Computational Approaches for Analyzing Ancient Genomes and Modern Metagenomes

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät der
Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Dipl.-Inform. (Bioinformatik)
Nico Weber
aus Oberhausen

Tübingen

2013

Tag der mündlichen Prüfung:	06.02.2014
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Daniel H. Huson
2. Berichterstatter:	Prof. Dr. Johannes Krause

Abstract

Modern genomics entered a new era with the invention of next-generation sequencing techniques. Technical progress, high throughput and reasonably cheap costs of the systems enable us to look into the genomic sequences of whole communities or even extinct species. In the first part of this work we present and discuss state-of-the-art methods for analyzing metagenomes efficiently. As the assignment of sequencing reads to known species or functions is one key element in the analysis we discuss currently used methods. Those methods are usually either slow or do not provide all necessary information, such as genome alignments, for a detailed analysis. Here we present a novel approach, which is faster compared to previous methods while still providing genome alignments. Database composition and the assignment of database entries to species or functions is an equally important step during a metagenomic analysis. We inspect how well the taxonomy is covered by commonly used databases such as the NCBI-NR database. We also evaluate the efficiency of assignment methods using either plain text or RefSeq accession numbers to map reference sequences to taxa or functions. In this context we present a method using a the GenBank identifier for classifying reference sequences. Validation using an in vitro simulated metagenomic dataset shows that the new approach can assign more reads to function or taxa. At the same time the new approach is more specific than the previously used methods.

The huge amounts of data and the steadily increasing number of samples require an initial investment of time and effort to be able to analyze the incoming data efficiently. Interdisciplinary work and external collaboration partners emphasize the need for a flexible approach to present intermediate steps during the analysis and sharing of the final results. Here we present a local instance of the workflow system galaxy which was used in the different projects throughout this thesis.

In the second part of this thesis we analyze ancient DNA samples which are suspected to be infected with ancient M. tuberculosis. Ancient strains have the potential of giving insight into evolution and distribution of extinct pathogens. Screening for potentially interesting samples was done using a whole genome shotgun approach. An additional screening was performed by sequencing samples which were enriched for four specific genes. For the final analysis we performed a genome wide enrichment prior to sequencing as ancient samples often yield only very low amounts of DNA. Design of the enrichment chip is discussed as well as the subsequent analysis. In the end of the analysis consensus sequences for three ancient strains are calculated. Single nucleotide polymorphisms are determined as a base for a downstream phylogentic analysis.

Acknowledgments

First I'd like to thank my supervisor Prof. Daniel Huson for the opportunity of undertaking a PhD in his group. I am very thankful for the constant guidance while at the same time having the freedom to explore.

I am also grateful to Prof. Johannes Krause for the chance to take part in such an exciting project and for agreeing in reviewing this thesis. I very much appreciate the inspiring discussions.

While pursuing my PhD I had the chance to work with many people in Tübingen and around the globe. I have greatly benefited from the fruitful debates throughout this time.

I also acknowledge the Landesgraduiertenförderung for my initial funding.

During my time in Tübingen I met many interesting people either in or outside the University. I want to thank all of them for providing a friendly atmosphere and the shared experiences.

I am deeply grateful to my family and friends for all the time spend together and for reminding me what really matters.

But most of all I owe my deepest gratitude to Lena for her cheerfulness and constant support.

In accordance with the standard scientific protocol, I will use the personal pronoun *we* to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

"The story so far:
In the beginning the Universe was created.
This has made a lot of people very angry and been widely regarded as a bad move."

Douglas Adams, The Restaurant at the End of the Universe

CHAPTER 1

Introduction

All known forms of life on our planet rely on a macromolecule called *Deoxyribonucleic acid (DNA)* to define their complete genome. The genome of the organism determines all reactions within each single cell and is therefore responsible for the whole appearance, the fitness and even up to some part the behavior of the organism in question. To achieve this the DNA encodes for a huge number of other macromolecules which are mediating between, code for or interact with proteins and regulatory structures. Even with the observed diversity of life in all habitats of earth, organisms share the same genomic content up to some point. The size of the shared information varies from only single genes over basic cell functions or pathways up to complete body features. The information shared between different organisms depend on their common evolutionary history.

More than 150 years after the initial discovery of the DNA by Friedrich Miescher in 1869, we are able to read the complete genomic content of a single organism in reasonable time and costs. Especially in the last 20 years DNA sequencing made huge steps forward catapulting modern biology into a new area.

Projects like the *Human Genome Project* (HGP) [1, 2] helped in focusing money and scientists and were a major motor in the development of new technologies. With the HGP finished in 2003 after 13 years of runtime the race for the *$1,000 Genome* was started. In January 2013 the Archon Genomics X-Price [3] started offering a reward of $10 million for the first team or company which manages to achieve this ambitious goal.

