2 of the number of reads (per 10 million metagenomics reads) mapped to each resistance gene at 95% similarity cutoffs. '*' or '#' indicates that a particular resistance gene is significantly enriched ($p \leq 0.05$, Fishers exact test) in case or control samples, respectively. MLS represents macrolide, lincosamide and streptogramin B antibiotics. It is important to note that antibiotic resistance genes are universally found in both cases and controls and cannot be exclusively linked to periodontal disease, although the relatively abundances could be different.

our development efforts with researchers actively involved in antibiotic resistance research as well as with the developers of biological ontologies and of databases storing related information (such as virulence factors or toxins). As part of these efforts we aim to refine the structure of our database, better determine the types of information stored and identify additional requirements for the user interface.

Future efforts will also target the development of new approaches for cataloguing and characterizing polymorphisms correlated with resistance, as well as for annotating changes to cellular regulatory networks that underlie the mechanisms of drug tolerance.

Chapter 5

Deep sequencing of the oral microbiome reveals signatures of periodontal disease

This project is accomplished through extensive collaborations with several external research groups, including Salomon Amar, Daniel Segre, Simon Kasif from Boston University and Colin Stine from University of Maryland, School of Medicine. The methods used in this study and results generated from this study have been published in [82]. My specific contributions to this project are mainly data analyses about taxonomic diversity and composition Figure 5.1 and Figure 5.3, statistical comparison of metabolic networks Figure 5.2, genome assembly (Table 5.2, Table 5.3 and Figure 5.4) and genomic variation analysis Figure 5.5. I was also extensively involved in the writing of the paper.

5.1 Background

Understanding the role of microbial communities in human health is emerging as one of the most important and fascinating biomedical challenges of our times [76, 128]. Our body harbors an enormous amount of microbial cells, estimated to exceed the number of human cells by an order of magnitude. These microbes are organized into complex communities specifically adapted to inhabit different niches of the human body, such as the skin, and the respiratory, gastrointestinal, and urogenital tracts. Such ecosystems carry a broad range of functions indispensable for the wellbeing of the host. At the same time, the rise of pathogens within such communities, causing infection and inflammation, constitutes an ongoing challenge in biomedical research. This is especially true in light of the slow rate at which new antibiotics are discovered, and the increase in the number of microbes that can resist treatment. In contrast to the traditional view of individual pathogens being responsible for disease onset, recent microbial ecosystem diversity analyses seem to point to a new perspective in which the transition from health to disease is attributed to a shift in the global balance of the microbial flora rather than to the specific appearance of individual pathogens. However, the mechanisms that underlie the connection between disease or infection and the dynamics of the host-associated ecosystems are still poorly understood.

In this work, we focus on the role of the oral microbial ecosystem in periodontal disease. Periodontal disease is the most common infectious disease affecting tooth-supporting structures. Left untreated, periodontitis can lead to, or aggravate existing systemic conditions such as cardiovascular disease, diabetes, pulmonary diseases, and obesity. In dentistry, understanding the changes in the oral microbiome that foretell the early stages of periodontitis and dental caries, the most prevalent chronic oral diseases, may allow the better diagnosis and treatment before the appearance of the telltale clinical manifestations of these diseases (such as tissue damage in periodontal pockets or dental hard tissue loss). The emergence and evolution of antibiotic resistance in periodontal pathogens has affected the therapeutic success rates for this disease. New approaches are urgently needed to help regain control over periodontal disease, and microbiome studies offer a promising new angle of attack. Unraveling the complex interactions that define the oral microbiome is a fundamental, but complex component of this endeavor.

Recent developments in systems biology make it possible to perform quantitative modeling of genome-scale metabolic networks for individual microbial species [31] and have been recently extended to explore small microbial consortia, possibly paving the way for future quantitative studies of the microbiome. However, at the ecosystem level, current modeling efforts and quantitative analyses are heavily limited by the unavailability of relevant data. Towards this goal, increasingly accessible metagenomic sequencing approaches hold the promise to enable a global systemic view of the human oral microbiome. Recent advances in sequencing technology are enabling scientists to generate billions of nucleotide bases at a fraction of the cost per base of traditional methods. This deep sequencing has revealed an unexpectedly high diversity of the human oral microbiome: dental plaque pooled from 98 healthy adults comprised about 10,000 microbial phylotypes [62] - an order of magnitude higher than the previously reported 700 oral microbial phylotypes as identified by cultivation or traditional cloning and sequencing. The total diversity of the global oral microbiome can be estimated to be around 25,000 phylotypes. To date, however, we do not know how many of these microbes contribute to periodontal disease, what metabolic functions are key players in the transition from health to disease, or how common or exclusive are the oral microbiomes of unrelated healthy individuals.

Here we combine the collection of whole-community sequencing data with a number of computational analyses to provide a snapshot of the microbial component of periodontal disease at a high resolution. Specifically, we collected subgingival plaque samples from healthy and periodontally affected patients and subjected them to 16S rRNA analysis and deep sequencing in order to explore their microbiome. Our analyses reveal a number of trends in genomic diversity and biological function enrichment during disease that allow us to formulate a novel hypothesis on the nature of periodontal disease. We also demonstrate the power of high-throughput sequencing approaches by reconstructing an unculturable member of the TM7 group, complementing an initial analysis that relied on single cell genomic approaches. We also characterize several regions of variation within one of the dominant members of the oral cavity, Actinomyces naeslundii. This paper describes a genomic and metabolic examination of the differences between the healthy and diseased periodontal microbiome.

Note that this project is accomplished through extensive collaborations with several external research groups, including Salomon Amar, Daniel Segre, Simon Kasif from Boston University and Colin Stine from University of Maryland, School of Medicine. My specific contributions to this project are mainly data analyses about taxonomic diversity and composition Figure 5.1 and Figure 5.3, statistical comparison of metabolic networks Figure 5.2, genome assembly (Table 5.2, Table 5.3 and Figure 5.4) and genomic variation analysis Figure 5.5.

5.2 Results and Discussion

5.2.1 A deep look at the oral microbiome in health and disease

Current knowledge of the composition and functional spectrum of the human oral microbiome is limited by the difficulty to culture the majority of microbes that populate the oral cavity. We used deep sequencing technology to overcome this limitation, and produce a substantial genomic dataset for the human microbiome under health and periodontal disease conditions. Specifically, we generated both 16S rRNA and whole metagenomic data from five subjects (3 periodontally healthy [H] and 2 chronic peridontitis [P] patients, see Table 5.1). A total of 495,195 16S rRNA sequences were generated with the 454 FLX sequencing technology, yielding an average of 30,000 sequences per sample after removing low-quality sequences (roughly 3-times more sequences per sample than generated in a recent survey of oral microbes [10]). A total of 272,709,876 sequence reads were generated using the Illumina GAII platform, 76bp, paired-end run (mean library size 207bp) from the whole metagenome of four of the above-mentioned subjects (2 H and 2 P). The low quality nucleotides were trimmed from all sequences and fragments matching to the human genome reference (NCBI release GRCh37.p1) were removed from further analysis. The level of human DNA contamination varied between different samples averaging about 87% of the sample, i.e. the oral microbiome represents just one eighth of the entire dataset or a total of 33,681,771 (12.4%) sequences (Table 1). This level of contamination is consistent with that observed in other studies, such as the Human Microbiome Project (see hmpdacc.org and upcoming manuscript).

Despite the moderate yield (in terms of fraction of microbial sequences in the dataset) our results show that valuable biological insights can be derived from the data, thus indicating that informative and clinically relevant whole-metagenomic analyses of the oral microbiota can be conducted in a cost-effective manner.

Phenotype	Subject (tooth)	Clinical	16S rRNA		Shotgun		% Human
			# Reads	Sample	# Reads	Sample	
Disease	1(14)	advanced	51,056	P11	9,725,937	P1	68.86
	1(19)	moderate	20,149	P12			
	1(30)	moderate	41,355	P13			
	2(30)	moderate	46,444	P21	4,893,057	P2	81.98
Healthy	3(1)	healthy	23,702	H11	12,357,917	H1	60.61
	3(2)	healthy	44,869	H12			
	3(3)	healthy	32,405	H13			
	3(4)	healthy	56,116	H14			
	4(3)	healthy	6,205	H21	6,704,860	H2	89.78
	4(14)	healthy	35,356	H22			
	4(19)	healthy	14,110	H23			
	4(30)	healthy	25,662	H24			
	5(3)	early	12,295	H31	NA	NA	NA
	5(19)	healthy	30,891	H32			
	5(30)	healthy	12,605	H33			

Table 5.1: Summary of sample information including high-quality read counts, taxonomic assignment of most abundant genus in each sample, and level of human contamination. The clinical labels represent: ‘healthy’ - healthy periodontal pocket; ‘early’ - early periodontal disease (bleeding under probing but no attachment loss), ‘moderate’ - moderate periodontal disease; ‘advanced’ - advanced periodontal disease.

5.2.2 Beyond the taxonomical view of periodontitis

The standard view of periodontitis, largely based on traditional microbiological approaches, associates the disease with the rise and damaging action of a small set of well-characterized pathogens. A first question we wanted to address using our data is whether, and to what extent, this traditional view still holds from the vantage point of metagenomic sequencing. Taxonomic profiling of the samples, whether derived from targeted 16S rRNA sequencing or from whole-metagenomic data (WGS) (see Methods and Figure 5.1) reveals a community dominated, on average, by the bacterial phyla Firmicutes, Actinobacteria, Bacteroidetes, Fusobacteria and Proteobacteria, consistent with previous studies [1,10]. Together, these groups account for 80-95% of the entire oral microbiome. At the genus level we identify a total of 55 distinct genera in the 16S rRNA data and 58 distinct genera in the WGS data that are present at an abundance of 0.1% or higher (an additional 73 and 62 rare genera can be found in the 16S rRNA and WGS data, respectively). The most abundant genera comprise previously characterized oral bacteria: Actinomyces, Prevotella, Streptococcus, Fusobacterium, Leptotrichia, Corynebacterium, Veillonella, Rothia, Capnocytophaga, Selenomonas, Treponema, and TM7 genera 1 and 5.

The TM7 division was prevalent in our samples (11 out of 15 samples contain this division at $> 2\%$ abundance), averaging 5.7% (standard deviation 7.2) of the entire population in the 16S rRNA data (WGS-based estimates also range about 6%), and up to 26.8% in sample P11. This division was statistically enriched in diseased samples ($p \leq 0.05$, Metastats [133], Figure 1). TM7 is a novel candidate bacterial division with no cultivated representatives, and previous studies have shown microbes from this division to be commonly found in the human oral flora but at relatively low abundance, generally around 1% of the population [12,98], though abundances as high as 8% were previously reported [138]. The high abundance of TM7 microbes

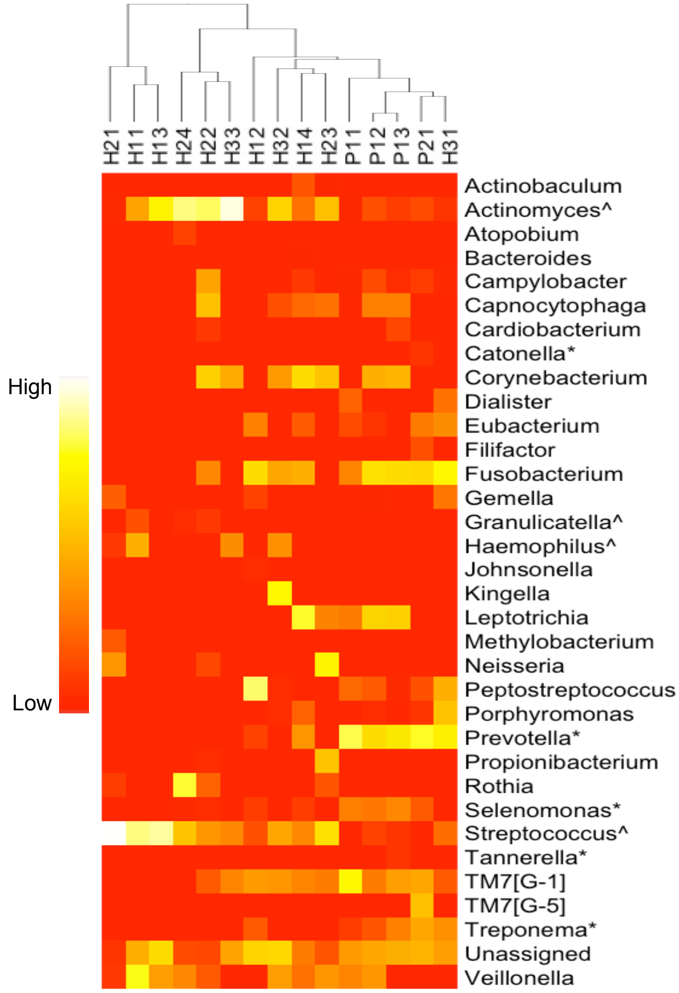


Figure 5.1: Relative abundance of genera in the samples. ˆ - genus significantly enriched in cases; $^{\#}$ - genus significantly enriched in controls ($p \leq 0.05$, Metastats [133]). Only genera with >1% abundance in at least one sample included. Colors reflect relative abundance from low (red) to high (white). Sample H31 (control) clusters together with the diseased samples, confirming clinical observations of early symptoms of periodontal disease (Table 5.1).

present in our samples, and their correlation with periodontal disease, indicate that the prevalence of this poorly studied bacterial division within the oral cavity, and its role in disease, have yet to be fully appreciated.

When comparing healthy and diseased samples we observe a shift in the composition of the oral microbiota (see Figure 5.1), supporting the well characterized transition ($pvalue < 10^{-15}$ using Fisher's exact test) from a gram-positive dominated community in the healthy samples, to a gram-negative dominated community in periodontal disease. On one hand, our findings recapitulate prior results that indicate that the gram-negative genera Selenomonas, Prevotella, Treponema, Tannerella, Haemophilus and Catonella are significantly enriched in periodontal disease. Further, we have found a set of gram-positive genera that are significantly enriched in healthy samples: Streptococcus, Actinomyces, and Granulicatella. Surprisingly, however, neither Fusobacterium, nor Porphyromonas were found to be significantly more abundant in the periodontal disease samples, despite being previously implicated in this disease. This is likely due to the high variance in the abundance of these organisms across our samples, as well as the small sample size which affects our statistical power.

Clustering analysis 5.1 reveals sample H31 (a control) to have a microbiota most similar to the diseased samples. This observation prompted a careful analysis of the clinical data collected during sampling. The data revealed some symptoms of mild periodontal disease (such as bleeding at probing time, see Materials and Methods for more details) that were not found in any of the other healthy samples, indicating that the microbiota may shift into a disease state before the full clinical symptoms of the disease are apparent. Also note that the diseased samples (including H31) cluster together tightly while the healthy samples are more widely distributed. This phenomenon is discussed in more detail below.

Taxonomic enrichment, however, cannot fully explain the etiology of periodontal disease. All organisms that exhibit an enrichment in either healthy or diseased samples are present in all the samples, irrespective of disease status, i.e. the mere presence of pathogens in the periodontal pocket is not sufficient to trigger periodontitis. The disease might be correlated with the presence of specific virulence factors within the genomes of particular pathogens, or might be initiated once the abundance of one or more pathogens crosses a specific threshold. The mechanisms that keep pathogenic bacteria in check during health but allow them to bloom during disease are not yet understood. These observations support our suggestion that a full understanding of periodontal disease requires whole-genome and whole-system analyses.

5.2.3 Metabolism, virulence factors and drug and metal resistance as disease signatures

In addition to providing a taxonomic overview, our metagenomic sequencing data contain high-resolution functional information. We annotated the function of genes identified in the assembled whole-metagenome data according to the KEGG Orthology, and used the resulting data to compare the functional potential of the oral microbiome in health and disease. The metabolic profiles of healthy and diseased samples differ in a number of important ways (Figure 5.2). The diseased microbiome is enriched in metabolic functions that are consistent with a parasitic lifestyle made possible by the availability of nutrients derived from the degradation of host tissue and from bacterial cells destroyed by the host immune response. Among these are functions for fatty acid metabolism and acetyl-coenzyme A degradation, aromatic amino acid degradation, ferrodoxin oxidation, and energy-coupling factor (ECF) class transporters. The periodontal pocket has been previously shown to be enriched for such nutrients in patients with periodontitis [18]. Several of these metabolic functions have also been associated with an intracellular lifestyle (e.g. fatty acid metabolism [32], or with anaerobic metabolism (e.g., ferrodoxin oxidation, and acetyl-CoA degradation), highlighting the diversity of survival strategies employed by the microbes inhabiting the periodontal pocket during disease. Also enriched in disease are a number of virulence factors such as the presence of conjugative transposons, type IV secretion systems, and the biosynthesis of toxic factors (e.g., acetone, butanol, and ethanol biosynthesis), as well as the Lipid-A of lipopolysaccharide (LPS) biosynthesis. LPS is a group of molecules known to trigger host immune response and inflammation and their enrichment in disease provides a possible explanation for the systemic impact of periodontitis on the human host.

Finally, the periodontal disease samples are enriched in a number of functions related to drug and metal resistance (mercury, cobalt-zinc-cadmium). Mercury resistance has been previously characterized as a common feature of oral bacteria, even in the absence of mercury-containing amalgam, and is frequently associated with antibiotic resistance [97]. The role drug resistance plays in disease is, however, unclear as antibiotic resistance factors are present in both healthy and diseased samples.

Comparatively, only a few pathways are significantly enriched in the healthy microbiome (or depleted in the diseased microbiome), including pathways for fatty

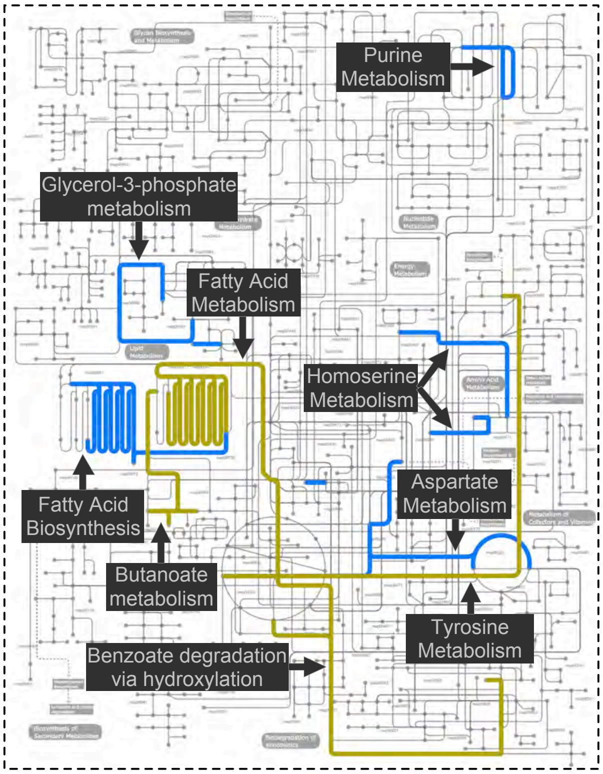


Figure 5.2: Metabolic pathways that are significantly enriched (found by MetaPath [84]) in healthy samples (dark blue; $p < 0.05$), and that are significantly enriched in diseased samples (dark yellow; $p < 0.05$). Figure is constructed with iPath [75].

acid biosynthesis, purine metabolism, and glycerol-3-phosphate metabolism. Certain fatty acids have been shown to have a protective role in periodontal health [63] and it is possible that some of these are synthesized by the healthy microbiota. However, most of what is known about the role of fatty acids in periodontal health is based on nutritional studies and the contribution of the oral microbiota has yet to be characterized. Glycerol-3-phosphate is a lipid metabolite that has been shown to occur in higher concentration in periodontal disease samples [7]. Our study hints that a possible explanation for this observation is a decrease in the ability of the disease microbiome to metabolize this compound. Also enriched are genes related to homoserine metabolism, possibly related to quorum sensing functions within the healthy microbiome, as homoserine lactones are frequently used as quorum sensing molecules in oral bacteria. The enrichment, within our healthy samples, of the reactions downstream of homo-serine lactone pathway may indicate a fully functioning quorum sensing system, allowing for the communication between organisms that is the hallmark of a healthy biofilm system. In poly-microbial biofilms it has been shown that mutants lacking quorum sensing molecules, while able to construct biofilms, are unable to obtain the correct structure and thickness. The depletion of pathways related to quorum sensing in our diseased samples may indicate a possible cause of disease progression due to the inability of the healthy microbiome to maintain a protective biofilm.

5.2.4 A systems level perspective on oral disease

The functional characterization reported above suggests that, beyond the taxonomic details, one can identify ecosystem-level signatures of periodontal disease consistent with its clinical manifestations. However, from the above analysis, it is still not clear whether these signatures reflect isolated instances of disease-related molecular processes, or fit into a coherent picture of the disease as a predictably different state of the whole oral microbial flora. We addressed this question by performing additional analyses at different levels of resolution, and found that a major systemic change seems to be identifiable between the healthy and diseased microbiomes. The diseased samples harbor a more diverse microbial community (as measured by the Shannon diversity index, Figure 5.3A), yet clustering analysis at the taxonomic level (Figure 5.1 and Figure 5.3B) and in terms of enzyme content (Figure 5.3C), as well as pairwise comparisons of individual healthy and diseased samples based on tetramer (subsequences of length 4) frequencies (Figure 5.3D), all indicate that disease samples are more similar to each other than the healthy samples. In other words, the diseased state appears to be associated with a constrained and predictable region in the space of all possible states a microbiome can take. Thus, although the periodontal disease microbiomes are more diverse in terms of community structure, that structure is quite similar across different patients. In contrast, the healthy microbiome in any individual patient has relatively lower taxonomic diversity, but its exact composition differs significantly across patients.

5.2.5 De novo assembly of oral microbes

The analyses we presented above have focused either exclusively on organisms (16S rRNA diversity) or biological function (metabolic analysis), thus ignoring the important link between organisms and the functions they perform. This connection can only be made by reconstructing partial or entire organisms from the community through metagenomic assembly. Currently, no practical genome assemblers exist that are specifically designed for large-scale metagenomic assembly, thus we relied on a hybrid assembly approach that combined de novo assembly using SOAPdenovo (assembler used in a recent metagenomic analysis of gut microbes [101]), and alignments against a collection of oral microbes (see Methods). The results shown in Table 5.2 demonstrate the power of this hybrid approach, which leads to an average

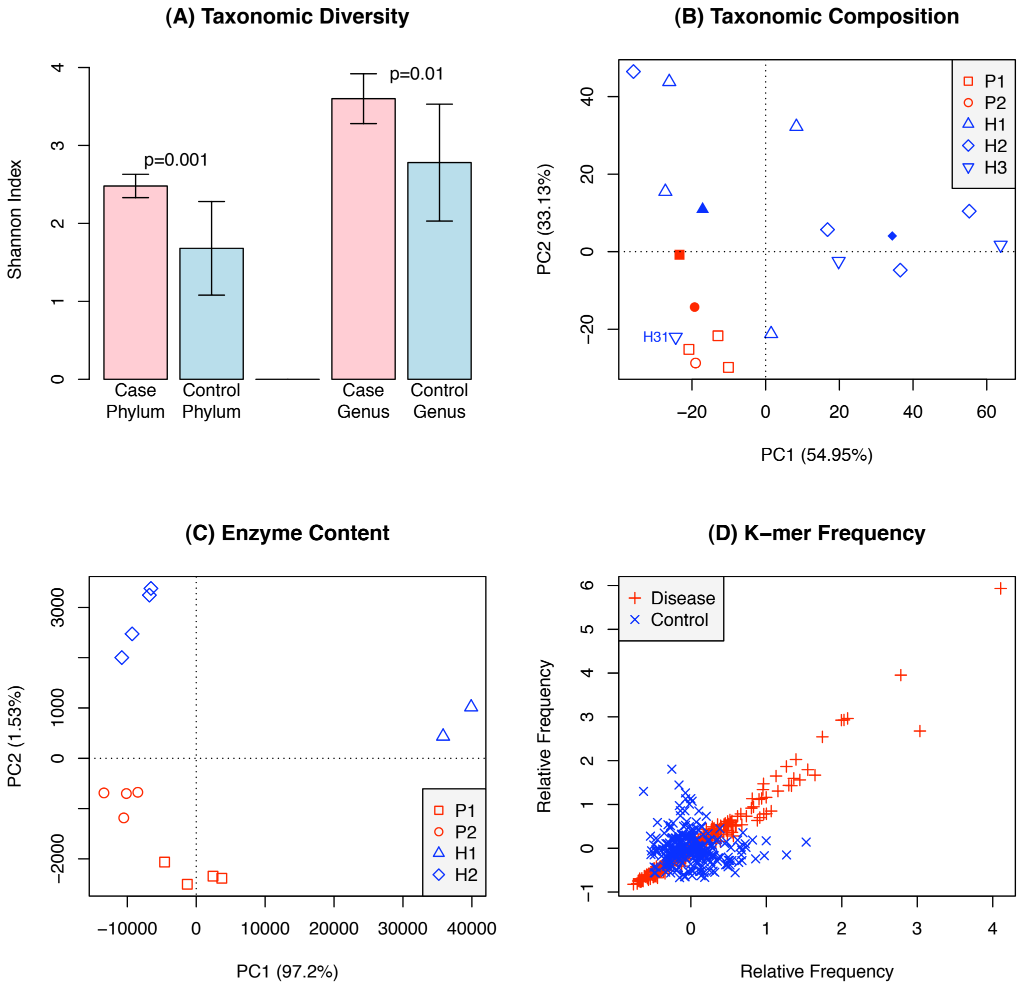


Figure 5.3: Systems-level analysis reveals the disease state to be an attractor in the space of possible states for the microbiome. A - Shannon diversity is significantly higher in diseased samples (community is more diverse). B - Disease samples cluster together in PCA analysis of the taxonomic composition of the samples. Sample H31 (tooth with incipient periodontal disease from an otherwise healthy patient) appears in the top right corner, clustering together with the disease samples; C - Disease samples cluster together in the PCA analysis of the enzyme content of the samples; D - Comparison of tetramer relative frequencies indicates disease samples are more similar to each other than controls. Results in panel C and D are generated from Daniel Segre group.

of 4.4 and 2.1 times larger (in terms of N50 contig size) assemblies than de novo assembly and comparative assembly, respectively. Despite the relatively low level of coverage in our data, we obtain fairly contiguous assemblies (average N50 contig size of 3.5Kbp), and are able to assemble up to about 50% of the total number of reads in our data-set. Furthermore, consistent with our previous observation that the periodontal disease samples are more diverse, the corresponding assemblies are also more fragmented (average N50 contig size is 1.2Kbp in diseased samples versus 5.8Kbp in healthy samples). In addition, a pooled assembly of all four samples results in dramatically increased contig sizes (max contig size is 16.9Kbp in pooled assembly versus 7.6Kbp in individual assemblies), indicating these samples contain closely related organisms.

Sample	Assembly	# contigs	Length(Mbp)	Max(Kbp)	N50(bp)	N90(bp)	# reads assembled	
							#(M)	pct
P1	SOAPdenovo	22,226	11.8	12	583	368	1.2	12.45
	Comparative	26,464	16.7	16	1113	598	1.3	13.21
	Hybrid	37,213	24.6	16	1829	1025	2.3	23.42
P2	SOAPdenovo	12,966	6.3	3.3	352	0	6.7	14.23
	Comparative	13,841	8.5	35.2	490	0	5.7	11.69
	Hybrid	21,835	12.5	37.6	647	396	10.5	21.39
H1	SOAPdenovo	45,658	3.1	22.6	3042	1648	5	40.2
	Comparative	46,036	3.3	18.6	2437	1559	3.5	28.21
	Hybrid	63,688	5.1	19	7567	3953	6.7	53.18
H2	SOAPdenovo	18,048	10.6	12.7	616	352	1.7	25.51
	Comparative	16,107	13.6	26.8	1543	689	2.2	32.33
	Hybrid	20,339	17.6	110	3934	1099	3.1	45.88
Pool	SOAPdenovo	98,051	54.9	15.7	2035	1342	8.1	24.12
	Comparative	63,506	60.1	44.6	8415	5474	8.4	24.89
	Hybrid	115,718	93.4	229.8	16896	9245	13.4	39.87

Table 5.2: Assembly statistics of metagenomic shotgun reads for contigs that are ¿= 300bp using (1) SOAPdenovo, (2) comparative assembly and (3) a hybrid approach that uses MINIMUS to combine the contigs from the previous two methods. 'Pool' represents the assembly of all four samples together. N50 or N90 is defined as the contig length such that equal or longer contigs produce 50% or 90% of 10Mbp.

5.2.6 Assembly of a TM7 genome

As described above, we detected a higher presence of TM7 organisms in our samples than previously reported in literature. TM7 is a novel candidate bacterial phylum without cultivated species, and previous studies have shown its high prevalence in human oral flora but with very low abundances [12, 98]. The first sequence of a TM7 organism (TM7a) was generated through single-cell isolation in a microfluidic device, followed by whole genome amplification [85]. Due to the artifacts of the whole genome amplification approach, the resulting assembly is fairly fragmented (see row 1 in Table 3). Here we relied on a hybrid assembly approach to reconstruct a more complete version of this genome, using the corresponding shotgun sequences generated in our project. Briefly, we started with the pooled assembly of all our samples and extracted all contigs that are mapped to the previously sequenced TM7a genome, and scaffolded these contigs using Bambus 2 [65]. Finally, we merged our TM7 assembly with the previously published assembly, derived from single-cell sequencing, in order to construct the most complete (to date) assembly of an organism from the TM7 group. The final assembly is still highly fragmented, comprising over 1,500 contigs (Table 3), however it contains almost 50% more sequence than the single-cell derived assembly (2.3Mbp versus 1.7Mbp), and the N50 contig size is two times larger (790 bp versus 389 bp). These results highlight the power of combining single-cell and metagenomic approaches when reconstructing the genomes of unculturable organisms from metagenomic samples (5.4).

Assembly	# contigs	Length(bp)	Max(bp)	3Mbp	
				N25(bp)	N50(bp)
HOMD TM7a reference genome	1,780	1,691,166	17,479	1,898	389
Hybrid assembly	1,340	1,478,421	13,917	1,753	NA
Scaffolds	874	1,593,887	20,925	5,093	482
Combine reference	2,222	2,209,727	17,514	2,904	790
Scaffolds	1,593	2,310,536	33,748	7,178	1,751

Table 5.3: Assembly statistics (calculated on contigs ≥ 300bp) for HOMD TM7a reference sequences (row 1), hybrid assembly from metagenomic shotgun reads (row 2), Bambus scaffolding of hybrid assembly (row 3), assembly from combining hybrid assembly and the HOMD reference sequences (row 4), and Bambus scaffolding of 'combined reference' (row 5). The N25 and N50 are calculated assuming a 3Mbp genome size.

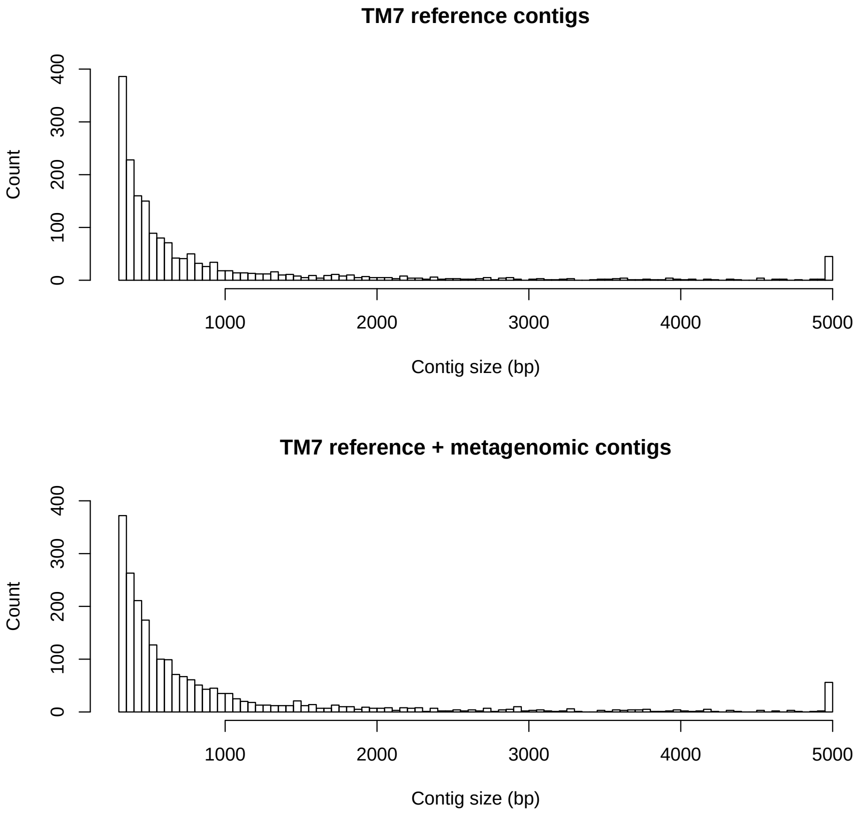


Figure 5.4: Distribution of contig sizes from TM7 reference genomes and the assembly from our metagenomic sample. The upper plot shows the distribution of contig size from TM7 reference genome, which is assembled from single-cell sequencing. The lower plot shows the distribution of contig size from the assembly that combines the contigs from TM7 reference genome and metagenome. Contig sizes that are >= 5000bp are plotted as 5000bp.

In addition, this improved TM7 genome assembly allows us to identify 703 genes that were not present in the original assembly (see Methods for details). In order to evaluate the additional information contained in these genes, we annotated them using the COMBREX [56] system (Supplementary Table 4). The analysis revealed several potential virulence genes including an EmrB/QacA family drug resistance transporter gene (Gene ID: 681_1) and two phage proteins (Gene IDs: 386_2 and 1828_4). These genes are not necessarily omissions from the original assembly, rather they could represent de novo insertions into the TM7 genome present in our sample. The set of novel TM7 genes does, however, included several housekeeping genes (e.g., 10 ribosomal protein genes not present in the original assembly) which should be conserved across TM7 genomes, thereby indicating that our assembly improves upon our current understanding of the structure of the TM7 genome in addition to revealing strain-specific genomic variants.

5.2.7 Genomic variation in Actinomyces naeslundii

Close analysis of one of the most abundant organisms in our samples (present at 24- and 6-fold coverage in samples H2 and H1, respectively), a relative of Actinomyces naeslundii MG1 (sequence ID SEQF1063 in the HOMD database), provides evidence for structural variations distinguishing this strain from the reference strain originally isolated from a patient with mild gingivitis. The average similarity between the assembled metagenomic contigs from our project and the reference sequence is 96.2% and 95.2% for samples CT1 and CT2, respectively (second and sixth ring in Figure 5.5). A number of genomic deletions with respect to the reference strain are apparent in our samples, several of which contain potential virulence factors. These differences could be explained by the fact that the reference genome was isolated from a patient with gingivitis, while in our samples the Actinomyces strains are predominantly associated with healthy samples. Most striking is a deletion at 2120 kbp containing a putative mobile element encoding a mercury resistance locus (including a mercury resistance gene, a site-specific recombinase, and an integrase). Mercury resistance is commonly found in oral bacteria, frequently associated with antibiotic resistance [97]. Interestingly, gene set enrichment analysis of the entire metagenomic data-set reveals an enrichment of mercury resistance genes in the diseased samples, possibly due to the association of these genes with virulence loci. Several other deletions also appear to encode virulence factors - a drug transporter (at position 580kbp in the reference strain) and an alcohol dehydrogenase gene (at position 165 kbp) further underscoring the difference between the pathogenic reference strain and the presumably commensal Actinomyces strains found in our samples. Another two deletions (at positions 20kbp, and 1010kbp) contain genes predicted to encode proteins involved in secretion and response regulation. These deletions occur at slightly different locations in the two samples we analyzed, suggesting they may be subject to rapid evolution.

Further evidence of the adaptation of Actinomyces to the oral environment is revealed by the analysis of single nucleotide polymorphism (SNP) densities. In Figure 5.5 (rings 6 and 8) we highlight the regions of the genome that have higher than

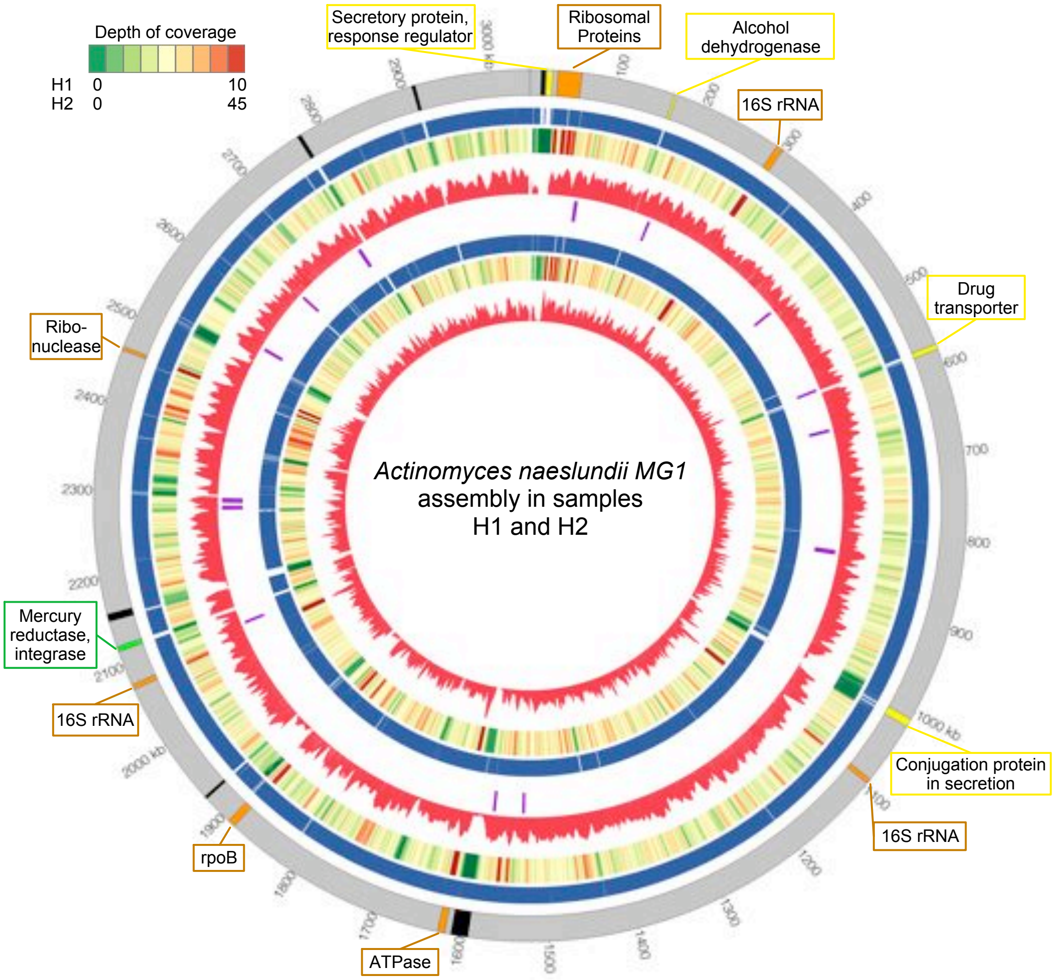


Figure 5.5: Systems-level analysis reveals the disease state to be an attractor in the space of possible states for the microbiome. A - Shannon diversity is significantly higher in diseased samples (community is more diverse). B - Disease samples cluster together in PCA analysis of the taxonomic composition of the samples. Sample H31 (tooth with incipient periodontal disease from an otherwise healthy patient) appears in the top right corner, clustering together with the disease samples; C - Disease samples cluster together in the PCA analysis of the enzyme content of the samples; D - Comparison of tetramer relative frequencies indicates disease samples are more similar to each other than controls. Results in panel C and D are generated from Daniel Segre group.

expected SNP densities (> 2 standard deviations from the mean). The most polymorphic regions correspond to genes known to be involved in the adaptation of an organism to its environment: transcriptional regulators, known to evolve rapidly in bacteria, and ABC transporters. Another highly-polymorphic region occurs within the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, a virulence-associated protein originally identified in Streptococci, which plays an important role in the colonization of periodontal pockets by interacting with plaque-forming bacteria. GAPDH was also shown to mediate the interactions between Streptococci and Porphyromonas gingivalis fimbriae, possibly contributing to the colonization of the subgingival pocket by P. gingivalis. These observations are consistent with previous findings of high-SNP densities within genomic regions surrounding recombination events.

5.3 Conclusion

Our study represents a important step towards characterizing the genomic composition of the microbial communities associated with periodontal disease. We have demonstrated that the subgingival microbiome can be effectively interrogated through high-throughput sequencing, and that the resulting data provide valuable insights into the molecular underpinnings of periodontal disease.

Despite a relatively small amount of bacterial sequence data recovered from our samples (primarily due to the high level of human DNA contamination), a combination of comparative and de novo assembly approaches was able to reconstruct large genomic segments from several dominant organisms in our samples, thereby allowing a better reconstruction of an unculturable TM7 organism (in conjunction with data generated through single cell genomic approaches), and providing a glimpse at the genomic variation (and possible association with virulence) within Actinomyces genomes. Better assemblies were possible in samples that were sequenced more deeply (e.g., sample H1), indicating the need to sequence the oral environment more deeply than has been done in this study. Furthermore, assembly quality roughly correlated with disease status, partly confirming our observation (based on 16S rRNA data) that diseased samples had a higher microbial diversity. This observation also highlights a limitation of existing assembly tools in dealing with genomic diversity, further underscoring the need for the development of metagenomic-specific genome assemblers.

The analysis of the TM7 and Actinomyces genomes revealed signatures consistent with recombination events possibly associated with virulence factors. Lateral transfer of virulence determinants through phages and recombination is well documented in the bacterial world, leading to a partial separation between function and phylogeny, thus, suggesting the need for metagenomic and functional analyses as a complement to taxonomic surveys of host-associated microbiota.

Taxonomic analyses of the data we generated are consistent with a well established community shift from a gram-positive dominated healthy microbiome to a gram-negative dominated diseased microbiome, which is also enriched in a number of oral pathogens. The molecular mechanisms that underlie and cause this transition are, however, unknown. Here we have shown that functional information derived from whole-metagenomic data provides a valuable complement to the taxonomic data and allows us to develop a novel theory of periodontal disease. The healthy state is highly regulated by the host immune system and interactions between community members to maintain a community dominated by few good microbes, usually gram-positive Actinobacteria or Streptococci. The transition to periodontal disease involves a disruption of the host-microbiome interactions that results in a more even community structure composed by a broad range of organisms that can thrive in the oral environment. The presence of pathogens within this community can lead to the clinical manifestations of periodontal disease, which in turn can lead to additional changes in the community due to the increased availability of nutrients released by the damaged tissue. As a result, the periodontal disease microbiome eventually settles into a state characterized by a diverse population of microbes adapted to a parasitic lifestyle made possible by the disrupted host homeostasis. One of the samples from our study was characterized by a microbiota typical of a diseased state, yet the corresponding tooth was just starting to show some of the clinical symptoms of disease. This observation implies that dysbiosis precedes the clinical manifestation of disease, and that the oral microbiota could be a potential tool for the early diagnosis of periodontitis.

The large variability we observe between healthy samples, and even between different teeth of a same person, highlights the limitation of using data derived from cross-sectional studies to define what the core normal microbiome means. Furthermore, case-control studies are likely insufficient to determine the causal agents of periodontal disease the organisms found to dominate the diseased microbiome (the usual suspects commonly described in the literature) may simply be a symptom of the disrupted subgingival environment rather than the primary cause of disease. The usual suspects approach considers presence and absence of specific bacteria to be the critical precondition for disease, however, our data support a more nuanced approach that considers quantitative and genomic differences as the critical factors when moving from health to a diseased state. Longitudinal studies are necessary to characterize the dynamic changes that occur in the oral microbiome in response to environmental changes (food intake, changes in the host, etc.) and to track the transition between the healthy and diseased states, and the return to health after treatment.

It is important to note that the analyses described above are a preliminary pilot project with limited sample size, and our observations must be confirmed in more extensive studies. Furthermore, we focus on whether the microbiome has the potential to perform certain biological functions, and on determining the relative fraction of the microbial population that can perform a particular function. These results (as well as those of similar metagenomic projects) must be complemented by experimental studies aimed at determining whether the biological processes statistically enriched in disease are actually active in the subgingival pocket.

As others have previously reported, and as observed in the data we have shown here, periodontal disease is the result of a disruption of the complex interactions occurring within the subgingival microbiome and between the microbiome and the host. A full understanding of the etiology of periodontal disease will only be possible through further in-depth systems-level analyses of the host-microbiome interaction.

5.4 Materials and Methods

5.4.1 Subject population

The subject population consisted of 5 patients who were in good general health and were recruited between August and November 2009 at the Clinical Research Center, Boston University Goldman School of Dental Medicine. Written informed consent was obtained from all enrolled individuals. The study protocol was reviewed and approved by the Institutional Review Board at the Boston University Medical Center. All subjects had at least 12 natural teeth with >20 years of age (age range, 28-45 years). Subjects diagnosed with chronic periodontitis (n=2) were selected among those who had at least six sites with probing depth 6 mm and attachment loss 5 mm. Subjects in the control (periodontally healthy) group had no pockets > 3 mm and no attachment loss > 2 mm at any site with no signs of periodontal inflammation characterized by bleeding on probing, redness, edema, and attachment loss, with the exception of subject 5 where one of the teeth (# 3, sample H31) exhibited mild bleeding at probing time consistent with initial periodontal disease. Sites with gingivitis were characterized by bleeding on probing, redness, edema, but no attachment loss (pocket depth 4 mm). Sites from chronic periodontitis subjects further characterized as mild, moderate or advanced periodontitis sites based on the pocket depth. Mild periodontitis was characterized with pockets > 4 mm but not more than 5 mm; moderate periodontitis was characterized with pockets > 5 but < 7 mm while advanced periodontitis was characterized with pockets > 7 mm. Healthy group consisted of subjects of Asian, Caucasian and African American origin, while periodontitis subjects were of Caucasian and African-American origin. Exclusion criteria included pregnancy, lactation, systemic conditions that could affect the progression or treatment of periodontal diseases. In addition, none of the subjects had received systemic antibiotics or periodontal therapy in the previous 6 months.

5.4.2 Data collection, sequencing and preprocessing

16S rRNA sequences were processed and filtered based on quality with an in-house pipeline as follows. First, sequences containing at least one unrecognizable base-pair ('N'), and that were too short ($<$ 75 cycles of the 454 instrument) were excluded from further analysis. Then, barcode sequences were deconvoluted and removed.

Metagenomic sequencing was performed on pooled DNA from multiple teeth in order to obtain sufficient DNA concentrations for library construction. Metagenomic shotgun sequences were obtained from the Illumina instruments in fastq format and were trimmed for quality using the FastX Toolkit (Hannon Lab, CSHL) with the following parameters: (1) minimum length 25, and (2) q-value cutoff 20. Sequences containing at least one ambiguity character ('N') were also removed. The remaining sequences that passed the quality trimming outlined above were mapped to the human genome reference (NCBI build 37 v 1) downloaded from NCBI using Bowtie with parameters (-v 3; at most 3 mismatches) If one of the sequences from a paired end matched the human genome, then both sequences were removed from the dataset. The remaining reads were mapped against the human sequences in the NCBI nr database using BLAST in order to remove human sequences not present in the NCBI human genome reference. For this additional check we required at least 95% global identity (since BLAST is a local alignment algorithm, our calculation also takes into account the length of the unaligned segments flanking the hit reported by BLAST).

5.4.3 Clustering and annotation of 16S rRNA sequences

The entire set of trimmed 16S rRNA sequences were clustered into Operational Taxonomic Units (OTUs) with the program DNACLUST [42], using a 1% radius (-r 2). To obtain the taxonomic identities, the OTU centers were aligned using BLAST to the RDP database [19] augmented with oral clones from the HOMD database [16], and were annotated using the lowest common ancestor approach (similar to the approach in [55]). The assignment process is conservative: (1) only sequence with at least global 98% identity with the reference is classified; (2) if there are more than one equally good best hits, then the sequence is classified using the lowest common ancestor approach; (3) otherwise it is classified as unknown. Finally, the taxonomic label of the OTU center is transferred to the sequences from the same OTU cluster. The resulting data was organized in a collection of tables at different taxonomic levels containing each taxonomic group as a row and each sample as a column. These tables formed the substrate for the further statistical analyses.

5.4.4 Assembly

We mapped and assembled the samples against reference sequences for oral microbes extracted from the Human Oral Microbiome Database (HOMD, http://www.homd.org) as follows:

1. We used MUMmer [70] (-maxmatch -l 20 -b) to map the individual reads against the HOMD reference database.

2. Reads that mapped with higher than 80% global identity were then assembled based on the mapped coordinates of the reads.