Nevertheless reading and understanding the genomic code and its function are two separate sides of the coin, so voices increasingly reminding us to keep the big picture in mind (e.g. "The $1,000 genome, the $100,000 analysis?"[4]).

This thesis discusses two fields descending from classic genomics: *Metagenomics* and *ancient DNA research*.

Metagenomics

Compared to genomics the field of metagenomics is a pretty young field and after 15 years of the first metagenomic study [5] published it is still vastly changing. Especially the introduction of newly developed sequencing methods enabled even smaller labs to undertake their own metagenomic studies.

Metagenomics is often defined as the science of analyzing the genomic content of an environmental sample. Popular locations are either nature samples such as soil and water or host-based (e.g. gut, nose etc.).

One motivation of metagenomics is the fact that only a fraction of the microbes can be cultured at a lab. Estimates suggest that 99.6% of the human microbiota can not be cultured through lab techniques [6]. The sole investigation of the remaining 0.4% for sure neglects the full potential, wastes a huge amount of information and hinders the full understanding of the whole community as it is. Another aspect is the chance of understanding microbial communities and interaction on various levels as seen in their natural environment. Early and popular projects such as the *Sargasso Sea Survey* [7] already employed whole genome shotgun (WGS) sequencing, but also targeted sequencing approaches exist. Targeted sequencing approaches often focus on phylogenetic markers such as 16S rRNA to reconstruct taxonomic composition of the samples whereas WGS projects have the advantage of possibly identifying novel genes and function previously not known.

Big projects such as the *Human Microbiome Project* [8] investigate microbe-community-host-interaction and are pushing general development and interest forward.

Sampling the environment and library preparation of the sample have their own challenges, but globally seen there are typical questions in each metagenomic project which need to be addressed by the downstream analysis.

Over the years various computational tools have been published to help analyzing metagenomes. The range of tools varies from database comparison tools [9, 10] over phylogentic predictors [11, 12] and visualisation tools [13] and finally to complete analysis pipelines [14, 15].

Despite the developments there is still need for new analysis methods as samples, databases, data volumes and complexity tend to grow rapidly [16]. This thesis will cover the different problems in detail and discusses novel approaches to solve them.

Ancient DNA Research

Nearly twice as old as the field of metagenomics is the field of ancient DNA (aDNA) research. The first publications now date back nearly 30 years [17, 18] using clonal amplification before the invention of PCR. Those early studies of aDNA already revealed that recovered aDNA mainly consists of damaged and short fragmented DNA in very low concentration. High copy number regions such as the mitochondrial DNA were yielding the best results. Protocols using PCR enabled the investigation of very low amounts of (endogenous) aDNA but also increase the potential of possible contamination during lab processes. The introduction of modern sequencing technologies literally amplified the impact that PCR made to the field and enabled various exciting publications. Publications describing the (draft) genomes of the extinct *wolly mammoth* [19] and *Neanderthal* [20] demonstrate that increased throughput now allows nearly full genome sequencing of ancient samples. But even with new technologies contamination with modern DNA, DNA concentration and DNA damage are major concerns of modern aDNA research [21, 22].

The ability to investigate extinct species is especially interesting with respect to pathogen evolution. Recent studies [23, 24] were able to reconstruct ancient strains enabling an unique view into the past. This work will investigate multiple ancient samples of remains which are suspected to be infected with *Mycobacterium tuberculosis* and discuss challenges met along the way. This includes different wet-lab approaches as well as computational solutions.

Overview

Metagenomics and ancient DNA research both challenge current analysis methods with the complexity of the sample and the huge amount of data.

In the first part of the thesis we present and discuss current and novel computational approaches for analyzing metagenomes.

For this we first introduce current sequencing technologies and modern methods in genomics in Chapter 2. We will also focus on specific metagenomic needs during the analysis. Chapter 4 will introduce methods for quality control to assure a high standard analysis. Assigning reads to specific species or functions which is a demanding task is discussed in Chapter 5 where also a novel hybrid method for classification of metagenomic reads is presented.

Database accuracy and mapping efficiency of downstream analysis is evaluated in Chapter 6 including suggestions for further improvements.

The last section of the first part introduces a workflow management system to enable multiple investigators easy and reproducible access to the same data.

In the second part of the thesis we focus on the investigation of ancient DNA. DNA in all non-living organisms is subject to change forced on it by the environment resulting in specific needs during the analysis. Chapter 8 will introduce systematic modifications encountered during aDNA research. This chapter will also introduce current theories of the evolution of *Mycobacterium tuberculosis*. In Section 8.2 we investigate aDNA samples using three different approaches including a specifically designed capture array to optimize the amount of sequenced aDNA of interest.