3. This process was repeated using a 90% similarity threshold, but mapping the reads against the assemblies generated at step 2, rather than against the HOMD database.

4. The resulting contigs were then combined with the results of a de novo assembly of the data, as described in more detail below.

We used SOAPdenovo V. 1.04 with parameters -K 23 and -M 3, as previously used by the MetaHit project to assemble gut microbiome data. Contig sequences longer than 100bp, which were generated by our customized comparative assembly pipeline and by SOAPdenovo, were combined and assembled using MINIMUS with the following parameters: (1) minimum overlap length 40bp, and (2) overlap error rate is 0.1.

5.4.5 Estimation of SNP rates and genetic diversity

After assembly, the shotgun reads were mapped back to the contigs using Bowtie [73] allowing at most 3 mismatches. To avoid sequencing and mapping errors we used a conservative approach as suggested in [68]: we only retained SNPs occurring in regions with a depth of coverage higher than 4, and with each individual haplotype represented in at least two different reads. The SNP rate was calculated using a 5Kbp window and a 100bp step size.

We adapted the approach used in [47, 59] to infer the genetic diversity θ from metagenomic shotgun sequencing data using composite likelihood estimators while accounting for a constant sequencing error rate. First we classified the nucleotide positions of the assembled contigs into k groups, where k is the maximum depth of coverage of the contigs, and positions within the same group have the same depth of coverage. The number of nucleotide positions in each group is denoted by $n_1, n_2, ..., n_k$. Considering the large number of bacteria in the sampled community relative to the number of reads sequenced, the probability that each read derives from a different individual microorganism is close to one. Thus, we have a population size equal to the depth of coverage at every site in the assembled contigs [59]. Consequently, the estimator can be obtained by calculating the expected number of true SNPs and false SNPs due to sequencing errors [47]. Then for a particular nucleotide group with the same depth d, assuming an infinite sites model, the expected number of segregating sites

$$\theta n_d \sum_{i=1}^{d-1} = S_d - S_d^{error} \tag{5.1}$$

where S_d is the observed number of segregating sites (SNPs) from data and S_d^{error} is the expected number of segregating sites induced by sequencing errors. For a nucleotide position with depth of coverage d, the probability of at least two mutations ($x >= 2$; considered as true SNPs) induced by sequencing error (e) is:

$$\sum_{x=2}^{d} B(d,e) = \sum_{x=2}^{d} \binom{d}{x} e^x (1-e)^{d-x} = 1 - (1-e)^d - ne(1-e)^{d-1} = e_d \qquad (5.2)$$

Since there are n_d such sites with depth of d, the expected number of segregating sites induced by error is $S_d^{error} = n_d e_d$. Hence the estimated $\hat{\theta}$ for regions with d depth of coverage is

$$\hat{\theta} = \frac{S_d - S_d^{error}}{n_d \sum_{i=1}^{d-1} \frac{1}{i}} \qquad (5.3)$$

Finally, summing over all groups we get the equation

$$\hat{\theta} = \frac{S - \sum_{d=4}^{k} S_d^{error}}{\sum_{d=4}^{k} n_d \sum_{i=1}^{d-1} \frac{1}{i}} = \frac{S - \sum_{d=4}^{k} n_d [1 - (1-e)^d - ne(1-e)^{d-1}]}{\sum_{d=4}^{k} n_d \sum_{i=1}^{d-1} \frac{1}{i}} \qquad (5.4)$$

In our calculation, we assume a constant sequencing error rate $e = 0.01$. The θ value is calculated using a 1Kbp window moving average (which is roughly the average gene size in bacteria) with 100bp step size. Regions of the genome that had a value of θ more than 2 standard deviations higher than the average were flagged as potential polymorphism hotspots.

Chapter 6

MetaCompass: Comparative Assembly of Metagenomic Sequences

The methods, algorithms and experiments in this study originated from discussions between Dr. Mihai Pop and me. I developed the program and performed the experiments. Dr. Mihai Pop and I write the manuscript together. At the time of writing this dissertation, the paper is under preparation.

6.1 Introduction

Microorganisms comprise the majority of Earth's ecological diversity, and they play important functional roles in virtually all ecosystems. Particularly, human-associated microbial communities play a critical role in health and disease [53]. In many environments, however, more than 99% of the bacteria cannot be cultured by standard laboratory techniques [126]. Metagenomics is a new scientific field that involves the analysis of organismal DNA sequences obtained directly from an environmental sample, enabling studies of microorganisms that are not easily cultured in a laboratory. Metagenomic studies, pioneered in the early 2000s, have recently increased in number and scope due to the rapid advances of high-throughput sequencing technologies, e.g., GS FLX system from 454 Life Sciences and Genome Analyzer from Illumina company, which permit large amounts of DNA to be sequenced quickly and cheaply. For example the MetaHit consortium has generated about 500G raw sequences from 124 human gut samples in its initial analysis [101], the Human Microbiome Project has produced more than 7 trillion base pair DNA reads [20], and the newly initiated Earth Microbiome Project is planning to sequence 200,000 samples with 6 billion base pairs per sample, totaling 1200 trillion bp DNA (http://www.earthmicrobiome.org/). There is a pressing need to extract meaningful biological information from the deluge of metagenomic sequences.

So far, one of the biggest challenges in analyzing next-generation sequences is the short read length. Although many sophisticated computational tools have been specifically developed for the short reads, e.g., MetaGene [92] for gene prediction, PhyloPythia [86] and PhymmBL [11] for taxonomic classification, the performance is still not very satisfactory in many applications. For example, the accuracy of PhymmBL (one of the most accurate classifier for genomic fragments) for 100bp reads is about 60% at the genus level. All these limitations, however, can be alleviated through metagenomic assembly. Genome assembly is the process of piecing together short DNA fragments, which are randomly extracted from a sample, to form a set of longer contiguous stretches of DNA strings called contigs, which form the foundation for many important downstream analyses, e.g., gene prediction, genomic variation discovery, etc.

In traditional single-genome projects, the most difficult part of genome assembly is handling repetitive regions (DNA repeats), which are the major cause of misassembly. The assembly of metagenomes, however, brings about additional assembly challenges in the form of non-uniform read depth due to non-uniform distribution of species abundance, multi-strain population (quasispecies) in the natural environment, and the potential for the coassembly of reads originating from different species (chimeric contigs). Hence, the performance of *de novo* assemblers is significantly compromised when assembling complex metagenomic samples. Despite these challenges, various *de novo* assemblers, based on either overlap-layout-consensus approach or De Bruijn graph, have been developed and applied to the assembly of metagenomes from massive amounts of short reads. For example, on average only about 40% of the reads in human gut metagenomic sample can be assembled into contigs that are longer than 500bp using SOAPdenovo [101]. The performance of these *de novo* assemblers [74, 79, 119, 139] is far from satisfactory when assembling metagenomic sequences. Mainly because most of the whole-metagenome shotgun projects are sequenced by Illumina technology producing a huge amount of short DNA reads, which are usually computational unmanageable using the traditional overlap-layout-consensus approach. Instead, they are assembled using De Bruijn graph based approach. The performance of De Bruijn graph assemblers relies on counting the abundance of k-mers. However, in metagenomic sample, usually the genome of one species is represented by millions of small variants of a consensus genome. Hence, this will potentially create many low-abundance k-mers and might significantly reduce the power of De Bruijn graph.

Instead of building another *de novo* assembler to compete with existing ones, we try to improve metagenomic assembly from another perspective: constructing a robust comparative assembler for metagenomics, which further can be used to complement *de novo* assemblies of the same data. Comparative assembly simply works as follows: at the beginning the short DNA reads are aligned to a reference genome of a closely related species, then contigs are built from the reads following the guidance of the reference genome. Previously, The AMOS comparative assembler [100] has been developed specifically for reference assisted assembly, but is generally only applicable for the analysis of individual genomes when the references are known, and not directly suitable for comparative assembly in metagenomics, for which usually no prior information about the reference genomes are available and at least dozens of reference genomes are needed per sample. Other previously developed tools usually only use comparative assembly to improve *de novo* assembly in very limited setting instead of building a full-fledged comparative assembler, e.g., bridge gaps [29, 127] and orient contigs into bigger scaffolds [54].

So far, thousands of bacterial genomes have been sequenced, and the number is expected to grow rapidly in the next few years, e.g., accordingly to the Genomes Online database (`http://www.genomesonline.org`) 1416 bacterial genomes have been completed and 6114 are ongoing. These sequenced genomes provide a great resource for performing comparative assembly of metagenomic sequences, because with the guidance of a reference genome, it is relatively easier to construct contigs spanning hypervariable regions and repeats within a composite population of genomes.

In this paper we develop algorithms and a software tool to improve metagenomic assembly through comparative genomics and hybrid combination. We show that comparative assembly can achieve comparable results to the state-of-the-art *de novo* assembler in terms of assembly statistics. Additionally, we show that comparative and *de novo* assemblies are complementary to each other, and the combination of these two can significantly improve the final metagenomic assemblies.

6.2 Methods and Materials

6.2.1 Methods overview

Figure 6.1 shows an overview of the MetaCompass comparative assembly pipeline. First we use MetaPhyler [80] to estimate the genomic compositions for all available reference genomes that might present in a metagenomic sample. Then the genomes with high depth of coverage are used as reference genomes for further comparative assembly. Afterwards, the metagenomic shotgun reads are mapped to the reference genomes using MUMmer-map developed in this study, or any other alignment tool that produces results in SAM format (e.g., Bowtie 2 [72] and BWA ([77])). Subsequently, based on the read alignments we build contigs using a minimum set cover algorithm, and use an iterative mapping approach to improve the contigs. Finally, the contigs are combined with assembly from *de novo* approaches to produce final results. Each step will be described in detail in the following sections.

6.2.2 Selecting reference genomes

In contrast to comparative assembly for single genome, a big challenge for metagenomics is the fact that we do not know what reference genomes we should use from thousands of genomes available in public databases. One simple solution is to use all available genomes as reference genomes, but this approach has several limitations. Building efficient indexes for large numbers of reference sequences

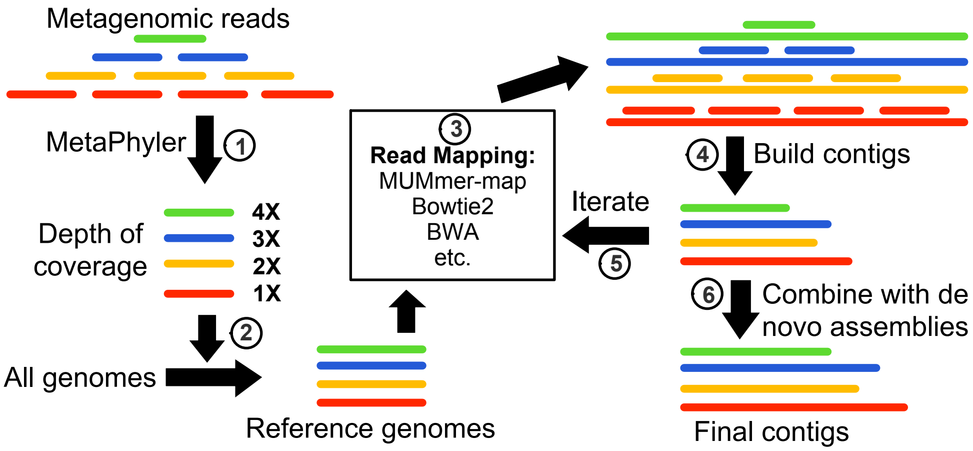


Figure 6.1: Overview of MetaCompass comparative assembly pipeline. First, we estimate the abundances of all available genomes in the sample using MetaPhyler. Next, genomes with high depth of coverage are extracted and used as reference genomes for assembly. Afterwards, metagenomic shotgun reads are aligned against these genomes using MUMmer-map, Bowtie 2 or BWA. Subsequently, we build consensus contigs using the minimum set cover algorithm and use an iterative mapping approach to improve the consensus contigs. Finally, the contigs are combined with assembly from *de novo* approaches to produce final results.

is computationally expensive and challenging, e.g., the index size limit for both Bowtie [73] and BWA [77] is about 4Gbp, substantially smaller than the total size of reference genomes available today. In addition, using all the available genomes requires a significant amount of memory during mapping, which may limit the usability of the tool in practice. Essentially, we only need to use reference genomes whose close relatives are present in the sample at high enough coverage to allow assembly (i.e., we only care about abundant reference genomes). To identify these genomes, we first run MetaPhyler to estimate the taxonomic composition, and then only genomes with high enough depth of coverage are used as references for further comparative assembly. MetaPhyler relies on phylogenetic marker genes as a taxonomic reference to quickly and accurately estimate the taxonomic composition. The depths of coverage of the reference genomes are estimated based upon the reads that are successfully classified.

6.2.3 Aligning reads to reference sequences

Comparative assembly starts with aligning shotgun reads to reference genomes, and typically this step is very computationally expensive because a massive amount of alignments need to be performed. Read mapping is a very fundamental problem in many computational biology problems, and plenty of read mappers [72,73,77] are available, especially for re-sequencing analysis of large eukaryotic genomes. Many of them are very fast for aligning reads that have very few mutations (e.g., less than 4 substitution errors). But the performance drops quickly when there are many mutations, especially insertions and deletions (e.g., bwa-short is mainly designed for sequencing error rates below 2%), which are quite common in metagenomic settings, because (i) bacterial species mutate and evolve fast, (ii) metagenomes usually contain a population of similar genomes for a particular species, (iii) the reference genomes diverge from the environmental sequences because of the vast diversity of unknown bacteria in the environment. In order to achieve high sensitivity for metagenomic read mapping, we should allow more differences than commonly encountered in the analysis of eukaryotic genomes.

In this project, as an alternative to Bowtie 2 [72], we developed a metagenomic read mapping tool (named as MUMmer-map) based on finding maximal exact seeds [70] and Smith-Waterman alignment extension (Figure 6.2). We first build a suffix tree [70] for the reference genomes. Then to align the reads, between each query read and genomes, a set of maximal exact matches that are longer than a threshold (18bp by default) are identified and are used as alignment seeds for next step. Afterwards, the left and right surrounding sequences between the query and reference are aligned through the Smith-Waterman algorithm.

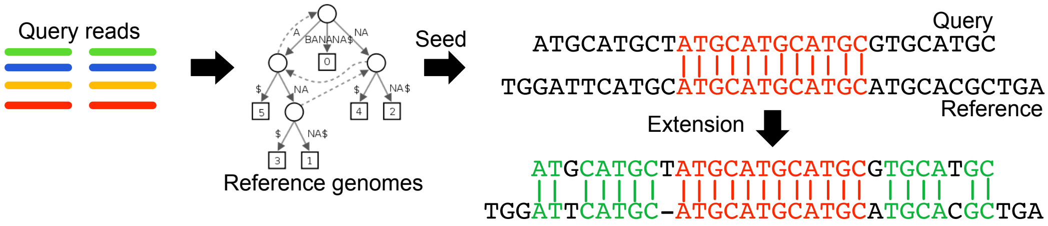

Query reads
A
BANANA$
NA
$
NA
$
NA$
$
NA$
0
5
4
2
3
1
Reference genomes
Seed
Query
ATGCATGCTATGCATGCATGCGTGCATGC
TGGATTCATGCATGCATGCATGCATGCACGCTGA
Reference
Extension
ATGCATGCTATGCATGCATGCGTGCATGC
TGGATTCATGC-ATGCATGCATGCATGCACGCTGA

Figure 6.2: Mapping reads to reference genomes. First, exact alignment seeds (maximal exact matches defined in MUMmer) are found by aligning shotgun reads to the suffix tree of reference genomes. Then surrounding sequences of the seeds are aligned between the query and reference through Smith-Waterman algorithm.

Since this paper is not about short read alignment, performance comparisons with other read mappers are beyond the scope of this paper. Our whole comparative assembly software is designed and developed in a modular way such that any read mapping tools can be used without interrupting the whole assembly process. Specifically, MetaCompass supports the SAM alignment format [78] produced from many popular tools, e.g., Bowtie and BWA.

6.2.4 Building contigs

As described above, at the beginning, we use MetaPhyler to choose a set of candidate reference genomes. This process may retrieve too many genomes as we may have multiple reference genomes that are very similar to each other from the same species. When building contigs, if a metagenomic read is mapped to more than one reference genomes equally well, we need to decide which reference genome(s) to use.

Figure 6.3a shows that if we use all the read mapping records disregarding whether the read is uniquely mapped or it is mapped to multiple reference genomes equally well, then potentially we will create many redundant contigs because of the similar genomic regions. Figure 6.3b shows that if we randomly pick one of the multiply-mapped alignments, then we may create many fragmented small contigs, which ideally should form one continuous big contig. Figure 6.3c shows that we can assign a read to the genome with highest depth of coverage, because the more abundant this genome is in the sample, the more likely that this read comes from it. This approach produces much better results than that from 6.3a and 6.3b. A fourth solution we propose is to pick a minimum set of genomes such that all of the reads can be mapped to at least one of them (Figure 6.3d).

Once we have identified this set of genomes, they will be ranked by the number of reads that can be mapped to each of them. The assemblies are built using the reference genomes one by one according to the rank. Note that once a read is used in a reference genome with higher rank, it will be discarded and not available for genomes with lower rank. The basic idea here is to use as few reference genomes as possible (being parsimonious), but at the same time trying to assemble more reads, construct longer contigs and avoid redundancies.

Formally, this problem can be framed as a set-cover problem, which is NP-hard and is an optimization problem that models many resource-selection problems [22].

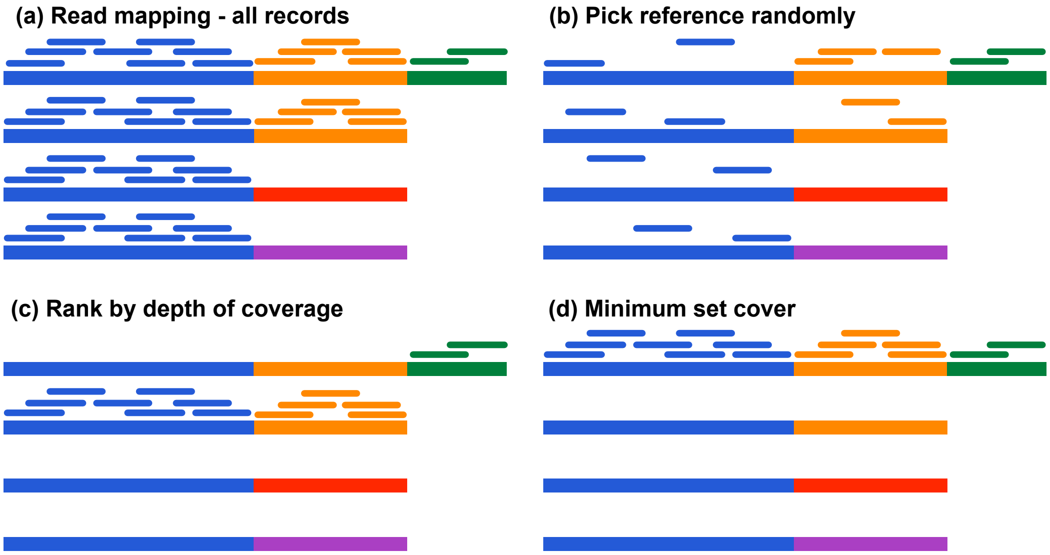

(a) Read mapping - all records
(b) Pick reference randomly
(c) Rank by depth of coverage
(d) Minimum set cover

Figure 6.3: Building contigs from read mapping records. Shorter bars represent shotgun reads; longer and thicker bars represent reference genomes (4 genomes in this figure). Regions with the same color in the reference genomes represent highly similar sequences. (a) All read mapping records. A read may be mapped to several reference genomes equally well, e.g., 5 yellow reads are mapped to both of the first two genomes. (b) For each read, if it is mapped to more than one reference genomes, we randomly pick one record. (c) A read is assigned to a reference with highest depth of coverage. (d) We pick the minimum number of reference genomes, to which all reads can be mapped.

An instance $(X, \mathcal{F})$ of the set-covering problem consists of a finite set X (the shotgun reads that can be mapped to reference genomes) and a family $\mathcal{F}$ (reference genomes) of subsets of X, such that every element of X (each shotgun read) belongs to at least one subset in F (reference genome):

$$X = \bigcup_{S \in \mathcal{F}} S \tag{6.1}$$

where each subset S is a set of reads mapped to a particular reference genome. The problem is to find a minimum-sized subset $\mathcal{C} \subseteq \mathcal{F}$ (minimum number of reference genomes) whose members cover all of X:

$$X = \bigcup_{S \in \mathcal{C}} S \tag{6.2}$$

Here “a reference genome covers a read” means that this read can be mapped to this reference genome. To solve this problem, we use a greedy approximation algorithm (see below) by picking, at each stage, the set S (genome) that covers the greatest number of remaining elements that are uncovered. The algorithm works as follows. The set U contains, at each stage, the set of remaining uncovered elements (reads). The set $\mathcal{C}$ contains the cover being constructed (reference genomes that are picked). Line 4 is the greedy decision-making step. We choose a subset S that covers as many uncovered elements as possible with ties broken randomly. After S is selected, its elements are removed from U, and S is placed into $\mathcal{C}$. When the algorithm terminates, the set $\mathcal{C}$ contains a subfamily of $\mathcal{F}$ that covers X. It can be shown that this greedy algorithm is the best-possible polynomial time approximation algorithm for the set cover problem, under plausible complexity assumptions.

Algorithm 2 Greedy approximation for minimum set covering problem.

Input: a finite set X; a family $\mathcal{F}$ of subsets of X.
Output: a subset $\mathcal{C} \subseteq \mathcal{F}$ whose members cover all of X.

```
1: $U \leftarrow X$
2: $\mathcal{C} \leftarrow \emptyset$
3: while $U \neq \emptyset$ do
4:     select an $S \in \mathcal{F}$ that maximizes $|S \cap U|$
5:     $U \leftarrow U - S$
6:     $\mathcal{C} \leftarrow C \cup S$
7: end while
8: return $\mathcal{C}$
```

Given a set of reference genomes, a set of shotgun reads and the alignment between each read and reference genome, the process of creating contigs is straightforward. For each nucleotide base of each reference genome, we look at the bases from the reads that are mapped to this locus, and pick the nucleotide with the highest depth of coverage as the consensus (Figure 6.4). In addition, to introduce an insertion, its depth of coverage should be higher than half of that of its neighbor nucleotides (Figure 6.4). Minimum depth of coverage and length for creating contigs can be specified through the program command-line options.

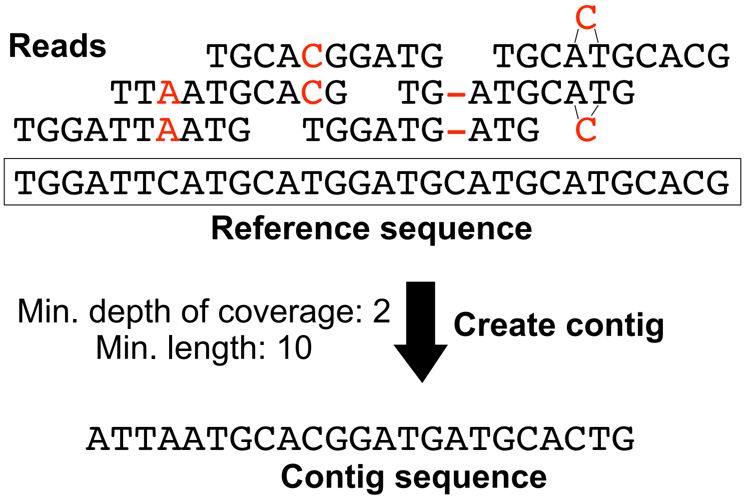


Figure 6.4: Creating consensus contig sequences from reads that are mapped to reference genomes using the majority rule.

6.2.5 Improving assembly through iteration

In metagenomic studies, it is common for a genome in the sample to not have very close relatives in the currently available reference genomes, or to have a hypervariable region that can not be aligned to the reference very well. To assemble such genomes, previous studies [29,127] have shown that iterative assembly is a promising approach which can increase both the contig size and the number of assembled reads. Iterative assembly simply works as follows (Figure 6.5): (i) map shotgun reads to the original reference genomes; (ii) create contigs based on the reads that are aligned;

(iii) use the newly created contigs and their surrounding sequences from the original genomes as new reference sequences, and iterate previous two steps multiple times until the assemblies can not be improved significantly.

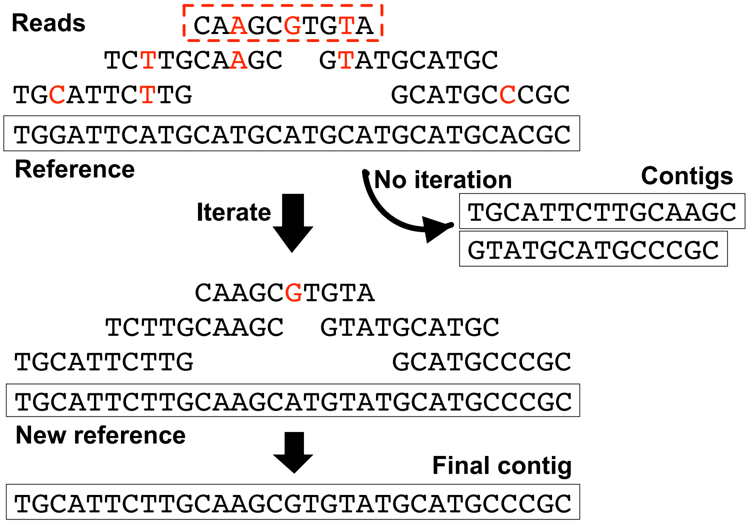


Figure 6.5: Iterative mapping can potentially improve comparative assembly. Here in this toy example, we assume that a read is considered as mapped when it has less than 3 mismatches. Hence the top read enclosed in red rectangle will not be mapped to the reference at the beginning. If no iteration is performed, two short contigs will be created from the comparative assembly. If iteration is performed, the top read will be successfully mapped to the new reference, producing a single long contig.

6.2.6 Combining comparative and *de novo* assemblies

As we have discussed before, comparative assembly is quite sensitive to genome recombinations, insertions or deletions, which unfortunately occur in bacterial genomes even between closely related species. Given that the bacterial genomes we have sequenced so far form only a very small subset of all bacterial species in our earth ecosystem, using a comparative approach will lead to fragmented assemblies or may fail to reconstruct novel species. *de novo* assembly, on the other hand, does not rely on any previous information, so it can, in principle, recover any genomes from the sample. In contrast, de novo assembly can reconstruct novel genomes but is affected by sequencing errors and repeats. Combining the two approaches allows us to partly mitigate their respective weaknesses and benefit from their complementary strengths. By leveraging previous available tools in genome assembly, we use a light-weight assembler MINIMUS [120] to directly combine the contigs generated from the two approaches. MINIMUS has been used successfully in [48] to combine contigs from Velvet and Edena. We should note that we do not handle assembly errors very well during the combination at this point even if one assembly may be able to correct the other one. The parameters we used are: (i) minimum overlap length is 100bp; (ii) overlap similarity cutoff is 95%.

6.2.7 Materials

Even and staggered metagenomic samples of mock community from the Human Microbiome Project were downloaded from the NCBI Short Read Archive with project ID 48475. The Illumina reads of 728 metagenomic samples from the Human Microbiome Project were downloaded from the Data analysis and Coordination Center (`http://www.hmpdacc.org`).

6.2.8 Parameters used in data analysis

In all results and analysis in this paper, only contigs longer than or equal to 300bp are considered. Boxplots were created using default settings in R. All software tools used in this study were run under default parameters, unless specified otherwise.

6.3 Results and Discussions

In this section, we first evaluate MetaCompass on two artificial metagenomic samples for which we know the genomic composition, and compare its performance with four other assemblers: AMOScmp [100], Meta-IDBA [96], MetaVelvet [139] and SOAPdenovo [79]. Then, we compare their performance using one real tongue dorsum metagenomic sample, which was generated by the Human Microbiome Project (HMP) and has been used previously in Meta-IDBA [96] for evaluation. Finally, to demonstrate the performance and applicability of MetaCompass on large-scale metagenomic data sets, we analyzed 728 metagenomic samples from the HMP project, and compared its performance with SOAPdenovo and a hybrid assembly approach.

6.3.1 Assembly evaluation using two artificial metagenomic samples

We evaluated the performance of MetaCompass and four other assemblers on two synthetic (also known as mock) microbial communities, which were created from purified genomic DNA of 20 bacteria and eukaryotes for which finished genome sequences are available (Table 6.1) [20]. The two artificial communities were created such that the 16S rRNA copies of the organisms have a uniform distribution for the even community and long-tail power law distribution for the staggered community. After sequence quality trimming, 18.6 and 33.2 million Illumina reads with average length 98bp and 95bp were obtained for even and staggered samples, respectively. The taxonomic composition of the sequenced metagenomic reads from these two samples was estimated using MetaPhyler [80] (Figure 6.6). The taxonomic composition of the mock even community is more uniform than that of the staggered community (entropy values are 3.39 versus 1.66), but the abundance distribution of the mock even community is still far from uniform, even after correcting for 16S rRNA copy variations. This may reflect the experimental biases coming from the sample preparation, sequencing technology, etc.

When testing the performance of MetaCompass on these two mock samples, we created two scenarios. In the first we assume the genomic composition of the mock communities is given (denoted as MC-Ref in figures) , and we run MetaCompass directly using the 22 true reference genomes. The assembly results can be considered as an upper bound on MetaCompass performance, because we know the exact genomes for the metagenomic reads to be assembled. In the second scenario, we run MetaCompass in normal settings (no prior information about genomic composition) on these two samples, and the reference genomes are selected according to the depth of coverage estimated from MetaPhyler. To compare the performance, we also assembled these two mock data sets using three previously developed *de novo*

Genome	DNA mass (grams)	
	Even	Staggered
Acinetobacter baumannii ATCC 17978	1.60E-10	1.60E-11
Actinomyces odontolyticus ATCC 17982	7.82E-11	7.82E-13
Bacillus cereus ATCC 10987	3.73E-11	3.73E-11
Bacteroides vulgatus str. ATCC 8482	1.52E-10	1.52E-12
Candida albicans ATCC MY-2876	3.27E-11	2.92E-11
Clostridium beijerinckii ATCC 51743	3.81E-11	3.81E-11
Deinococcus radiodurans ATCC 20539	1.76E-09	1.76E-11
Enterococcus faecalis str. ATCC 47077	2.22E-11	2.22E-13
Escherichia coli ATCC 70096	2.71E-11	2.71E-10
Helicobacter pylori ATCC 700392	4.50E-11	4.50E-12
Listeria monocytogenes ATCC BAA-679	1.53E-11	1.53E-12
Lactobacillus gasseri str. ATCC 20243	3.98E-11	3.98E-12
Methanobrevibacter smithii ATCC 35061	9.50E-11	9.50E-10
Neisseria meningitidis ATCC BAA-335	6.87E-11	6.87E-12
Propionibacterium acnes DSM 16379	1.39E-10	1.39E-11
Pseudomonas aeruginosa ATCC 47085	1.80E-10	1.80E-10
Rhodobacter sphaeroides ATCC 17023	1.30E-10	6.97E-11
Staphylococcus aureus ATCC BAA-1718	6.97E-11	1.31E-09
Staphylococcus epidermidis str. ATCC 12228	1.31E-10	1.83E-11
Streptococcus agalactiae ATCC BAA-611	1.83E-11	4.70E-10
Streptococcus mutans ATCC 700610	4.70E-11	8.11E-13
Streptococcus pneumoniae ATCC BAA-334	8.11E-11	6.97E-11

Table 6.1: Genomic composition of even and staggered mock samples [20]. The two artificial communities are created such that the 16S rRNA copies of the organisms have a uniform distribution for the even community and long-tail power law distribution for the staggered community.

assemblers: SOAPdenovo [79], MetaVelvet [90] and Meta-IDBA [96], and one comparative assembler AMOScmp [100]. When running AMOScmp, we assume that the 22 reference genomes are known, because this software was designed for single genome comparative assembly, and does not have the ability to automatically select the reference genome. The three *de novo* assemblers were run under default parameters. In the following, we compare the assembly performance using three measurements: contig size, assembly error, and coverage of reference genomes (see Methods for details).

Figure 6.7 shows the comparison of contig sizes between different assemblies on even and staggered mock data sets. The trajectory shows the contig length (y axis) plotted against the cumulative size (x axis) of the contigs whose lengths are smaller

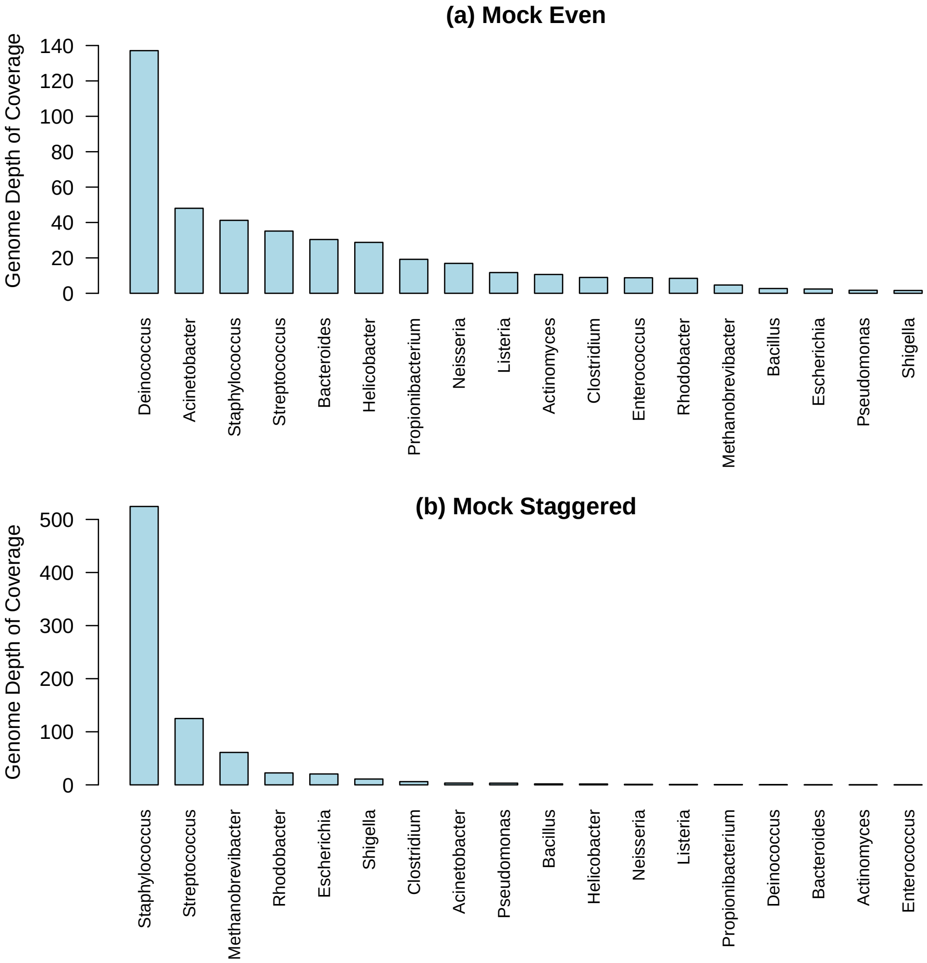


Figure 6.6: Taxonomic compositions of even and staggered mock metagenomic samples estimated from metagenomic shotgun reads using MetaPhyler. Genome depth of coverage (y axis) is estimated by MetaPhyler based on classified phylogenetic marker genes.

than or equal to the contig length represented on the y axis. The y axis is represented in log scale. As expected, MetaCompass with known true reference genomes creates the longest contigs among all approaches (red curves in Figure 6.7). MetaCompass without prior information about the genomic composition performs second best (blue curves in Figure 6.7), and it performs better than AMOScmp (AMOS-Ref; brown curves), even though AMOScmp uses the true reference genomes. Especially, from Figure 6.7 we can tell that MetaCompass creates some very long contigs (almost completely covers the abundant genomes), and this will be very helpful for some further genome-scale analysis. For the staggered metagenomic sample (Figure 6.7b), MetaCompass is consistently better than AMOS-Ref. For the even metagenomic sample in Figure 6.7a, MetaCompass is significantly better than AMOS-Ref when cumulative size is smaller than 30Mbp, and drops below AMOS-Ref afterwards. The reason is the reference genomes selection procedure wherein, reads are assigned to genomes with a greedy approach according to the rank (see Methods for details). Reads coming from similar genomic regions (e.g., 16S rDNA) in different genomes will all be assigned to the genome of highest rank. Hence, this will produce fewer contig sequences, and fragment the assemblies for lower-rank genomes into smaller pieces.

Next, we compare the quality of the contigs generated from different assemblers by looking at the number of mis-assembled contigs. There are many different measures of accuracy, and we take a simple approach that is similar to the Feature Response Curve proposed by [91]: a contig is counted as one error if it can not be aligned to the true reference genomes contiguously. Here we ignore the number of break points required to align the error contig to reference. We first sort the contigs in decreasing order by length, and plot the cumulative contigs length (y axis) against the number of misassembled contigs (Figure 6.8). MetaCompass (MC-Ref in Figure 6.8) makes very few errors when using true reference genomes; in contrast, AMOScmp (AMOS-Ref in Figure 6.8) creates a lot more errors even given the true genomes. MetaCompass (blue curve in Figure 6.8) without knowing true reference

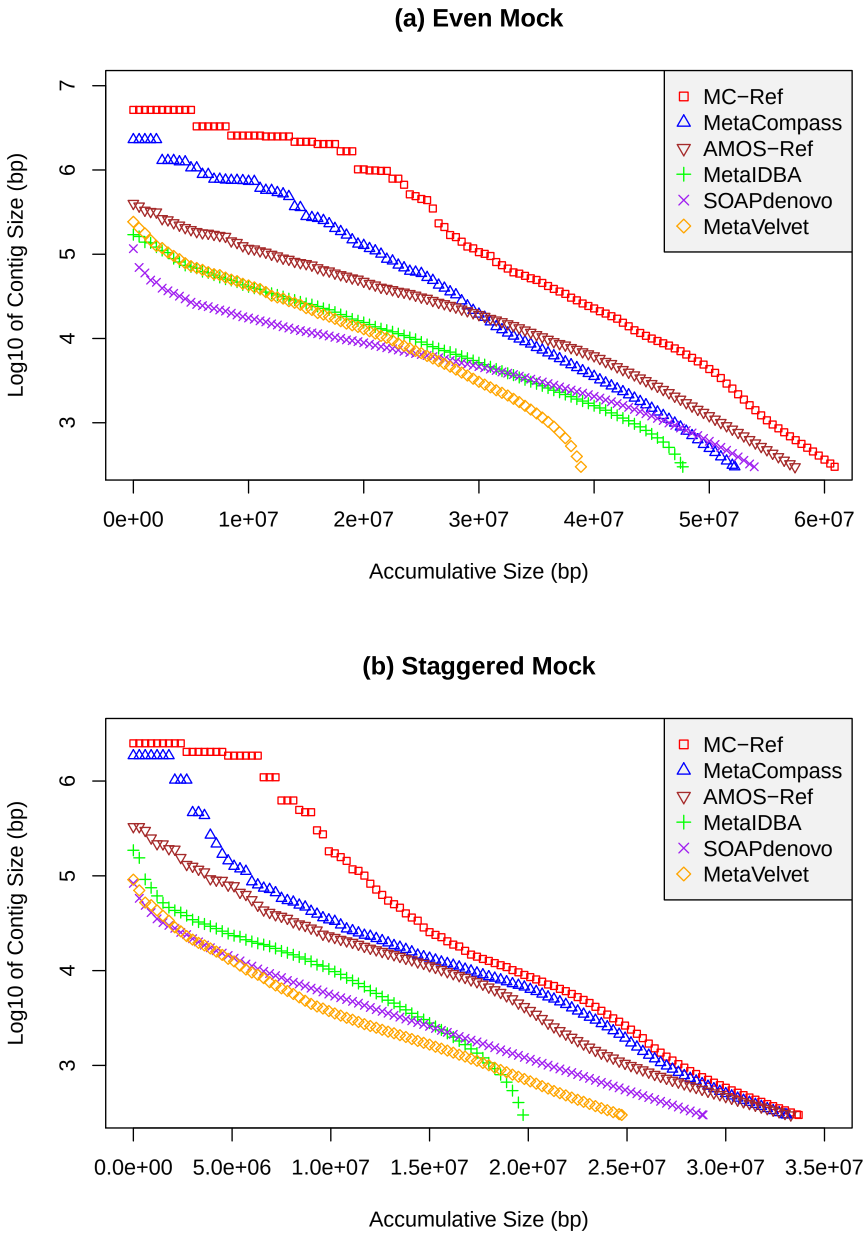


Figure 6.7: Accumulative contig size distribution of different assemblies for even and staggered mock metagenomic samples. The contigs are sorted in order of decreasing length. The contig length (y axis) is plotted against the accumulative size (x axis) of the contigs whose lengths are smaller than or equal to y axis. The accumulative contigs size (x axis) is plotted against the contig size (y axis). Note that the y axis is in log scale, so a difference by one indicates a 10 times difference in real contig length. MC-Ref and AMOS-Ref assume the true reference genomes are given when running MetaCompass and AMOScmp assemblers.

genomes is still better (makes fewer errors) than AMOScmp. Also from Figure 6.8, we can tell that almost all errors made by MetaCompass come from small contigs; whereas other assemblers make much more errors in large contigs. Since the errors from MetaCompass are in small contigs, they will only affect the quality of further genomic analysis very slightly. Among the three *de novo* assemblers, we see that SOAPdenovo produces more contig sequences than the other two, but at the same time, also produces much more errors.

The above two evaluation measurements can not recognize assembly that produces redundant contigs. To address this issue, for each assembly we map the contigs back to the true reference genomes and calculate the regions that are covered (Figure 6.9). Again, MetaCompass with known reference genomes (MC-Ref in Figure 6.9) covers the largest amount of reference genomes. MetaCompass covers slightly fewer reference genomes than MC-Ref does, and AMOS-Ref demonstrates similar performance with MetaCompass. Among the *de novo* assemblers, SOAPdenovo assembly covers much more reference genomes than the other two do, but as we have observed before (Figure 6.8), it also creates much more errors.

Given that the genomes in the metagenomic sample have been sequenced before, reference based comparative assembly indeed can produce longer contigs than *de novo* assemblers do. However, keep in mind that in real metagenomic samples, closely related reference genomes are frequently not available, especially for samples

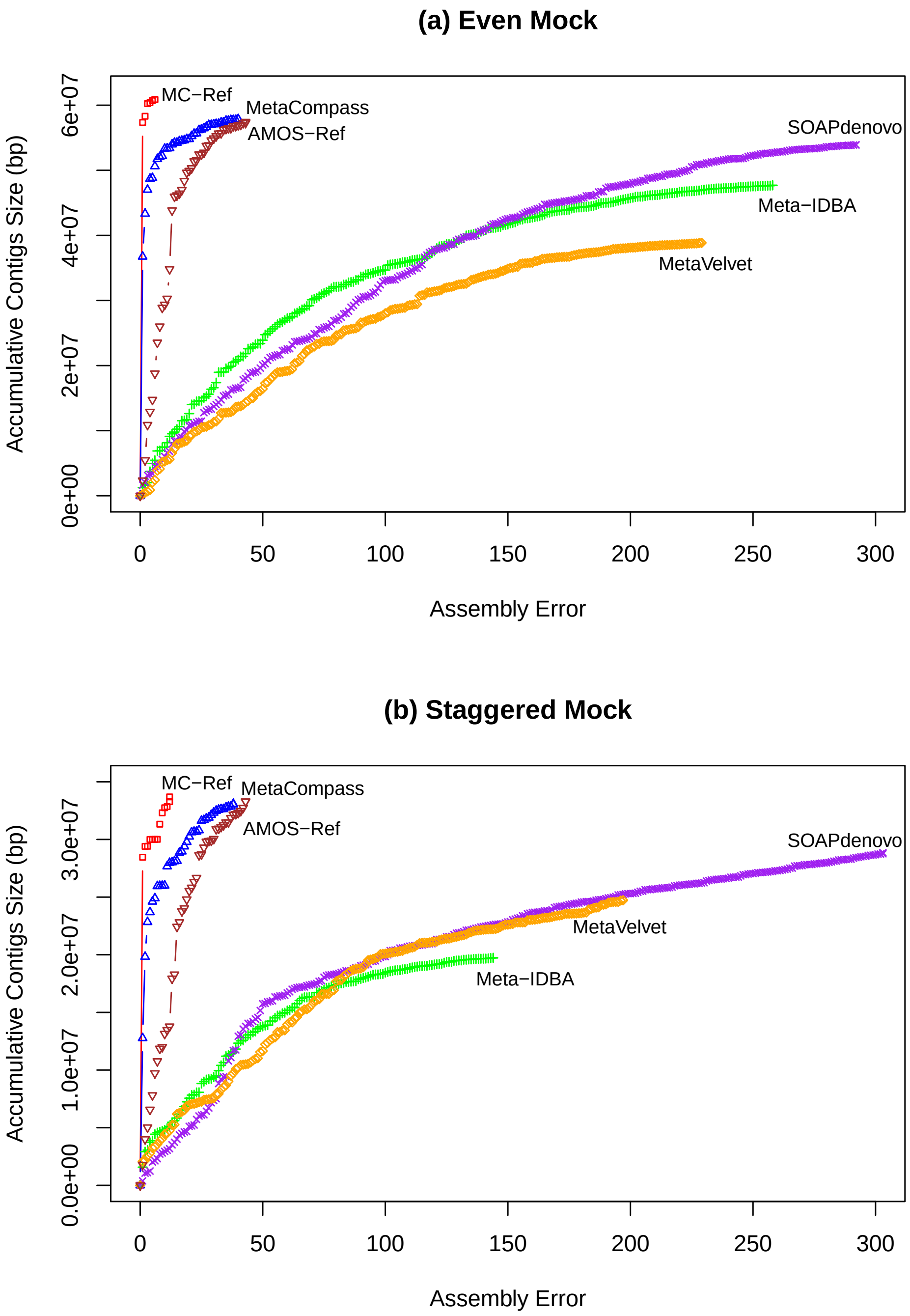


Figure 6.8: Distribution of assembly errors generated by MetaCompass and four other assemblers on the two mock metagenomic data sets. The contigs are sorted in decreasing length, and the cumulative contigs size (y axis) is plotted against the number of erroneous contigs. MC-Ref and AMOS-Ref assume the true reference genomes are given when running MetaCompass and AMOScmp assemblers.

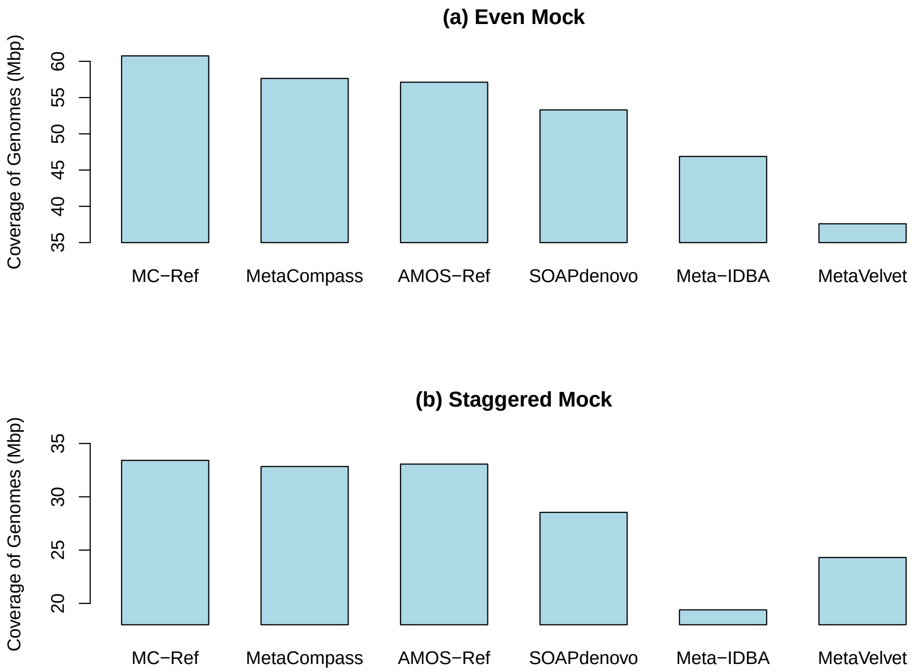


Figure 6.9: Number of bases of the reference genomes that are covered by assembled contigs. MC-Ref and AMOS-Ref assume the true reference genomes are given when running MetaCompass and AMOScmp assemblers.

from novel environments. So, in this section we are not suggesting that MetaCompass should substitute *de novo* assemblers, rather that when good reference genomes are available, MetaCompass can produce significantly longer and more accurate assemblies. Comparative assembly by MetaCompass should also be performed beside *de novo* assemblies to achieve best overall metagenomic assembly.