Conclusion and outlook in Part IV will summarize our findings and discusses potential fields for further research.

Part I

Background on Sequencing, Genomics and Metagenomics

This part introduces the technologies and methods building the foundation for genomics and metagenomics. In the beginning we will present state-of-the-art sequencing technologies as a base for all following analysis and discuss the advantages and drawbacks of the current methods. We will introduce common approaches and challenges in the field of metagenomics. The last chapter of this part focuses on special computational challenges.

Chapter 2

Technology: Sequencing

Completing the first full sequence of the human genome was a milestone for modern genomics. Only 10 years later modern genomics has evolved into a field with a many interesting subjects at least as exciting as the beginning questions.

In this chapter we will present the history and state of the art methods of DNA sequencing and give an introduction into genomics and metagenomics. After discussing basic questions and challenges we focus on challenges related to the computational analysis.

2.1 Sequencing Technologies

The genomic content of an organism is described by a sequence consisting only of the four bases *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T). Reading and determining the order of the recurring bases forming the genetic code is referred to as DNA sequencing.

In the end of the 1970s the first two methods for DNA sequencing were developed by Maxam and Gilbert [25] as well as Sanger et al. [26]. Up to now the procedures advanced and were supplemented or replaced by new technologies. By the end of 2012 we have multiple sequencing technologies available ranging from the *first generation* of Sanger Sequencing, over massively parallel sequencing up to single molecule sequencing.

Only with few exceptions the basic approach for sequencing is to mimic or use the function of DNA polymerase to synthesize a new strand based on a DNA template out of modified *deoxyribonucleotides (dNTPs)*. The different technologies now use various techniques to determine the order and amount of nucleotides incorporated and therefore can reconstruct the original sequence of the template DNA strand. Detection is mainly based on optical systems, but also a few different promising systems are on the market. Although most systems share the basic idea, they differ in the exact execution as well as preparation, efficiency and potential biases.

2.1.1 Sanger sequencing

The original Sanger sequencing method [26] is based on chain termination techniques which uses modified bases missing a 3'-hydroxy group inhibiting further extension of the DNA strand. Those *dideoxyribonucleoside triphospates (ddNTPs)* where originally radioactively labeled and mixed with unmodified dNTPs, which don't stop the elongation of the strand. Only a fraction of ddNTPs were mixed with dNTPs. This proportion ensures that the copy process was stopped only on a random basis. Using a DNA polymerase together with a primer and the mix the sequence of interest gets amplified. During this process the DNA polymerase gets repeatedly stopped by ddNTPs, resulting in multiple strands of different length with the last base labeled. Using a gel electrophoresis the fragments can be ordered by their length, representing the complement DNA sequence of the template. Over the years Sanger sequencing was modified further including non radioactive labeling, automating of the process and various other improvements. Up until today for some people Sanger sequencing is still the gold standard in sequencing [27].

2.1.2 Massively Parallel sequencing

With the advent of new sequencing technologies the term *next generation* or *second generation sequencing* was coined to describe new technologies trying to succeed Sanger sequencing. The rapid developments with more and more new techniques makes it difficult to draw the line between second generation sequencing and *third* or even *forth* generations. Most new methods have in common that template DNA is shattered into fragments and then amplified to a specific amount to facilitate detection of the signals generated during nucleotide incorporation. For this the fragments need to be immobilized to ensure that multiple copies of the strands are at close physical range. The most common techniques for amplification of the template DNA are emulsion-PCR [28] and bridge amplification [29]. Emulsion-PCR binds single strand DNA templates to beads and encapsulates them into tiny bioreactors within an emulsion. The DNA is then amplified on the surface of the bead. After amplification the beads are placed into micro-wells where the sequencing reaction will take place. Bridge amplification uses adaptors to attach template strands to a surface on which they are clonally amplified. The surface later contains clusters of DNA representing a single template at high spatial concentration, amplifying the signal generated during sequencing.

With the new approaches in development a new sub-type of sequencing was developed and defined as mate-pair or paired-end sequencing. This model is supported by multiple sequencing technologies and based on the library preparation prior to amplification. The new method has the advantage of getting pairs of reads from the same strand with a known orientation and distance between them. Downstream analysis strongly benefits from mate-pair libraries, because e.g. regions longer than the single read length can be resolved. This can be beneficial if there are repetitive elements included in the original strand.

Besides their similarities and differences in template preparation and amplification the modern technologies use different methods to read the DNA template. We therefore describe different technologies based on their underlying method used for sequence detection.