6.3.2 A tongue dorsum metagenomic sample from HMP

We use a real human tongue dorsum metagenomic sample from the Human Microbiome Project [20] to evaluate the performance of assemblers in practice. The reason we choose this data set is because it had been used previously by Meta-

IDBA [96]. Figure 6.10 shows the taxonomic composition at the genus and phylum level, as estimated by MetaPhyler.

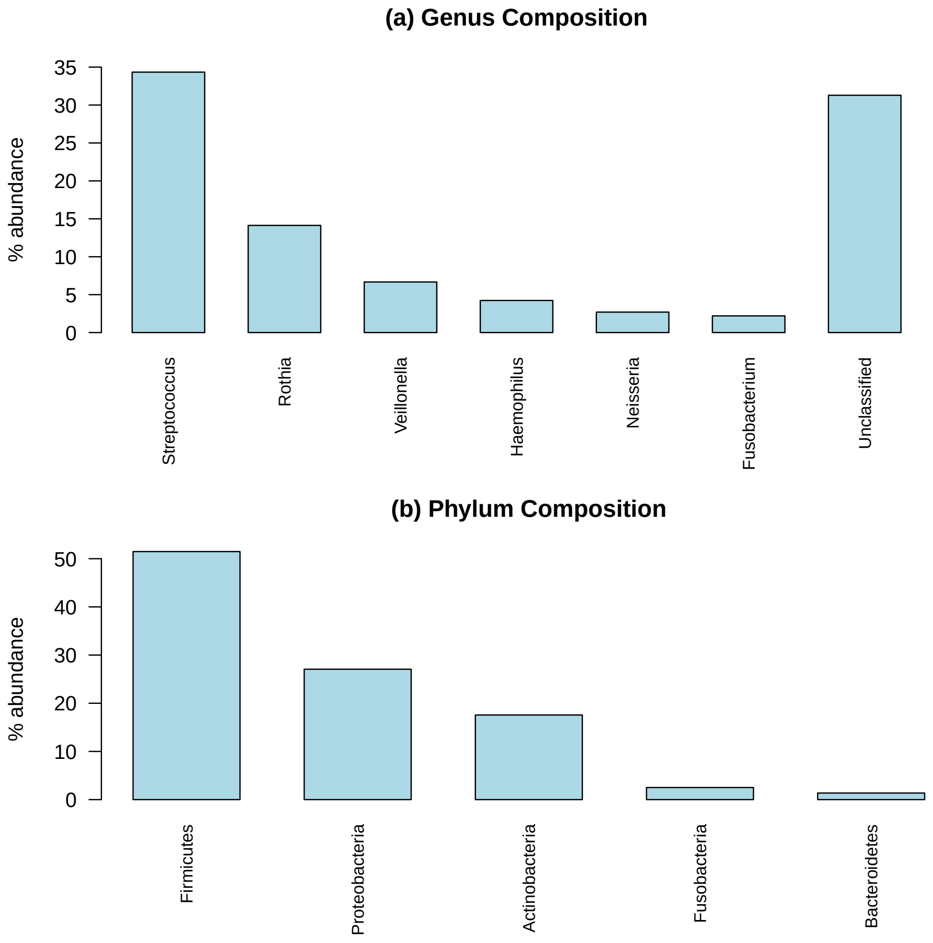


Figure 6.10: Taxonomic composition of a tongue dorsum metagenomic sample estimated by MetaPhyler. Only genera and phyla, whose abundances are higher than 1%, are shown in this figure.

Figure 6.11 shows the comparison of contig sizes from different assemblies. In addition to the assemblers mentioned in the previous section, we also used MINIMUS [120], a light-weight assembler, to combine the contigs between MetaCompass, and MetaVelvet or SOAPdenovo. Individually, the performance of the assemblers do not differ significantly; MetaVelvet produces the largest assemblies; MetaCompass appears to be better than others between 20Mbp and 50Mbp of cumulative size (Figure 6.11). The combination of either MetaCompass and MetaVelvet (MC+MV in Figure 6.11), or MetaCompass and SOAPdenovo (MC+SOAP in Figure 6.11) produces significantly more in total size and longer contigs than any individual assemblers.

Redundant contigs can artificially inflate the contig size statistic used above. To prevent this situation, we calculate the number of reads that are assembled for each assembly by mapping the reads back to the contigs using Bowtie 2 (Figure 6.12). MetaVelvet assembles about 52.6 million reads (62.7%), which is more than any other assembler does; the two combination approaches (MC+MV and MC+SOAP in Figure 6.12) assembled 61.8 (73.5%) and 68.1 million reads (81.1%), which represent huge improvements over individual assemblers alone. In summary, in this real tongue dorsum metagenomic sample, the performance of MetaCompass is comparable to the state of art *de novo* assemblers, and the hybrid approach gives the overall best assembly.

6.3.3 728 metagenomic samples from HMP

We further evaluate the performance of MetaCompass on 35.7 billion Illumina shortgun reads from 728 metagenomic samples of the Human Microbiome Project [20]. These samples cover 15 different human body sites all coming from healthy

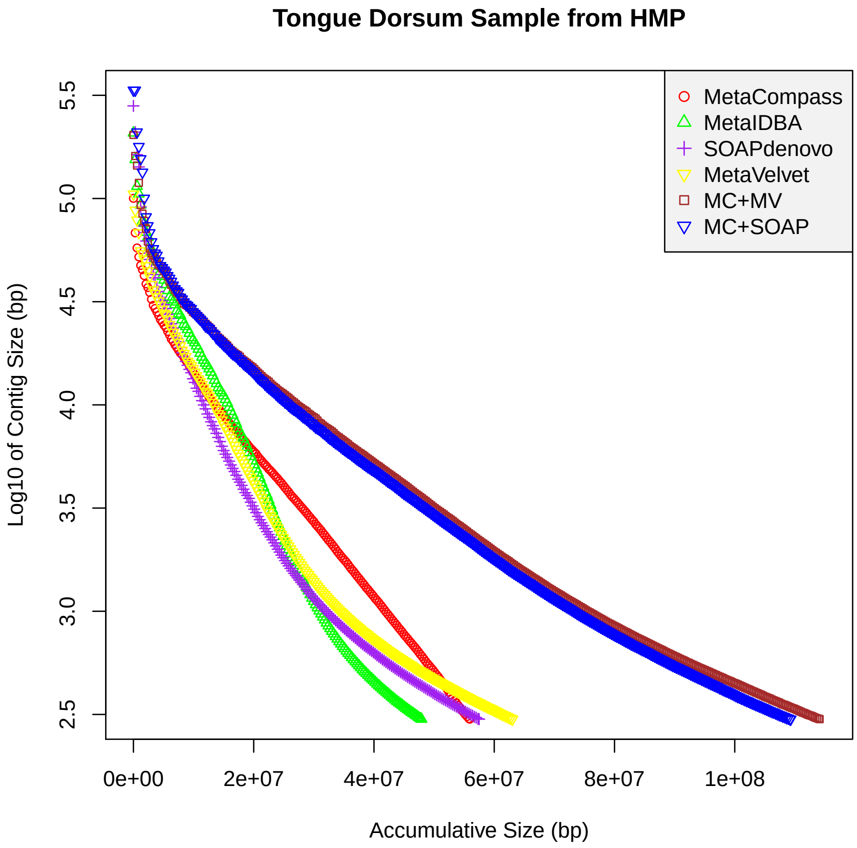


Figure 6.11: Accumulative contig size distribution of different assemblies for a tongue dorsum metagenomic sample. The contigs are sorted in order of decreasing length. The contig length (y axis) is plotted against the accumulative size (x axis) of the contigs whose lengths are smaller than or equal to y axis. The accumulative contigs size (x axis) is plotted against the contig size (y axis). Note that the y axis is in log scale, so a difference by one indicates a 10 times difference in real contig length. MC+MV represents combinations of contigs from MetaCompass and MetaVelvet; MC+SOAP represents combinations of contigs from MetaCompass and SOAPdenovo.

individuals. One major goal of HMP is to systematically characterize microbial communities found at different human body sites, and to provide genomic references for future human related metagenomic studies. Here we demonstrate the power of MetaCompass by analyzing all these samples, and combine the assembly with the contigs generated from SOAPdenovo (Figure 6.13). In total, MetaCompass

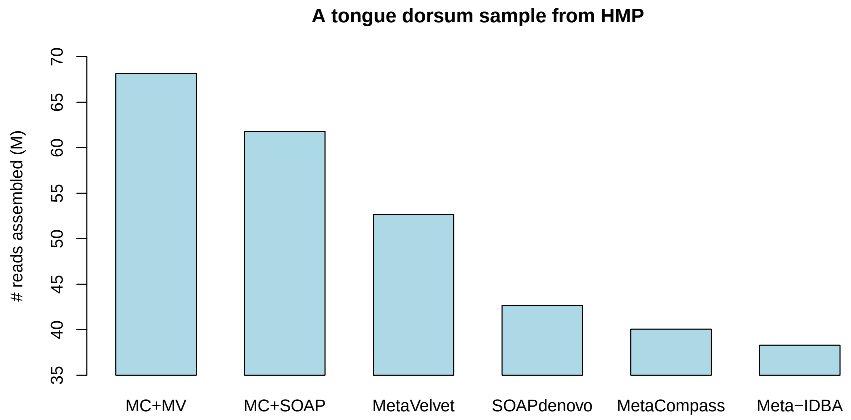


Figure 6.12: Number of reads assembled in different assemblies for a tongue dorsum metagenomic sample from HMP. The statistic is calculated by mapping the reads back to contigs.

assembled 33.3 billion base pairs (Gbp) in contigs that are longer than or equal to 300bp; SOAPdenovo produced 44.1Gbp contigs, which were assembled by the HMP data analysis consortium previously (http://www.hmpdacc.org/). It is expected to see that SOAPdenovo produced more contigs than MetaCompass does, because the reference genomes we have for the human microbiome only account for a tiny portion of the diverse population of bacteria species living with us. The contig sizes of MetaCompass, however, are bigger than that of SOAPdenovo, because the reference genomes can easily help us resolve repeats during comparative assembly (Figure 6.13). For example, the average maximum contig size across all samples from MetaCompass is 131Kbp compared with 86Kbp from SOAPdenovo.

Before we combine the contigs from MetaCompass and SOAPdenovo, we analyzed the redundany of the contigs from these two approaches to see if we can gain potential benefits from a combined approach. Specifically, we try to find the

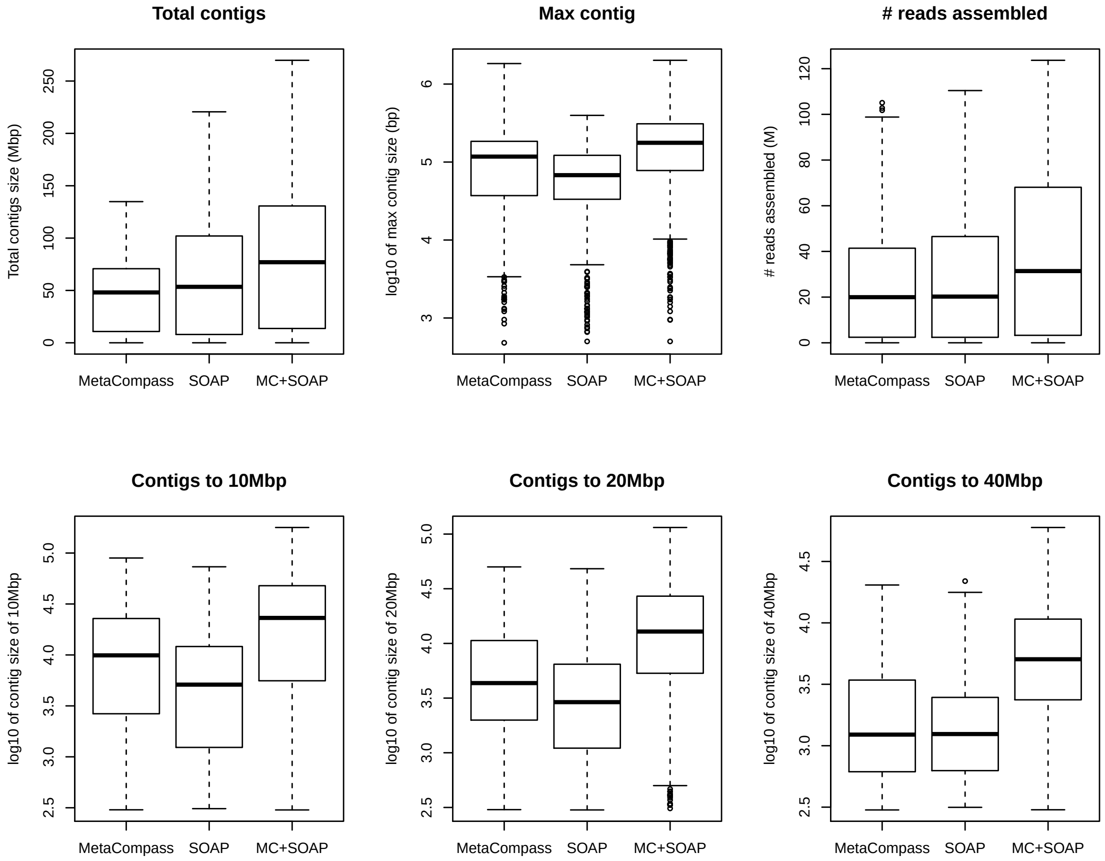


Figure 6.13: Assembly statistics of comparative assembly (MetaCompass), *de novo* assembly (SOAPdenovo) and a hybrid approach (MC+SOAP) for 728 HMP metagenomic samples. "# reads assembled" is calculated by mapping reads back to the contigs using Bowtie 2. "Contig size of nMbp" is computed as the contig length such that using equal and longer contigs produce nMbp.

answers to these two questions: (1) How many contigs from SOAPdenovo can be mapped to reference genomes selected by MetaCompass? (2) Are the reads assembled by MetaCompass a subset of the reads assembled by SOAPdenovo? To answer the first question, we aligned all the contigs generated by SOAPdenovo to reference genomes using BLAST (similarity cutoff: 90%; alignment length cutoff: 300bp), and extracted the contigs that were successfully matched. Figure 6.14 shows that only a very small portion can be aligned against the reference genomes, indicating that SOAPdenovo assembled many novel contigs relative to the reference genomes used by MetaCompass. Specifically, out of the 44.1Gbp contigs generated from SOAPdenovo assembly, only 17.8B can be mapped to reference genomes. To answer the second question, we compared the reads that were assembled by SOAPdenovo and MetaCompass by mapping reads back to the contigs. In total about 11.2 billion reads are assembled by both approaches; 9.3 and 7.6 billion reads are uniquely assembled by SOApdenovo and MetaCompass respectively (Figure 6.15). These two observations indicate that these two approaches should complement each other significantly.

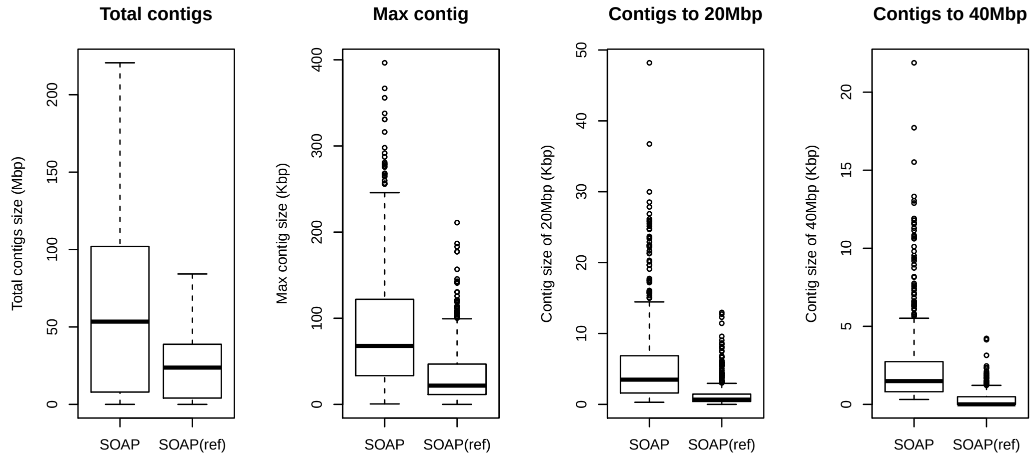


Figure 6.14: Assembly statistics of comparative assembly (MetaCompass), *de novo* assembly (SOAP) and contigs produced from SOAP that can be mapped to reference genomes (SOAP(ref)) on 728 HMP metagenomic samples.

To combine the contigs from MetaCompass and SOAPdenovo, we choose to use a previously developed overlap-layout-consensus based fast and lightweight assembler MINIMUS [120]. Figure 6.13 shows that this simple hybrid assembly approach

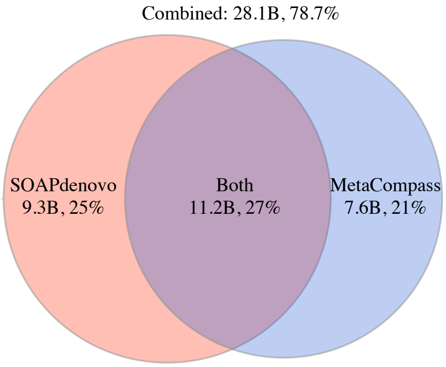


Figure 6.15: Venn diagram showing the number of reads assembled by SOAPdenovo and MetaCompass. In total, 28.1 billion (78.7%) out of 35.7 billion reads are assembled into contigs longer than or equal to 300bp by SOAPdenovo and MetaCompass combined. 11.2 billion reads (27%) are assembled by both approaches.

produces much better results than either MetaCompass or SOAPdenovo does individually. In total, hybrid assembly produces 58.4Gbp contigs, whereas MetaCompass and SOAPdenovo produces 33.3Gbp and 44.2Gbp contigs respectively. The significant increase in the amount of contigs assembled is not a result of redundant sequences, because the number of reads assembled also increases significantly to 28.2 billion reads (79% of all reads). Overall, the contigs assembled from this hybrid approach significantly improves the performance and increases our ability to analyze complex metagenomic samples.

In addition, we analyzed the assembly results separately for seven body sites (buccal, fonix, nares, retroauricular, stool, supragingival, and tongue), from which more than 10 samples were sequenced (Figure 6.16). The relative performance of MetaCompass and SOAPdenovo varies significantly across body sites. Regarding buccal samples, MetaCompass is better than SOAPdenovo in all categories, e.g., MetaCompass assembled 2.97Gbp contigs from 777 million reads, while SOAPdenovo generated 1.87Gbp in contigs from 551 million reads. For tongue samples, MetaCompass produced 7.3Gbp in contigs from 4.9 billion reads and SOAPdenovo produced 11.7Gbp in contigs from 5.4 billion reads. However the average maximum contig size of MetaCompass (130.1Kbp) is larger than SOAPdenovo (110.9Kbp). The performance of *de novo* assembly depends heavily on how complex the community is and the amount of repetitive sequences in the genomes; in contrast, the performance of comparative assembly depends on how close the reference genomes are to the genomes in the metagenomic sample.

6.3.4 Improvements from iterative assembly

In metagenomics, we might frequently see mutation hotspot regions, where our target genomes in the sample are not very similar with available reference genomes. The shotgun reads coming from these regions can not be mapped to the reference genomes because of the large differences, resulting in fragmented contigs. Fortunately, iterative assembly potentially can help us approach the consensus genome in the natural population, and walk through the hotspot genomic regions (see Methods for details). To verify the assembly improvements gained from iterative assembly, here we show the results from 728 metagenomic samples from the Human Microbiome Project. During the comparative assembly, we have performed three iterations, and Figure 6.17 shows the corresponding assembly statistics. The total number of reads assembled increases from 17.4 to 18.8 billion reads, and the average maximum contig size increases from 106Kbp to 131Kbp. Overall, the iterative approach does improve the assembly significantly as expected.

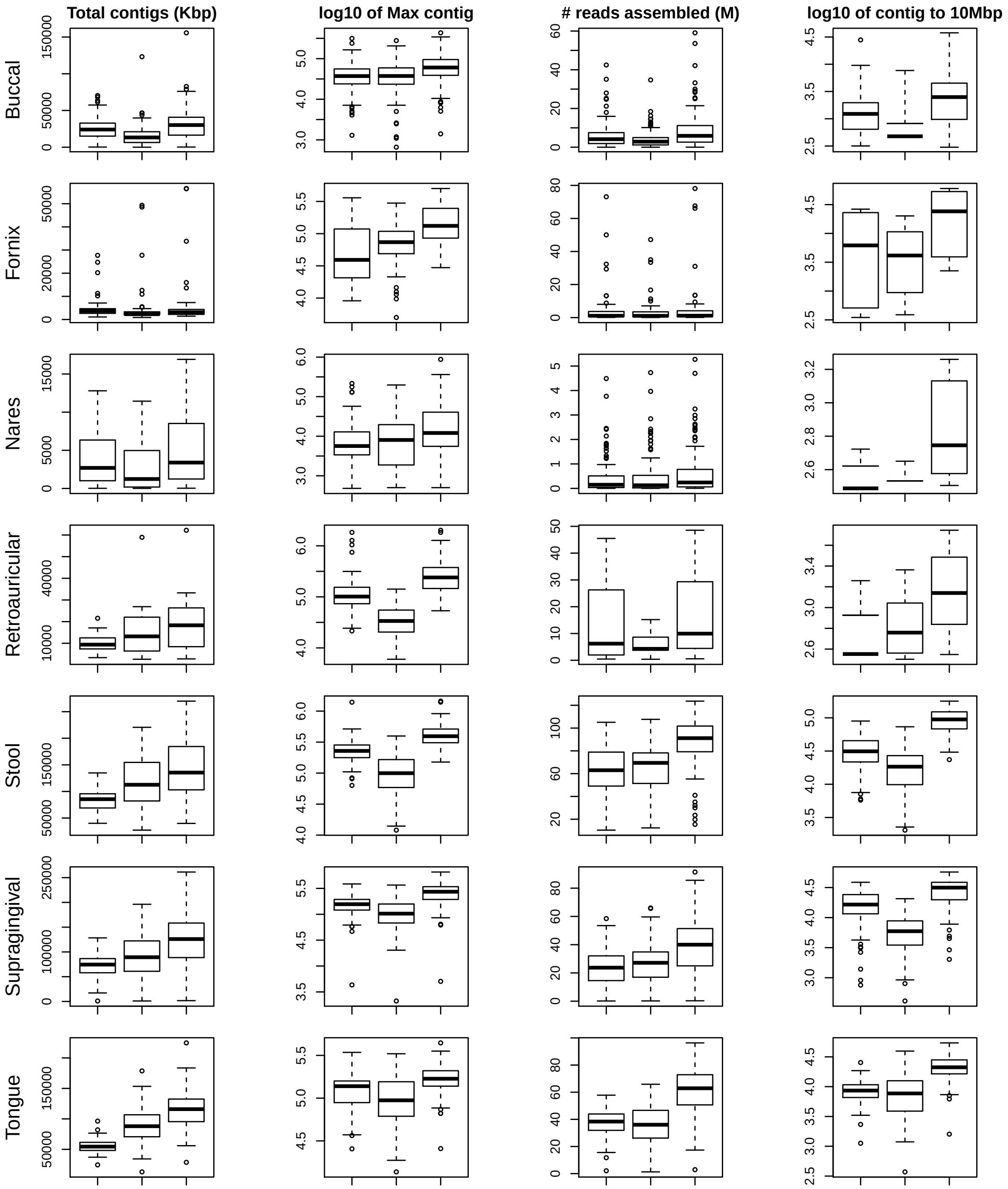


Figure 6.16: Assembly of HMP samples by body site. Each row represents a distinct body site. The three boxplots in each panel are MetaCompass, SOAPdenovo and combination (MC+SOAP).

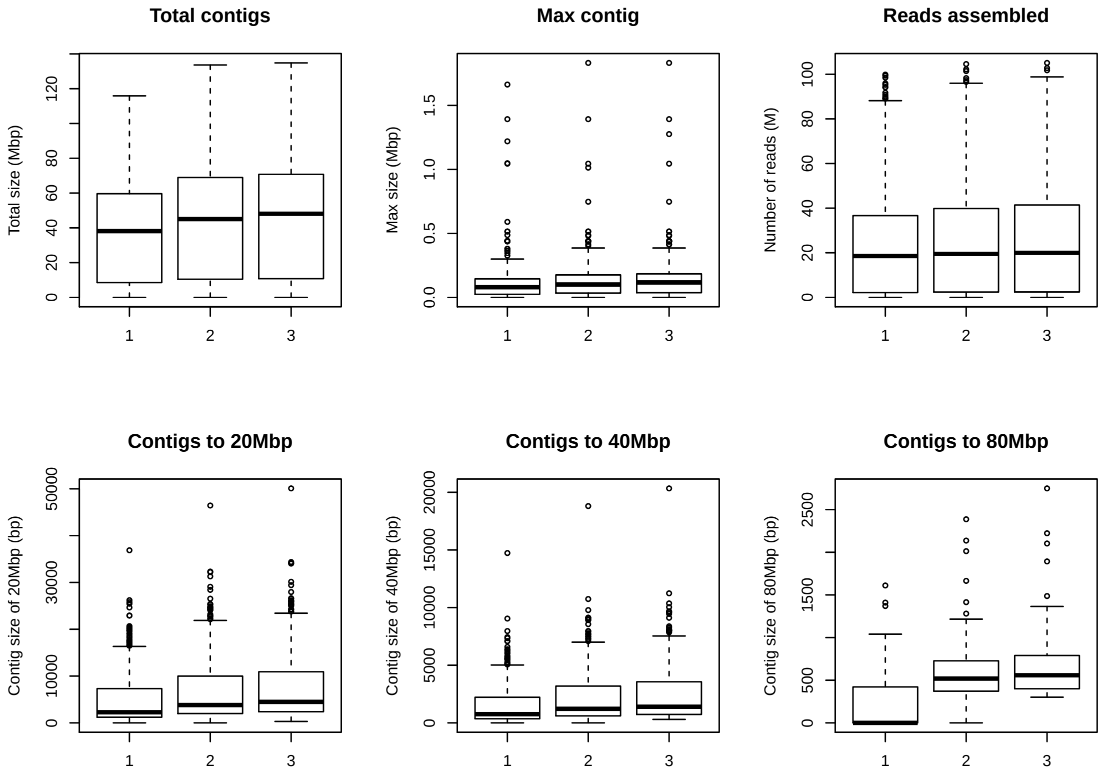


Figure 6.17: Improving comparative assembly through iterative mapping for 728 metagenomic samples from the Human Microbiome Project. Both the contig size and the number of assembled reads are improved significantly.

6.3.5 Discovering genomic variation

Comparative and *de novo* assemblies are two independent and complementary ways to reconstruct genomes from metagenomic shotgun sequences. As shown above, a hybrid approach improves the overall assembly significantly, and allows for better further computational analysis, e.g., taxonomic annotation, gene prediction and annotation, genomic variation, etc. In addition, since the contigs from comparative assembly are automatically aligned against the reference genome, it gives us an easy way to explore genomic variations. Here we show an anecdotal example of the discovery of a potential genomic recombination event.

Cmp1 and Cmp2 are two contigs assembled by the comparative approach from sample SRS011126 of the Human Microbiome Project, and the gap between them is about 1.3Kbp based on their coordinates relative to the reference genome *Treponema denticola* ATCC 35405 (Figure 6.18(a)). The gap could be caused by several reasons, e.g., deletion, recombination or low depth of sequencing coverage, and we hope that the hybrid approach could help us bridge this gap. In the hybrid assembly these two contigs are connected by a contig (named as Soap1) from the *de novo* assembly (Figure 6.18(b)), and the gap between them is about 0.5Kbp based on the coordinates relative to this new contig. This 0.5Kbp gap is much shorter than the original 1.3Kbp gap, and also these two gap sequences are not similar to each other (as determined by BLASTN alignment), indicating a potential genomic recombination event. Then we analyzed the gene annotations from the GenBank record (Figure 6.18(a) and Table 6.2) around these two contigs in the reference genome, and have identified genes that might cause this genome recombination event, e.g., XerD phage integrase family site-specific recombinase.

Gene name	Start	End	Strand	GenBank ID	Annotation
BspA	2792010	2792750	+	NP_973335	Cell surface antigen BspA.
PepF	2792903	2794633	+	NP_973336	Oligoendopeptidase F.
EamA	2794813	2795229	+	NP_973337	EamA-like transporter family.
RmsS	2795342	2795830	+	NP_973338	Type I restriction-modification system.
TprL	2796003	2796782	+	NP_973339	TPR repeat lipoprotein.
XerD	2796940	2797749	−	NP_973340	Phage integrase: site-specific recombinase.
RmsS	2797810	2798337	+	NP_973341	Type I restriction-modification system.

Table 6.2: Gene annotations around the contigs Cmp1 and Cmp2 based on reference genome *Treponema denticola* ATCC 35405.

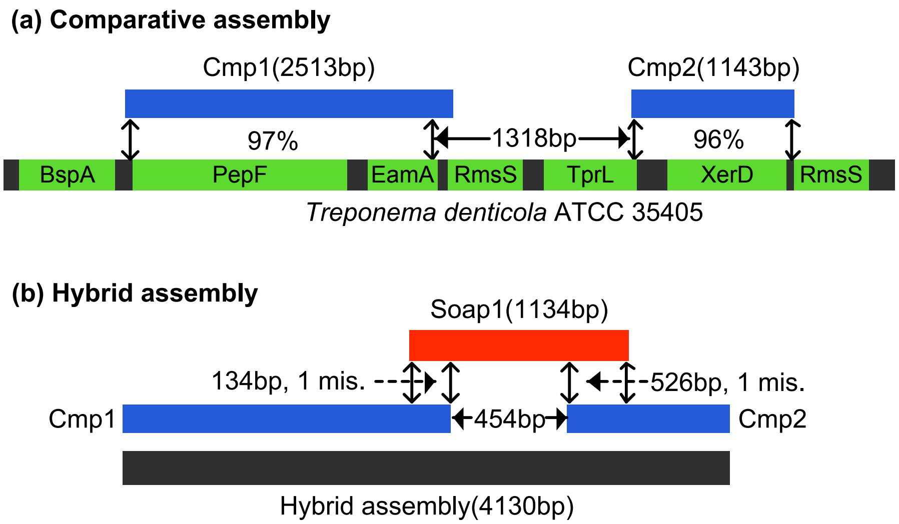


Figure 6.18: Hybrid assembly of contigs from comparative (Cmp1 and Cmp2) and *de novo* (Soap1) assemblies. Figure (a) shows the creation of contigs Cmp1 and Cmp2 guided by the reference genome Treponema denticola ATCC 35405. Figure (b) shows that Cmp1 and Cmp2 can be connected by contig Soap1 from *de novo* assembly. Soap1 overlaps with Cmp1 by 134bp with 1 mismatch, and overlaps with Cmp2 by 526bp with 1 mismatch.

6.3.6 Running time and memory usage

We compared the running time and peak memory usage of the assemblers (Figure 6.19) using the even mock metagenomic data set. The peak memory usage of MetaCompass is much smaller than that of other assemblers under comparison (Figure 6.19b). The memory usage of MetaCompass is invariant to the number of the metagenomic reads to be assembled, rather it linearly depends on the size of the reference genomes (about 8 bytes per base pair). The running time is worse than the three *de novo* assemblers (Figure 6.19a), but it grows linearly with the number of metagenomic reads, and can be easily run in multiple threads or processes to achieve linear speedup (parallel execution is available through command-line option). Also, during the initial reference genome selection by MetaPhyler, instead of using all the reads, the users can sample a subset of the reads to estimate the taxonomic composition. The reasons why MetaCompass is slower than MC-Ref (MetaCompass with given true reference genomes) are: we need to run MetaPhyler at the beginning, and MetaPhyler usually proposes a much larger reference genomes set than the truth.

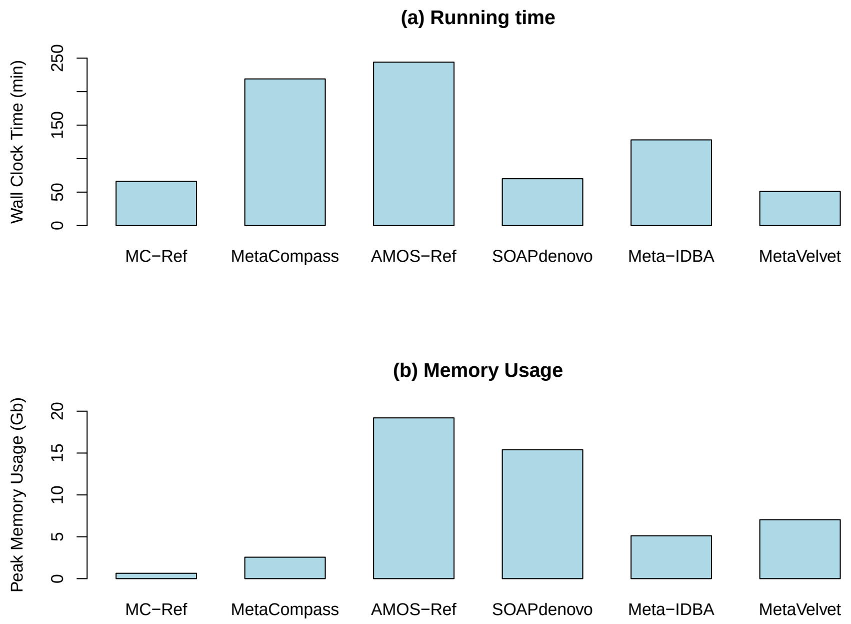


Figure 6.19: Comparison of running time and peak memory usage on the even mock metagenomic sample. Running time is calculated as the CPU usage time in a single thread. MC-Ref and AMOS-Ref assume the true reference genomes are given when running MetaCompass and AMOScmp assemblers.

6.4 Conclusions

In this project, we have developed a new comparative genome assembler specifically for metagenomic sequences. We show that MetaCompass can achieve comparable results to the state-of-the-art *de novo* assembler, and much better than

AMOScmp, a single genome comparative assembler. Furthermore, we show that MetaCompass and *de novo* assemblies are quite complementary to each other, and the combination of these two approaches can significantly improve individual performance. Overall, MetaCompass is highly applicable to the analysis of large-scale metagenomic data sets, and also can run on typical personal computers with small memory capacity. The algorithm, however, is not perfect, for example, although the performance of reference genomes selection in MetaCompass is very good in the mock metagenomic datasets, it is still significantly worse than when we know the reference genomes exactly. So one future direction for research is improved reference genome selection, e.g., by analyzing the depth of coverage of different regions across the genomes. This will avoid assigning the reads that come from a similar region across genomes to a single genome, and will prevent unnecessary gaps between contigs.

Chapter 7

Conclusion

During my PhD research, I have developed efficient algorithms and tools primarily for large-scale metagenomic sequences analysis. The tools and algorithms have been used by many people in the research community, and helped them understand the underlying biological systems from different perspectives, discover interesting patterns or phenomena in their own data sets. We have demonstrated the efficiency and capability of our tools by applying them to several real-world biological data sets, including the largest available so far, the Human Microbiome Project, which has generated near 10 terabytes of raw sequences. My research has also resulted in interesting biological findings.

In addition, I want to summarize my research in terms of contributions from a computational perspective (for details see Section 1.3). During the development of the above-mentioned computational algorithms and tools, we have employed, customized and improved previous algorithms significantly. MetaPath is a statistical algorithm that can be adapted to compare any weighted networks. Previous methods usually, partitioned the network into a set of discrete features. Also we have introduced another nonparametric p-value which takes into account the network structure. ARDB database is specifically designed for the information retrieval and analysis of a particular set of genes. The infrastructure, data integration and representation approach can be generalized to building databases for other biological genes. Because a lot of biomedical information are stored as semi-structured data, ARDB provides a framework for compiling and building specialized gene databases. MetaPhyler is a hierarchical training and classification software for biological sequences. It automatically learns the similarity distributions within each group induced by the given taxonomy. During classification, MetaPhyler computes a confidence score for each classification at each hierarchical level, and hence it can easily identify novel taxonomic group from the query. In addition to taxonomic classification, MetaPhyler can also be adapted to functional annotation of biological sequences. Furthermore, the algorithm can also be applied to text classification when a hierarchy is available (e.g., music has subtypes such as pop, country music; both music and sports belong to entertainment). MetaCompass is the first algorithm and tool that allows efficient comparative assembly of metagenomic sequences, and it improves the overall metagenomic assembly performance by about 40%.

We will continue improve our algorithms and maintain our tools to accommodate the fast-growing sequencing technologies, and we hope our tools will continue benefit the general research community.

Appendix A

Comparative analysis of antibiotic resistance genes reveals insights into the evolution of drug resistance in *Staphyloccus aureus*

A.1 Background

Staphylococcus aureus, a Gram-positive bacterium discovered in the 1880s, is a leading cause of hospital-acquired infections affecting as many as 2% of newly admitted patients [134]. S. aureus is the most frequent cause of surgical site infections, lower respiratory tract infections, and cardiovascular infections and is the second most frequent cause of healthcare - associated pneumonia and blood stream infections [103]. According to recent estimates, >400,000 S. aureus related hospitalizations occurred per year in the United States, causing 11,000 deaths annually [64]. In 1942 the first penicillin-resistant S. aureus isolate was observed in a hospital just two years after the introduction of penicillin in medical use. In 1961, two years after the introduction of methicillin, S. aureus developed resistance through the acquisition of mecA gene. During the last 45 years, various hospital-associated methicillin-resistant S. aureus (HA-MRSA) clones spread worldwide. Since the 1990s, more virulent community-associated MRSA (CA-MRSA) clones have emerged and become increasingly prevalent. The ability of S. aureus to develop resistance to virtually all classes of antimicrobial agents increasingly complicates efforts to prevent and treat infections, especially in hospitalized patients. Morbidity and mortality caused by drug resistant S. aureus strains have increased the financial burden on health care systems and lead to strong interest in studying the evolution (in terms of acquisition and distribution of antibiotic resistance genes) of this organism, which is poorly understood [93].

A.1.1 Antibiotic Resistance in Staphylococci

Methicillin resistance in staphylococci is conferred by the mecA gene encoding a penicillin-binding protein PBP2', which catalyses the cross-linking of cell wall peptidoglycan and has a low affinity for beta-lactam antibiotics. Staphylococcal cassette chromosome (SCC) elements are, so far, the only vectors described for the mecA gene, and this large fragment of mobile DNA is designated SCCmec. SCCmec elements can integrate into the genome of staphylococci at a unique site (attBSCC) near the origin of replication [57]. According to their putative cassette chromosome recombinase genes (ccr) and overall genetic compositions, six different types of SCCmec have been identified and characterized [94]. Staphylococcal genomes seem to change continuously as genetic elements move in and out, but no mechanism has been found responsible for transferring SCC elements between different staphylococcal species.

One of the key features of S. aureus has been its ability to readily acquire resistance to antibiotics, often through lateral gene transfer. The adaptability and resilience of S. aureus is highlighted by the large number of MRSA strains that plague our hospitals worldwide (HA-MRSA) and by new virulent MRSA clones that have emerged recently as a major cause of severe community-acquired (CA-MRSA) infections [33]. Differentiating between community and hospital-acquired strains is difficult. Two putative characteristic features of CA-MRSA are the presence of type IV SCCmec and PantonValentine leukocidin (PVL) gene. However, no studies have proven the exact biological roles of these elements in the pathogenesis of CA-MRSA. In fact, CA-MRSA strains exist that possess neither SCCmec IV nor the PVL locus [38, 109]. Although antibiotic resistance genes are important genetic factors in the pathogenesis of MRSA, no comprehensive analysis of these genes has been performed.

The rapidly increasing volume of genomic sequence data generated from staphylococci represents a tremendous resource for studies of bacterial diversity and evolution. At the time of writing, 18 complete and 5 draft Staphylococcus genomes are available from the NCBI genome database with many more sequencing projects underway. In this study we compare the sequenced Staphylococcus genomes on the basis of their antibiotic resistance profiles as predicted by a newly created database of antibiotic resistance factors [81]. Our analysis reveals a strong correlation of antibiotic resistance genes with the adaptation of S. aureus to community and hospital environments. By tracking the horizontal gene transfer events of antibiotic resistance genes and SCCmec elements in the sequenced Staphylococcus genomes, and mapping them to the phylogenetic tree, we show that these transposable elements can be acquired by different Staphylococcus lineages independently, strongly supporting the hypothesis that MRSA strains have emerged on multiple independent occasions. Finally, we conduct a survey of staphylococcal antibiotic resistance genes and demonstrate that different staphylococcal strains share a common reservoir of antibiotic resistance genes.

A.2 Results and Discussion

A.2.1 Correlation of Niche Adaptation (Ecological Fitness) with Antibiotic Resistance

Community acquired and hospital acquired MRSA have distinct phenotypic characteristics, e.g. CA-MRSA strains are more virulent, grow faster and contain fewer resistance genes. To better understand the genetic underpinnings of the differences between CA-MRSA and HA-MRSA, we compared 23 Staphylococcus genomes (Table A.1) on the basis of their antibiotic resistance genes, annotated using the ARDB database [81]. Although these different staphylococci have a relatively short evolutionary history, especially within the S. aureus lineage, the number of resistance genes varied dramatically. For example, S. aureus N315 carries 17 antibiotic resistance genes, but only three antibiotic resistance genes are identified in S. aureus MW2 (a CA-MRSA strain) and two are identified in S. aureus RF122 (a bovine isolate) (Figure A.1). Such large variation in antibiotic resistance gene content might play a role in the ecological adaptation of individual strains and pathogenic potential, especially because these strains are frequently subject to antibiotic treatment.

Organism	Strain Source	SCCmec Type	MLST Type	Clonal Complex
S. aureus COL	Early Isolate	I	250	8
S. aureus JH1	Hospital	IIb	105	5
S. aureus JH9	Hospital	IIb	105	5
S. aureus MRSA252	Hospital	IIa	36	30
S. aureus MSSA476	Community	NA	1	1
S. aureus Mu3	Hospital	IIa	5	5
S. aureus Mu50	Hospital	IIa	5	5
S. aureus MW2	Community	IV	1	1
S. aureus N315	Hospital	IIa	5	5
S. aureus NCTC8325	Lab Strain	NA	8	8
S. aureus Newman	Early Isolate	NA	254	8
S. aureus RF122	Bovine	NA	151	151
S. aureus USA300	Community	IV	8	8
S. aureus USA300 TCH1516	Community	IV	8	8
S. epidermidis RP62A	Hospital	IIb	NA	NA
S. haemolyticus JCSC1435	Hospital	N1	NA	NA
S. epidermidis ATCC 12228	Noninfectious Strain	NA	NA	NA
S. saprophyticus ATCC 15305	Community	NA	NA	NA
S. aureus CF-Marseille	Hospital	IVa	NA	NA
S. aureus JKD6008	Hospital	IIa	239	8
S. aureus JKD6009	Hospital	IIa	239	8
S. aureus Mu50-omega	Hospital	IIa	5	5
S. aureus TCH959	Community	NA	8	8

Table A.1: 18 complete and 5 draft (the last five rows) Staphylococcus genomes.

Hospital acquired staphylococci carry more antibiotic resistance genes (mean: 11.5, standard deviation: 2.3) than their community acquired counterparts (mean: 5.8, standard deviation: 3.4) or other staphylococcal strains (mean: 3.4, standard deviation: 1.2) (Figure A.1). This observation is unsurprising as HA strains need to survive in the antibiotic-intensive hospital environment. One exception is S. aureus TCH1516, a community acquired strain that acquired eight antibiotic resistance genes within a plasmid. S. aureus TCH1516 has the potential to colonize a hospital environment, consistent with its increasing prevalence. When comparing the number of antibiotic resistance genes only within the main chromosome, the separation between CA- (mean: 3.5, standard deviation: 0.5) and HA-MRSA (mean: 10.3, standard deviation: 2.4) strains becomes more significant. An explanation could be that HA-MRSA strains lose the environmental fitness to compete with CA-MRSA in a community environment because of the metabolic burden imposed by these resis-

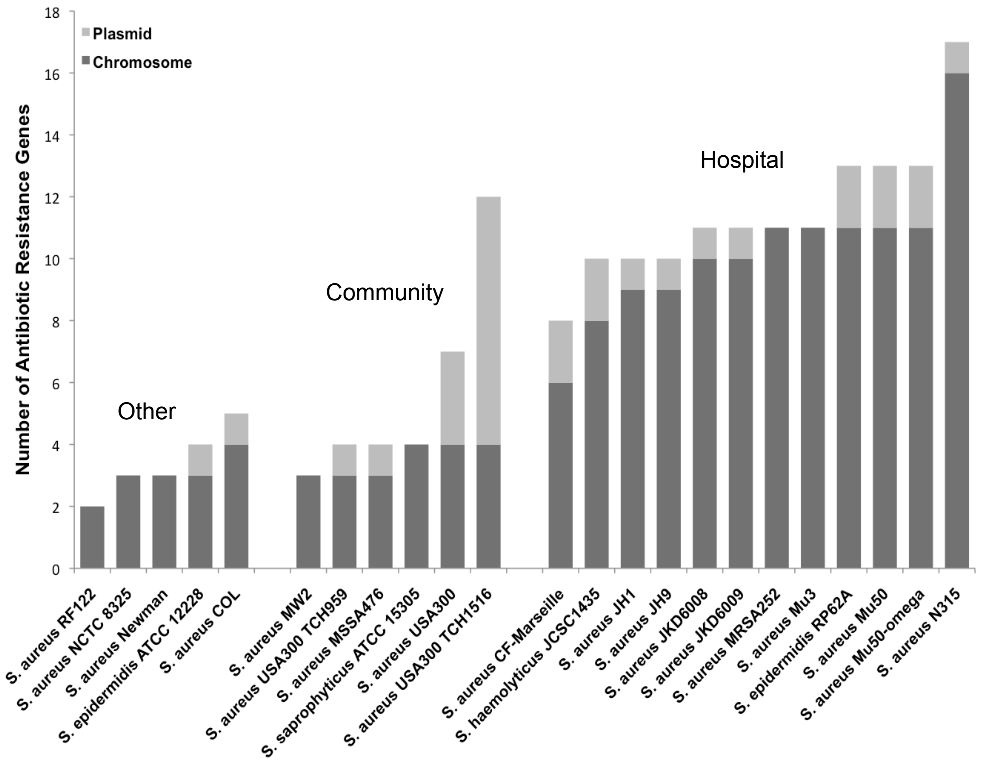


Figure A.1: Comparison of the number of antibiotic resistance genes in staphylococcal genomes isolated from different environments. The antibiotic resistance genes for each staphylococcal genome are computationally predicted and summarized. Based on the habitat source, staphylococcal strains are classified into three groups: hospital-acquired, community-acquired and other (non-human pathogens, lab strains, or uncertain origin). Dark grey bars represent antibiotic resistance genes found in chromosomes and light bars represent antibiotic resistance genes found in plasmids.

tance genes, which in turn are necessary to survive in a hospital setting due to high drug pressure. Compared with its close relative S. aureus USA300 (another community acquired strain), S. aureus TCH1516 has the same resistance genes within the main chromosome. The fact that most of the resistance genes within the S. aureus TCH1516 are on a plasmid, which is different from the plasmid found in S. aureus USA300, may allow it to easily modulate their expression in order to avoid high metabolic costs and therefore retaining fitness in an environmental setting.

Frequently, HA-MRSA strains contain many duplicated (more than one copy of a same resistance gene) or redundant (more than one gene conferring the same resistance profile) resistance genes , possibly providing an explanation for the higher abundance of resistance genes within HA strains. This dosage effect, however, does not fully explain the difference between hospital and community-acquired staphylococci. HA strains appear to also have more diverse resistance profiles, in terms of the spectrum of antibiotics they can tolerate. To assess this effect we compared the genomes on the basis of their resistance profile alone, ignoring the actual number of genes conferring a certain type of resistance. Specifically, each Staphylococcus genome was represented as a 0/1 characteristic string where each character indicates whether this strain is resistant (0 for sensitive; 1 for resistant) to a specific antibiotic or a group of antibiotics. We then constructed a parsimony tree (Figure A.2; see methods for details) from the resulting strings. In this antibiotic resistance profile tree, HA-MRSA and CA-MRSA strains are located in two distinct clusters, highlighting their different resistance profiles. As an outgroup we chose a hypothetical strain sensitive to all antibiotics (a character string with all zeroes). As seen in the figure, the hospital-acquired strains have resistance to a broader spectrum of antibiotics than their community-acquired counterparts. All HA-MRSA strains have the methicillin-resistant gene mecA, but only some CA-MRSA strains have this gene. For example, S. aureus TCH959 and S. aureus MSSA476 are CA-MRSA strains that lack the mecA gene. Most HA-MRSA strains are resistant to bleomycin (a glycopeptide antibiotic inhibiting DNA metabolism) conferred by the ble gene, while none of the CA-MRSA strains are resistant to this antibiotic. All HA-MRSA strains are resistant to macrolide, lincosamide and streptogramin B (MLSB) antibiotics, but only S. aureus USA300 and TCH1516 in CA-MRSA are resistant to MLSB, which explains why these strains cluster close to the HA-MRSA cluster (Figure A.2). In recent years, replacement of MRSA strains in some hospitals by CA-MRSA strains with enhanced drug resistance has been observed [116]. The resistance phenotypes of S. aureus USA300 and TCH1516 may indicate their potential to colonize hospital environment.

A.2.2 Horizontal Gene Transfer of Antibiotic Resistance Genes and SCCmec element

Two opposing theories have previously been suggested to describe the relationship between the first MRSA isolates and recent MRSA clones. The single-clone theory suggests that all MRSA clones have a common ancestor, and that SCCmec element was introduced only once into S. aureus [67]; the multi-clone theory hypothesizes that SCCmec was introduced several times into different S. aureus genetic lineages [35,93]. Distinguishing between these two theories requires an accurate reconstruction of the evolutionary history of S. aureus strains, which is difficult due to the short evolutionary time since the emergence of MRSA strains. Methods based on genome rearrangement and inversion [50] and multi locus sequence typing [34] have been proposed. However, S. aureus genomes do not contain sufficient genome rearrangement events [38]. Further, MLST sequences frequently do not have enough discriminatory power because of the short evolutionary history of

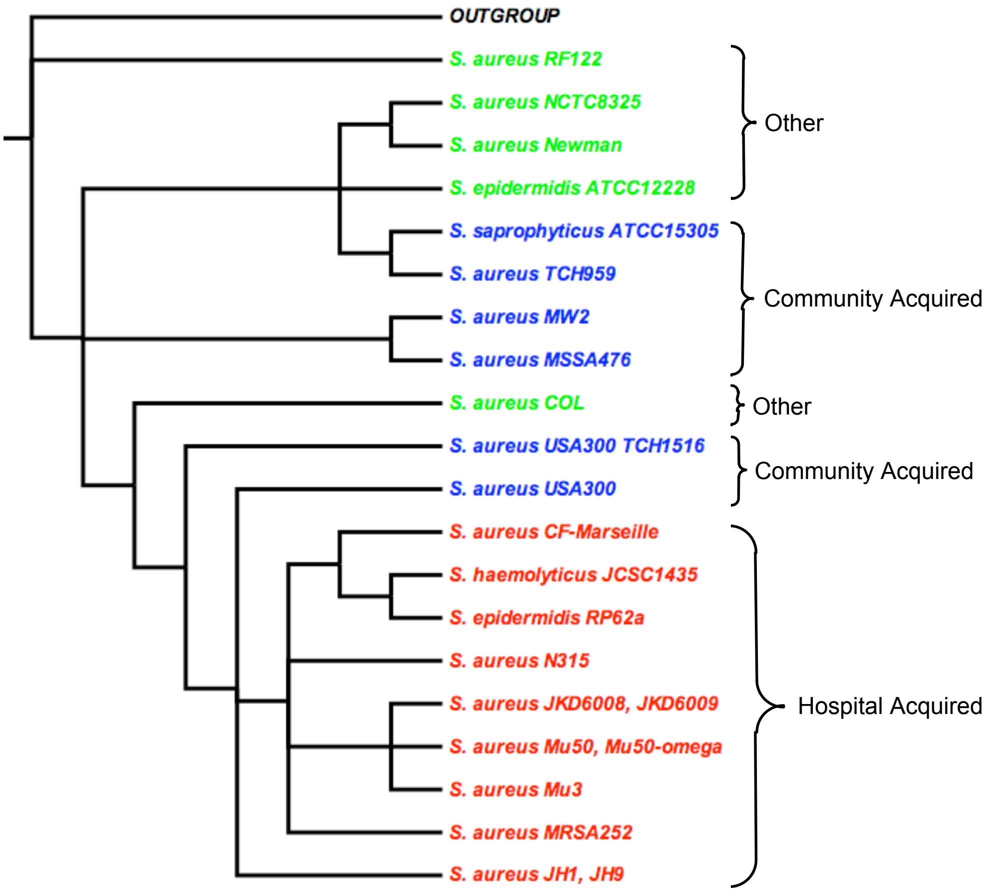


Figure A.2: Maximum parsimony tree of antibiotic resistance profiles for 23 staphylococcal genomes. The antibiotic resistance profile for each staphylococcal genome is computationally predicted and represented as a 0/1 characteristic string where each character indicates whether this strain is resistant (0 for sensitive; 1 for resistant) to a particular antibiotic. Species in blue represent CA-MRSA strains, in red represent HA-MRSA strains and in green represent non-human pathogens, lab strains or strains of uncertain origin.

MRSA. For example, both S. aureus TCH1516 and S. aureus NCTC8325 belong to ST8 (type 8 of Multi Locus Sequence Typing), but they have distinct antibiotic resistance phenotypes. To build a reliable phylogenetic tree of all sequenced staphylococcal genomes, we used 31 housekeeping genes as proposed in [136](See methods for details). The structure is highly consistent with the clonal complex classification (Table A.1), in which every member of a complex shares six or seven MLST alleles with at least one other member [37], indicating the phylogenetic tree correctly recovers the evolutionary history. Afterwards, we mapped putative HGT events of antibiotic resistance genes onto the phylogenetic tree using the following rule. If a resistance gene is inserted into the same locus within two chromosomes, then we suggest that these two bacteria share the same HGT origin. In other words, these two bacteria have diverged after the insertion of resistance gene into their common ancestor. For example, S. aureus Mu3 and S. aureus Mu50 contain the same tetM gene, and this gene has been inserted into the same locus of the two chromosomes.

Most of the acquisitions of antibiotic resistance genes and SCCmec elements occur near or at terminal leave of the phylogenetic tree (Figure A.3), indicating that these transferable elements are acquired by different staphylococcal lineages independently and these resistance genes are not native genes in the staphylococcal lineage. The only genes that deviate from this pattern are fosB and mepA resistance genes indicating that these genes are native to S. aureus strains. FosB is a metallothiol transferase that confers fosfomycin resistance and has about 60% percent identity with the fosB gene in Bacillus. MepA is a multi antimicrobial extrusion efflux (MATE) family protein, contributes to the multidrug resistance in staphylococci [60] and is only identified in S. aureus. These observations indicate that originally S. aureus strains only contained few resistance genes, and the highly resistant phenotype was developed through rapid evolution [52].

Different SCCmec elements have been imported eleven times into different phylogenetic lineages, supporting the hypothesis that MRSA has arisen on many in-

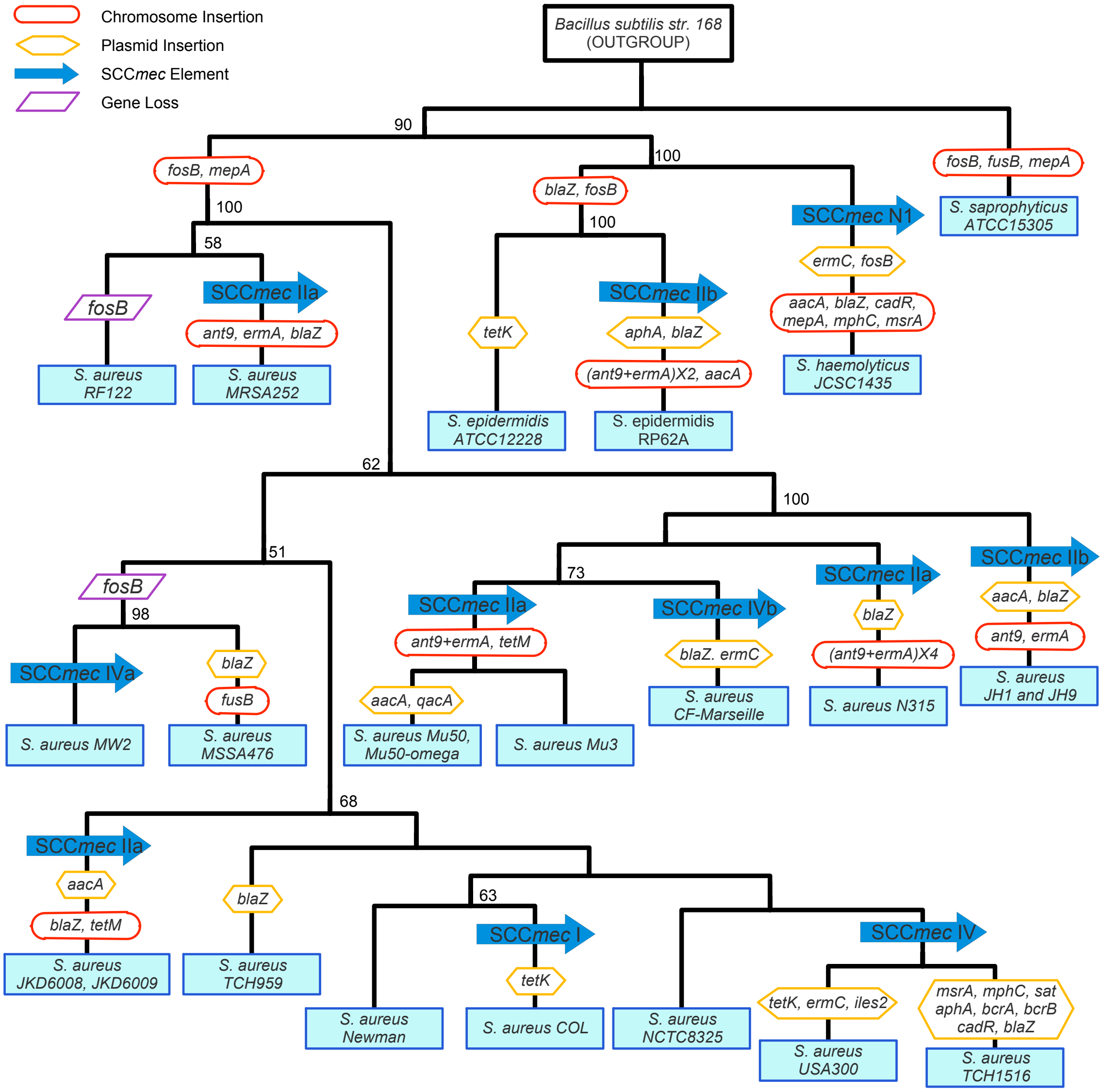


Figure A.3: Phylogenetic tree of 23 staphylococcal genomes (Bacillus subtilis str. 168 as outgroup) is constructed on the basis of 31 housekeeping marker genes using neighbor joining. The tree is annotated with horizontal gene transfer events of antibiotic resistance genes and SCCmec elements. Resistance genes associated with SCCmec elements are listed in Table S9. The numbers on the intersections indicate bootstrap value after 100 trials (500 trials performed and normalized to 100), and bootstrap values below 50 are not indicated. (ant9+ermA)X2 represents that ant9 and ermA resistance genes are located on a same transposon, and these two genes have been inserted twice in different loci of the chromosome. Blue arrows represent SCCmec element insertions, yellow hexagons represent plasmid resistance genes acquisition, red round rectangles represent chromosomal resistance genes acquistition and purple parallelograms represent resistance gene loss from chromosome.

dependent occasions through the local acquisitions of SCCmec elements, and could not have descended from the diversification of a single original MRSA clone. The resistance genes associated with each SCCmec element are shown in Table S9. The pUB110 plasmid, which encodes aadD and ble resistance genes, has been integrated into SCCmec IIa. In addition, both SCCmec IIa and IIb carry the Tn554, a high frequency and site-specific transposable element [8]. In some genomes (e.g., S. aureus N315), Tn554 has also been inserted multiple times into other loci. Virtually all S. aureus strains have the potential to acquire resistance to a broad range of antibiotics. For example, S. aureus MRSA252 is distantly related to other sequenced S. aureus strains [52], but it has acquired a type II SCCmec element, which is identical with that from S. aureus Mu3, S. aureus Mu50, and S. aureus N315, and a blaZ gene, which is widely distributed among S. aureus strains. Antibiotic resistance genes and SCCmec elements are not shared only within S. aureus strains; they have also been horizontally transferred to other commensals from the Staphylococcus genus. For example, compared with its close relative - the non-infectious strain S. epidermidis ATCC12228 - S. epidermidis RP62A has acquired several antibiotic resistance genes that have been found in S. aureus and a type II SCCmec element, which is identical with that from S. aureus JH1 and JH9. S. haemolyticus JCSC1435 has acquired a novel type SCCmec element, which only carries the mecA resistance gene. However, it has also acquired eight other antibiotic resistance genes conferring resistance to all major antibiotics. These observations indicate that HGT of antibiotic resistance genes among multiple staphylococcal species occurs commonly, allowing for adaptation to shifting host environments.

Figure A.3 demonstrates marked diversity in the distribution of antibiotic resistance genes, indicating that mobile DNA is exchanged readily in the S. aureus population. For example, the type IIa SCCmec element has been found in three different clonal complexes (CC5, CC30 and CC8; Table A.1). Sometimes it can be imported into other non aureus staphylococcal species. The heterogeneity of the genetic background in MRSA, shown in A.3, indicates that many if not all S. aureus have the potential to become drug resistant. One example is the observation of the re-emergence of early pandemic S. aureus strains that have acquired mecA and become a MRSA clone [106]. However, the spread of antibiotic resistance genes does not correlate with the phylogenetic relationship inferred from the phylogenetic tree A.3, suggesting that the mobile elements play an important role in facilitating the rapid changes of drug resistance potential in S. aureus.

A.2.3 HGT of Antibiotic Resistance Genes in Human Staphylococcal Pathogens Follows Founder Effect

Staphylococci have been shown to be able to acquire resistance genes from the local habitat, implying the presence of a resistance gene reservoir within staphylococci or other organisms that co-colonize the same environment or host [46]. To characterize this antibiotic resistance genes repository, we retrieved from the NCBI protein database all the resistance genes, belonging to several prevalent types found in staphylococci. The number and variability of antibiotic resistance genes present in the natural bacterial populations are huge [81], but only a few are identified in staphylococci, especially in human staphylococcal pathogens. In addition, resistance genes isolated from animal staphylococci are generally more diverse, containing novel resistance types that are not present in human staphylococcal pathogens. This phenomenon can be described by the 'founder effect' model in population genetics, which refers to the loss of genetic variation when a new colony is established by a very small number of individuals from a larger population.

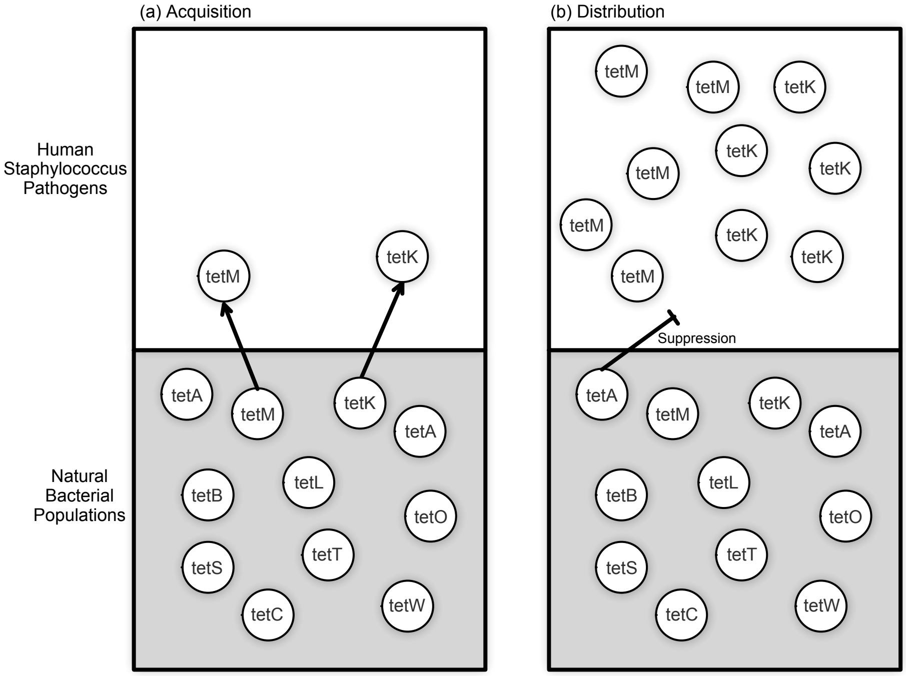


Figure A.4: Founder effect in the acquisition and distribution of antibiotic resistance genes in human staphylococcal pathogens. Bottom gray box represents natural bacterial populations, and top white box represents human Staphylococcus pathogens. Resistance gene is represented by circle with name in it. (a) Natural environment contains various tetracycline resistance genes, but only tetM and tetK are transferred into the human Staphylococcus pathogens. (b) tetM and tetK resistance genes are spread in human staphylococcal pathogens, and they also impede the entrance of other similar tetracycline resistance genes.

In natural environment, more than 30 different types of tetracycline resistance genes have been found belonging to two functional groups: major facilitator superfamily (MFS) efflux transporter and ribosome protection protein (RPP) [105].

Only tetK (MFS) and tetM (RPP) resistance types are found and spread in human staphylococcal pathogen population, although these pathogens have been isolated from a great variety of locations. TetK is only found and distributed in staphylococci, and all the tetK type genes have very high sequence similarities (>98%). TetM, a ribosome protection protein, has been identified in more than 20 different genera (the pairwise similarity ranges from 80% to 100%), but all tetM genes in staphylococci have very high similarities (>96%). Moreover, tetM resistance genes isolated from human staphylococcal pathogens from diverse locations have even higher similarity (>99%). For example, tetM gene NP_390922 isolated from Japan is exactly the same as gene ACH85960 from Denmark, and is 99% identical with genes AAA26678 and AAL27024 isolated from US and Poland respectively. TetL is another MFS type tetracycline resistance gene, and this gene has only been found in S. hyicus and S. epidermidis from animals. These facts indicate that the tetracycline resistance genes repository available for human staphylococcal pathogens is limited, despite the large diversity of these genes within the natural environment.

We further evaluated if the acquisition and distribution of other resistance genes follow the same pattern as tetracycline resistance genes. mecA is the only gene responsible for methicillin resistance, and this gene is only found in staphylococci. All mecA genes in human staphylococcal pathogens (including non-aureus species, such as S. epidermis RP62A and S. haemolyticus JCSC1435) have sequence similarities above 99%, although some mecA genes isolated from animals and other environments are more diverse. For example, mecA gene CAJ15584 in S. kloosii isolated from a horse has only 89% sequence similarity with those isolated from human staphylococci. Beta-lactamase resistance genes are generally classified into 4 classes belonging to more than 40 different types [13], but only two types, denoted as blaZ1 and blaZ2, have been identified in staphylococci. Only one blaZ2 gene (ABQ23577) has been isolated from staphylococci in an S. epidermis strain, and this gene is highly prevalent in Klebsiella pneumoniae, indicating this group of organisms as its probable source [118]. BlaZ1 genes are mostly found in staphylococci (with only two exceptions: AAA24777 and AAB40888 from Enterococcus faecalis) with high sequences similarities (above 95%), and are widely distributed in both human and animal staphylococcal pathogens. With regard to macrolide, lincosamide and streptogramin B antibiotics resistance, about 20 different types of rRNA methylases (erm) resistance genes are identified, but only four types are found in staphylococci.

These results indicate that human staphylococcal pathogens generally share a common source of antibiotic resistance genes, and the exchanges between this reservoir and other environments are rare. This agrees with the idea that the range and frequencies of HGT are constrained most often by selective barriers (Kurland, et al., 2003) between different environments. The reasons for this phenomenon could be that certain DNA structures favour the combination of certain elements; only transposable elements and resistance genes coexisting in the same environment are transferable; and the element conferring a selective advantage quickly spreads and predominates within the bacterial population under antibiotic threat.

A.3 Conclusion

We have investigated the correlation of antibiotic resistance with the adaptation of staphylococci in the community and hospital habitats. The analysis is based on a newly developed antibiotic resistance factors database [81]. Our results strongly indicate that HA-MRSA strains contain significantly more antibiotic resistance genes than CA-MRSA strains, especially within the main chromosome, which agrees with previous observations [25]. We propose antibiotic resistance genes in general as another characteristic marker for the HA- and CA-MRSA strains besides SCCmec and the PVL element. By mapping the HGT events of antibiotic resistance genes and SCCmec elements onto the phylogenetic tree of Staphylococcus genomes, we showed that these transposable elements are frequently acquired independently, providing the evidence that the emergence of MRSA strains has occurred repeatedly. In addition, we compared all the antibiotic resistance genes belonging to several types (tet, mecA, blaZ and erm), and found that only limited versions of resistance genes are distributed in human staphylococcal pathogens, despite the fact that some of the strains are isolated from distinct locations.

Since the emerging S. aureus pathogens create profound threats to public health, tools for rapid detection and characterization of the microbes are a critical necessity. Functional gene arrays targeting antibiotic resistance and virulence [58] have been developed as a potential solution. In addition, SCCmec typing [17] and multi locus sequence typing [36] have been widely used for S. aureus strains, however because of the rapid evolution of antibiotic resistance, these two methods can not fully discriminate and characterize different clones. The analyses described in this paper indicate that diagnostic assays based on resistance genes could become a better and more biologically meaningful approach. Because antibiotic resistance evolves rapidly under drug pressure, it is difficult if not impossible to capture these characteristics using other marker genes. Typing resistance genes directly not only provides more discriminatory power, but also gives us an overview of the resistance profiles, which can be used to design more efficient treatment strategies.

A.4 Materials and methods

A.4.1 Genome Sequences and Annotation

The DNA sequences of 23 Staphylococcus genomes were obtained from the NCBI genome database. Protein sequences for each genome were extracted from the corresponding features annotation, and antibiotic resistance genes were annotated using the tools from Antibiotic Resistance Genes Database [81].

A.4.2 Data Analysis

Using the identified antibiotic resistance genes and the associated resistance profile, the resistance phenotype for each Staphylococcus strain is represented by a 0/1 characteristic string based on the presence or absence of resistance genes in its genome. 0 represents sensitive to the antibiotic and 1 represents resistance. The outgroup is represented by a string with all 0s, indicating it does not contain any resistance gene. A maximum parsimony tree for the Staphylococcus genomes (Figure

A.2) was constructed based on these characteristic strings using PHYLIP software package (pars program), using a maximum parsimony algorithm that minimizes the number of character state changes, method widely used for morphological data.

The phylogenetic tree in Figure 3 was constructed on the basis of 31 marker genes [136], with Bacillus subtilis subsp. subtilis str. 168 as an outgroup. The protein sequences of the marker genes were concatenated accordingly for each genome, and were aligned using CLUSTALW [124]. Then the alignment was bootstrapped 500 times, and a neighbor-joining tree was constructed and summarized using protdist, neighbor and consense programs in PHYLIP. To visualize the HGT events of antibiotic resistance genes and SCCmec elements, all of them are painted onto the phylogenetic tree using the following rule. If two staphylococci share a same resistance gene with same locus on the chromosomes, then we define that these two bacteria share the same HGT origin. In other words, these two bacteria evolutionarily diverge after their common ancestor acquires the resistance gene. For example, S. aureus Mu3 and S. aureus Mu50 contain the same tetM gene, and this gene has been inserted into the same locus of the two chromosomes, then we define these two strains share the same tetM insertion origin.

To retrieve all the antibiotic resistance genes belonging to a general type (e.g. tet, mecA, blaZ and erm), a representative sequence is selected for each subtype of these four types. There are more than 30 tet subtypes, and we produce a representative gene sequence for each of them. We focus on subtypes as genes belonging to a same general type may have very low sequence similarities, although they confer the same antibiotic resistance phenotype.

Bibliography

[1] J. Ahn, L. Yang, B. J. Paster, I. Ganly, L. Morris, Z. Pei, and R. B. Hayes. Oral microbiome profiles: 16s rrna pyrosequencing and microarray assay comparison. *PLoS One*, 6(7):e22788, 2011.

[2] M. N. Alekshun and S. B. Levy. Molecular mechanisms of antibacterial multidrug resistance. *Cell*, 128(6):1037–50, 2007.

[3] H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder, and J. Handelsman. Functional metagenomics reveals diverse beta-lactamases in a remote alaskan soil. *ISME J*, 3(2):243–51, 2009.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[5] E. A. Bancroft. Antimicrobial resistance: it's not just for hospitals. *JAMA*, 298(15):1803–4, 2007.

[6] M. Barber. Methicillin-resistant staphylococci. *J Clin Pathol*, 14:385–93, 1961.

[7] V. M. Barnes, R. Teles, H. M. Trivedi, W. Devizio, T. Xu, M. W. Mitchell, M. V. Milburn, and L. Guo. Acceleration of purine degradation by periodontal diseases. *J Dent Res*, 88(9):851–5, 2009.

[8] M. C. Bastos and E. Murphy. Transposon tn554 encodes three products required for transposition. *EMBO J*, 7(9):2935–41, 1988.

[9] O. Beja, L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–6, 2000.

[10] E. M. Bik, C. D. Long, G. C. Armitage, P. Loomer, J. Emerson, E. F. Mongodin, K. E. Nelson, S. R. Gill, C. M. Fraser-Liggett, and D. A. Relman. Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J*, 4(8):962–74, 2010.

[11] A. Brady and S. L. Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods*, 6(9):673–6, 2009.

[12] M. M. Brinig, P. W. Lepp, C. C. Ouverney, G. C. Armitage, and D. A. Relman. Prevalence of bacteria of division tm7 in human subgingival plaque and their association with disease. *Appl Environ Microbiol*, 69(3):1687–94, 2003.

[13] K. Bush, G. A. Jacoby, and A. A. Medeiros. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob Agents Chemother*, 39(6):1211–33, 1995.

[14] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. Allpaths: de novo assembly of whole-genome shotgun microreads. *Genome Res*, 18(5):810–20, 2008.

[15] K. Chen and L. Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*, 1(2):106–12, 2005.

[16] T. Chen, W. H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)*, 2010:baq013, 2010.

[17] P. Chongtrakool, T. Ito, X. X. Ma, Y. Kondo, S. Trakulsomboon, C. Tiensasitorn, M. Jamklang, T. Chavalit, J. H. Song, and K. Hiramatsu. Staphylococcal cassette chromosome mec (sccmec) typing of methicillin-resistant staphylococcus aureus strains isolated in 11 asian countries: a proposal for a new nomenclature for sccmec elements. *Antimicrob Agents Chemother*, 50(3):1001–12, 2006.

[18] Y. Cicek, I. Ozmen, V. Canakci, A. Dilsiz, and F. Sahin. Content and composition of fatty acids in normal and inflamed gingival tissues. *Prostaglandins Leukot Essent Fatty Acids*, 72(3):147–51, 2005.

[19] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, A. M. Bandela, E. Cardenas, G. M. Garrity, and J. M. Tiedje. The ribosomal database project (rdp-ii): introducing myrdp space and quality controlled public data. *Nucleic Acids Res*, 35(Database issue):D169–72, 2007.

[20] The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, 06 2012.

[21] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 06 2012.

[22] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2nd revised edition edition, September 2001.

[23] P. Courvalin. Vancomycin resistance in gram-positive cocci. *Clin Infect Dis*, 42 Suppl 1:S25–34, 2006.

[24] V. M. D'Costa, K. M. McGrann, D. W. Hughes, and G. D. Wright. Sampling the antibiotic resistome. *Science*, 311(5759):374–7, 2006.

[25] H. de Lencastre, D. Oliveira, and A. Tomasz. Antibiotic resistant staphylococcus aureus: a paradigm of adaptive power. *Curr Opin Microbiol*, 10(5):428–35, 2007.

[26] J. A. Delaney, V. Schneider-Lindner, P. Brassard, and S. Suissa. Mortality after infection with methicillin-resistant staphylococcus aureus (mrsa) diagnosed in the community. *BMC Med*, 6:2, 2008.

[27] M. L. Diaz-Torres, R. McNab, D. A. Spratt, A. Villedieu, N. Hunt, M. Wilson, and P. Mullany. Novel tetracycline resistance determinant from the oral metagenome. *Antimicrob Agents Chemother*, 47(4):1430–2, 2003.

[28] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Muller. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, 24(13):i223–31, 2008.

[29] B. E. Dutilh, M. A. Huynen, and M. Strous. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics*, 25(21):2878–81, 2009.

[30] R. H. Eckel. Obesity and heart disease: a statement for healthcare professionals from the nutrition committee, american heart association. *Circulation*, 96(9):3248–50, 1997.

[31] J. S. Edwards, M. Covert, and B. Palsson. Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol*, 4(3):133–40, 2002.

[32] W. Eisenreich, T. Dandekar, J. Heesemann, and W. Goebel. Carbon metabolism of intracellular bacterial pathogens and possible links to virulence. *Nat Rev Microbiol*, 8(6):401–12, 2010.

[33] M. C. Enright. The evolution of a resistant pathogen–the case of mrsa. *Curr Opin Pharmacol*, 3(5):474–9, 2003.

[34] M. C. Enright, N. P. Day, C. E. Davies, S. J. Peacock, and B. G. Spratt. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of staphylococcus aureus. *J Clin Microbiol*, 38(3):1008–15, 2000.

[35] M. C. Enright, D. A. Robinson, G. Randle, E. J. Feil, H. Grundmann, and B. G. Spratt. The evolutionary history of methicillin-resistant staphylococcus aureus (mrsa). *Proc Natl Acad Sci U S A*, 99(11):7687–92, 2002.

[36] M. C. Enright and B. G. Spratt. A multilocus sequence typing scheme for streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology*, 144 (Pt 11):3049–60, 1998.

[37] E. J. Feil, B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt. eburst: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol*, 186(5):1518–30, 2004.

[38] Y. Feng, C. J. Chen, L. H. Su, S. Hu, J. Yu, and C. H. Chiu. Evolution and pathogenesis of staphylococcus aureus: lessons learned from genotyping and comparative genomics. *FEMS Microbiol Rev*, 32(1):23–37, 2008.

[39] M. R. Fokkema, H. A. Woltil, C. M. van Beusekom, A. Schaafsma, D. A. Dijck-Brouwer, and F. A. Muskiet. Plasma total homocysteine increases from day 20 to 40 in breastfed but not formula-fed low-birthweight infants. *Acta Paediatr*, 91(5):507–11, 2002.

[40] S. Gallistl, K. Sudi, H. Mangge, W. Erwa, and M. Borkenstein. Insulin is an independent correlate of plasma homocysteine levels in obese children and adolescents. *Diabetes Care*, 23(9):1348–52, 2000.

[41] W. Gerlach, S. Junemann, F. Tille, A. Goesmann, and J. Stoye. Webcarma: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10:430, 2009.

[42] M. Ghodsi, B. Liu, and M. Pop. Dnaclust: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, 12:271, 2011.

[43] T. A. Gianoulis, J. Raes, P. V. Patel, R. Bjornson, J. O. Korbel, I. Letunic, T. Yamada, A. Paccanaro, L. J. Jensen, M. Snyder, P. Bork, and M. B. Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A*, 106(5):1374–9, 2009.

[44] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–9, 2006.

[45] M. Hamady and R. Knight. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*, 19(7):1141–52, 2009.

[46] A. M. Hanssen and J. U. Ericson Sollid. Sccmec in staphylococci: genes on the move. *FEMS Immunol Med Microbiol*, 46(1):8–20, 2006.

[47] I. Hellmann, Y. Mang, Z. Gu, P. Li, F. M. de la Vega, A. G. Clark, and R. Nielsen. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res*, 18(7):1020–9, 2008.

[48] D. Hernandez, P. Francois, L. Farinelli, M. Osteras, and J. Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on

a desktop computer. *Genome Res*, 18(5):802–9, 2008. Hernandez, David Francois, Patrice Farinelli, Laurent Osteras, Magne Schrenzel, Jacques Research Support, Non-U.S. Gov't United States Genome research Genome Res. 2008 May;18(5):802-9. Epub 2008 Mar 10.

[49] C. F. Higgins. Multiple molecular mechanisms for multidrug resistance transporters. *Nature*, 446(7137):749–57, 2007.

[50] S. K. Highlander, K. G. Hulten, X. Qin, H. Jiang, S. Yerrapragada, Jr. Mason, E. O., Y. Shang, T. M. Williams, R. M. Fortunov, Y. Liu, O. Igboeli, J. Petrosino, M. Tirumalai, A. Uzman, G. E. Fox, A. M. Cardenas, D. M. Muzny, L. Hemphill, Y. Ding, S. Dugan, P. R. Blyth, C. J. Buhay, H. H. Dinh, A. C. Hawes, M. Holder, C. L. Kovar, S. L. Lee, W. Liu, L. V. Nazareth, Q. Wang, J. Zhou, S. L. Kaplan, and G. M. Weinstock. Subtle genetic changes enhance virulence of methicillin resistant and sensitive staphylococcus aureus. *BMC Microbiol*, 7:99, 2007.

[51] S. Hirsch, J. Poniachick, M. Avendano, A. Csendes, P. Burdiles, G. Smok, J. C. Diaz, and M. P. de la Maza. Serum folate and homocysteine levels in obese females with non-alcoholic fatty liver. *Nutrition*, 21(2):137–41, 2005.

[52] M. T. Holden, E. J. Feil, J. A. Lindsay, S. J. Peacock, N. P. Day, M. C. Enright, T. J. Foster, C. E. Moore, L. Hurst, R. Atkin, A. Barron, N. Bason, S. D. Bentley, C. Chillingworth, T. Chillingworth, C. Churcher, L. Clark, C. Corton, A. Cronin, J. Doggett, L. Dowd, T. Feltwell, Z. Hance, B. Harris, H. Hauser, S. Holroyd, K. Jagels, K. D. James, N. Lennard, A. Line, R. Mayes, S. Moule, K. Mungall, D. Ormond, M. A. Quail, E. Rabbinowitsch, K. Rutherford, M. Sanders, S. Sharp, M. Simmonds, K. Stevens, S. Whitehead, B. G. Barrell, B. G. Spratt, and J. Parkhill. Complete genomes of two clinical staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A*, 101(26):9786–91, 2004.

[53] L. V. Hooper and J. I. Gordon. Commensal host-bacterial relationships in the gut. *Science*, 292(5519):1115–8, 2001.

[54] P. Husemann and J. Stoye. r2cat: synteny plots and comparative assembly. *Bioinformatics*, 26(4):570–1, 2010.

[55] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. Megan analysis of metagenomic data. *Genome Res*, 2007.

[56] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–40, 2002.

[57] T. Ito, Y. Katayama, and K. Hiramatsu. Cloning and nucleotide sequence determination of the entire mec dna of pre-methicillin-resistant staphylococcus aureus n315. *Antimicrob Agents Chemother*, 43(6):1449–58, 1999.

[58] C. Jaing, S. Gardner, K. McLoughlin, N. Mulakken, M. Alegria-Hartman, P. Banda, P. Williams, P. Gu, M. Wagner, C. Manohar, and T. Slezak. A functional gene array for detection of bacterial virulence elements. *PLoS One*, 3(5):e2163, 2008.

[59] P. L. Johnson and M. Slatkin. Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res*, 16(10):1320–7, 2006.

[60] G. W. Kaatz, F. McAleese, and S. M. Seo. Multidrug resistance in staphylococcus aureus due to overexpression of a novel multidrug and toxin extrusion (mate) transport protein. *Antimicrob Agents Chemother*, 49(5):1857–64, 2005.

[61] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–4, 2008.

[62] B. J. Keijser, E. Zaura, S. M. Huse, J. M. van der Vossen, F. H. Schuren, R. C. Montijn, J. M. ten Cate, and W. Crielaard. Pyrosequencing analysis of the oral microflora of healthy adults. *J Dent Res*, 87(11):1016–20, 2008.

[63] L. Kesavalu, V. Bakthavatchalu, M. M. Rahman, J. Su, B. Raghu, D. Dawson, G. Fernandes, and J. L. Ebersole. Omega-3 fatty acid regulates inflammatory cytokine/mediator messenger rna expression in porphyromonas gingivalis-induced experimental periodontal disease. *Oral Microbiol Immunol*, 22(4):232–9, 2007.

[64] E. Klein, D. L. Smith, and R. Laxminarayan. Hospitalizations and deaths caused by methicillin-resistant staphylococcus aureus, united states, 1999-2005. *Emerg Infect Dis*, 13(12):1840–6, 2007.

[65] S. Koren, T. J. Treangen, and M. Pop. Bambus 2: scaffolding metagenomes. *Bioinformatics*, 27(21):2964–71, 2011.

[66] L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye. Phylogenetic classification of short environmental dna fragments. *Nucleic Acids Res*, 36(7):2230–9, 2008.

[67] B. Kreiswirth, J. Kornblum, R. D. Arbeit, W. Eisner, J. N. Maslow, A. McGeer, D. E. Low, and R. P. Novick. Evidence for a clonal origin of methicillin resistance in staphylococcus aureus. *Science*, 259(5092):227–30, 1993.

[68] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev*, 72(4):557–78, Table of Contents, 2008.

[69] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, and M. Hattori. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*, 14(4):169–81, 2007.

[70] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2):R12, 2004.

[71] B. Lai, R. Ding, Y. Li, L. Duan, and H. Zhu. A de novo metagenomic assembly program for shotgun dna reads. *Bioinformatics*, 2012.

[72] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9(4):357–9, 2012.

[73] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[74] J. Laserson, V. Jojic, and D. Koller. Genovo: de novo assembly for metagenomes. *J Comput Biol*, 18(3):429–43, 2011.

[75] I. Letunic, T. Yamada, M. Kanehisa, and P. Bork. ipath: interactive exploration of biochemical pathways and networks. *Trends Biochem Sci*, 33(3):101–3, 2008.

[76] R. E. Ley, D. A. Peterson, and J. I. Gordon. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4):837–48, 2006.

[77] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[78] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[79] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, H. Yang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–72, 2010.

[80] B. Liu, T. Gibbons, M. Ghodsi, T. Treangen, and M. Pop. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, 12 Suppl 2:S4, 2011.

[81] B. Liu and M. Pop. Ardb–antibiotic resistance genes database. *Nucleic Acids Res*, 37(Database issue):D443–7, 2009.

[82] Bo Liu, Lina L. Faller, Niels Klitgord, Varun Mazumdar, Mohammad Ghodsi, Daniel D. Sommer, Theodore R. Gibbons, Todd J. Treangen, Yi-Chien Chang, Shan Li, O. Colin Stine, Hatice Hasturk, Simon Kasif, Daniel Segr, Mihai Pop, and Salomon Amar. Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PLoS ONE*, 7(6):e37919, 06 2012.

[83] Bo Liu, T. Gibbons, M. Ghodsi, and M. Pop. Metaphyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 95 –100, 2010.

[84] Bo Liu and Mihai Pop. Identifying differentially abundant metabolic pathways in metagenomic datasets. In *ISBRA*, volume 6053 of *Lecture Notes in Computer Science*, pages 101–112. Springer, 2010.

[85] Y. Marcy, C. Ouverney, E. M. Bik, T. Losekann, N. Ivanova, H. G. Martin, E. Szeto, D. Platt, P. Hugenholtz, D. A. Relman, and S. R. Quake. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated tm7 microbes from the human mouth. *Proc Natl Acad Sci U S A*, 104(29):11889–94, 2007.

[86] A. C. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length dna fragments. *Nat Methods*, 4(1):63–72, 2007.

[87] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386, 2008.

[88] R. Mojtabai. Body mass index and serum folate in childbearing age women. *Eur J Epidemiol*, 19(11):1029–36, 2004.

[89] M. M. Mwangi, S. W. Wu, Y. Zhou, K. Sieradzki, H. de Lencastre, P. Richardson, D. Bruce, E. Rubin, E. Myers, E. D. Siggia, and A. Tomasz. Tracking the in vivo evolution of multidrug resistance in staphylococcus aureus by whole-genome sequencing. *Proc Natl Acad Sci U S A*, 104(22):9451–6, 2007.

[90] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. Metavelvet: an extension of velvet assembler to de novo metagenome

assembly from short sequence reads. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '11, pages 116–124, New York, NY, USA, 2011. ACM.

[91] Giuseppe Narzisi and Bud Mishra. Comparing de novo genome assembly: The long and short of it. *PLoS ONE*, 6(4):e19175, 04 2011.

[92] H. Noguchi, J. Park, and T. Takagi. Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*, 34(19):5623–30, 2006.

[93] U. Nubel, P. Roumagnac, M. Feldkamp, J. H. Song, K. S. Ko, Y. C. Huang, G. Coombs, M. Ip, H. Westh, R. Skov, M. J. Struelens, R. V. Goering, B. Strommenger, A. Weller, W. Witte, and M. Achtman. Frequent emergence and limited geographic dispersal of methicillin-resistant staphylococcus aureus. *Proc Natl Acad Sci U S A*, 105(37):14130–5, 2008.

[94] D. C. Oliveira, C. Milheirico, and H. de Lencastre. Redefining a structural variant of staphylococcal cassette chromosome mec, sccmec type vi. *Antimicrob Agents Chemother*, 50(10):3457–9, 2006.

[95] D. H. Parks and R. G. Beiko. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–21, 2010.

[96] Y. Peng, H. C. Leung, S. M. Yiu, and F. Y. Chin. Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–101, 2011.

[97] R. Pike, V. Lucas, P. Stapleton, M. S. Gilthorpe, G. Roberts, R. Rowbury, H. Richards, P. Mullany, and M. Wilson. Prevalence and antibiotic resistance profile of mercury-resistant oral bacteria from children with and without mercury amalgam fillings. *J Antimicrob Chemother*, 49(5):777–83, 2002.

[98] M. Podar, C. B. Abulencia, M. Walcher, D. Hutchison, K. Zengler, J. A. Garcia, T. Holland, D. Cotton, L. Hauser, and M. Keller. Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol*, 73(10):3205–14, 2007.

[99] M. Pop. Genome assembly reborn: recent computational challenges. *Brief Bioinform*, 10(4):354–66, 2009.

[100] M. Pop, A. Phillippy, A. L. Delcher, and S. L. Salzberg. Comparative genome assembly. *Brief Bioinform*, 5(3):237–48, 2004.

[101] Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[102] J. Raes, K. U. Foerstner, and P. Bork. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, 10(5):490–8, 2007.

[103] M. J. Richards, J. R. Edwards, D. H. Culver, and R. P. Gaynes. Nosocomial infections in pediatric intensive care units in the united states. national nosocomial infections surveillance system. *Pediatrics*, 103(4):e39, 1999.

[104] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–52, 2004.

[105] M. C. Roberts. Update on acquired tetracycline resistance genes. *FEMS Microbiol Lett*, 245(2):195–203, 2005.

[106] D. A. Robinson, A. M. Kearns, A. Holmes, D. Morrison, H. Grundmann, G. Edwards, F. G. O'Brien, F. C. Tenover, L. K. McDougal, A. B. Monk, and M. C. Enright. Re-emergence of early pandemic staphylococcus aureus as a community-acquired meticillin-resistant clone. *Lancet*, 365(9466):1256–8, 2005.

[107] B. Rodriguez-Brito, F. Rohwer, and R. A. Edwards. An application of statistics to comparative metagenomics. *BMC Bioinformatics*, 7:162, 2006.

[108] J. I. Ross, E. A. Eady, J. H. Cove, and W. J. Cunliffe. 16s rrna mutation associated with tetracycline resistance in a gram-positive bacterium. *Antimicrob Agents Chemother*, 42(7):1702–5, 1998.

[109] A. S. Rossney, A. C. Shore, P. M. Morgan, M. M. Fitzgibbon, B. O'Connell, and D. C. Coleman. The emergence and importation of diverse genotypes of methicillin-resistant staphylococcus aureus (mrsa) harboring the panton-valentine leukocidin gene (pvl) reveal that pvl is a poor marker for community-acquired mrsa strains in ireland. *J Clin Microbiol*, 45(8):2554–63, 2007.

[110] A. Sboner, X. J. Mu, D. Greenbaum, R. K. Auerbach, and M. B. Gerstein. The real cost of sequencing: higher than you think! *Genome Biol*, 12(8):125, 2011.

[111] J. Scaria, U. Chandramouli, and S. K. Verma. Antibiotic resistance genes online (argo): a database on vancomycin and beta-lactam resistance genes. *Bioinformation*, 1(1):5–7, 2005.

[112] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, and C. Huttenhower. Metagenomic biomarker discovery and explanation. *Genome Biol*, 12(6):R60, 2011.

[113] J. Sekiguchi, T. Miyoshi-Akiyama, E. Augustynowicz-Kopec, Z. Zwolska, F. Kirikae, E. Toyota, I. Kobayashi, K. Morita, K. Kudo, S. Kato, T. Kuratsuji, T. Mori, and T. Kirikae. Detection of multidrug resistance in mycobacterium tuberculosis. *J Clin Microbiol*, 45(1):179–92, 2007.

[114] L. A. Selvey, M. Whitby, and B. Johnson. Nosocomial methicillin-resistant staphylococcus aureus bacteremia: is it any worse than nosocomial methicillin-sensitive staphylococcus aureus bacteremia? *Infect Control Hosp Epidemiol*, 21(10):645–8, 2000.

[115] L. A. Seville, A. J. Patterson, K. P. Scott, P. Mullany, M. A. Quail, J. Parkhill, D. Ready, M. Wilson, D. Spratt, and A. P. Roberts. Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic dna. *Microb Drug Resist*, 15(3):159–66, 2009.

[116] U. Seybold, E. V. Kourbatova, J. G. Johnson, S. J. Halvosa, Y. F. Wang, M. D. King, S. M. Ray, and H. M. Blumberg. Emergence of community-associated methicillin-resistant staphylococcus aureus usa300 genotype as a major cause of health care-associated blood stream infections. *Clin Infect Dis*, 42(5):647–56, 2006.

[117] I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi. Pathway-based functional analysis of metagenomes. *J Comput Biol*, 18(3):495–505, 2011.

[118] H. H. Sheng, Y. Qu, X. M. Wu, Y. Y. Dong, X. T. Zeng, H. Cao, X. W. Huang, H. Q. Yin, Y. Q. Yu, Y. X. Ni, and H. S. Xiao. A new blalen-17 gene in a clinical isolate of staphylococcus epidermidis in shanghai, china. *Chin Med J (Engl)*, 121(3):272–5, 2008.

[119] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. Abyss: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–23, 2009.

[120] D. D. Sommer, A. L. Delcher, S. L. Salzberg, and M. Pop. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics*, 8:64, 2007.

[121] M. O. Sommer, G. Dantas, and G. M. Church. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*, 325(5944):1128–31, 2009.

[122] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5, 2003.

[123] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–6, 2000.

[124] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, 1994.

[125] S. G. Tringe and P. Hugenholtz. A renaissance for the pioneering 16s rrna gene. *Curr Opin Microbiol*, 11(5):442–6, 2008.

[126] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7, 2005.

[127] I. J. Tsai, T. D. Otto, and M. Berriman. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol*, 11(4):R41, 2010.

[128] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–10, 2007.

[129] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31, 2006.

[130] Turnbaugh, P. J. *et al*. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4, 2009.

[131] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–30, 2007.

[132] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, 73(16):5261–7, 2007.

[133] J. R. White, N. Nagarajan, and M. Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, 5(4):e1000352, 2009.

[134] W. Witte, C. Cuny, I. Klare, U. Nubel, B. Strommenger, and G. Werner. Emergence and spread of antibiotic-resistant gram-positive bacterial pathogens. *Int J Med Microbiol*, 298(5-6):365–77, 2008.

[135] J. C. Wooley and Y. Ye. Metagenomics: Facts and artifacts, and computational challenges*. *J Comput Sci Technol*, 25(1):71–81, 2009.

[136] M. Wu and J. A. Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):R151, 2008.

[137] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N. Baldassano, Andrey P. Anokhin, Andrew C. Heath, Barbara Warner, Jens Reeder, Justin Kuczynski, J. Gregory Caporaso, Catherine A. Lozupone, Christian Lauber, Jose Carlos Clemente, Dan Knights, Rob Knight, and Jeffrey I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, advance online publication:–, 05 2012.

[138] E. Zaura, B. J. Keijser, S. M. Huse, and W. Crielaard. Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol*, 9:259, 2009.

[139] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–9, 2008.

[140] C. E. Zhou, J. Smith, M. Lam, A. Zemla, M. D. Dyer, and T. Slezak. Mvirdb–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*, 35(Database issue):D391–4, 2007.