Pyrosequencing. In 1986 Nyrén [30] developed a method introducing bioluminescence to measure nucleotide incorporation during DNA synthesis. Nearly 20 years later Margulies et al. [31] published the first massively parallel approach using bioluminescence for DNA sequencing.

The approach is based on a complex enzymatic reaction resulting in light emission during strand elongation. With intermediate steps *adenosine triphospate* (ATP) is produced which is then used by the firefly enzyme luciferase to emit a light signal which can be detected. Alternating the base nucleotides each round, capturing the light signal and removing left over dNTPs one is able to determine the sequence of the template sequence. Pyrosequencing uses emulsion-PCR for amplification and immobilization of template DNA.

454 Life Sciences developed the commercial applications and distributes the Genome Sequencers (GS) Junior and FLX(+) using Pyrosequencing. Their newest system claims to have read lengths up to 1,000 bps with a total of approximately 700 mega-bases per run. The company was founded by Rothberg and is since 2007 part of Roche.

Reversible dye termination. The initial method is based on work by Turcatti et al. [32, 33] and was later adopted by the company Solexa in 2001 which entered the market in 2006 as the *Genome Analyzer*. Illumina bought Solaxa in the beginning of 2007. The company now offers various different sequencers called HiSeq, MiSeq or Genome Analyzer. The latest models reach read lengths up to 2x100 base pairs and a throughput of 600 giga-bases per run.

The reversible dye termination technology uses dye-labeled and further modified nucleotides that pause strand elongation after incorporation of one base. The incorporated nucleotide is detected by a optical device based on its color. After this, the dye and termination end of the nucleotide is removed leaving an unblocked 3′-end allowing the DNA polymerase to continue and the next cycle begins.

The underlying protocol uses bridge amplification for preparation of the template DNA.

Sequencing by ligation. Both previously presented techniques implement a sequence by synthesis (SBS) approach to read the nucleotide composition of the DNA strand. The next method presented uses a sequencing by ligation technique which uses DNA ligases instead of DNA polymerases to achieve this goal.

The commercial application *Sequencing by Oligonucleotide Ligation and Detection system* (SOLID) is based on the work of Shendure et al. published 2005 [34] and entered the market 2007.

The method uses special octamers acting as probes with two known nucleotides and six degenerated or universal nucleotides. After only perfectly hybridized probes are joined by the ligase the incorporated two nucleotides are detected by the dye the probe is label with. The SOLID system uses a degenerated color code with only four different dyes encoding for four different octamers. After dye detection the last three bases are removed and a new octamer is hybridized. This results in every fifth base to be read. Completing a specific number of cycles the whole primer and hybridization product gets removed and a new round starts with the starting position being shifted one base upstream.

SOLID is now distributed by Life Technologies after acquiring Applied Biosystems. The current models 5500 W and 5500xl W offer a read length of 75bp with a maximum of 320 giga-bases of throughput per run. Early ABI SOLID sequencers use emulsion-PCR similar to pyrosequencing. In 2012 Life Technologies announced the introduction of the "Wildfire" library preparation for their newest models. Wildfire is similar to bridge amplification used by Illumina's sequencers and claims to achieve a reduction in library preparation time and costs.

Ion Sensitive detection. The so called *Ion torrent* technology is base on detecting small voltage changes as results from the incorporation of nucleotides [35, 36]. The template strands are amplified using emulsion-PCR and put into micro-wells on a semiconductor chip. Each well includes a sensor which can measure free protons. The wells are flooded with a predetermined sequence of nucleotides so the sequencer knows the sequence of the incorporated base. Ion torrent is distributed by *Life Technologies* and claims that with the release of the new Ion Proton II chip a sample-to variant analysis with 20x fold coverage of the human genome will be possible within a single day.

Single-molecule sequencing

Besides the already presented massively parallel sequencing technologies there are ambitions to sequence a single DNA molecule in one read. Companies offering single molecule sequencing approaches are *Helicos Biosciences*, *Pacific Biosciences*, and *Oxford Nanopore Technologies*. Although their different approaches are highly interesting and promise to yield good results in the future single molecule sequencing has not managed to compete with the massively parallel approaches yet. Up to now single molecule sequencing has no broad audience in genomics and metagenomics and therefore we will not discuss these technologies further.

2.1.3 Sources of Error and Challenges

The previous sections illustrated different technologies for sequencing DNA molecules. Though they differ in their approach some technologies share a potential methodical bias.

The complex system of modern sequencers, the required library preparation, and constantly updated chemistry makes it difficult to asses the introduced biases completely. It has been shown that certain protocols have problems with various genetic features such as high GC-content, high AT-content or regions interfering with primer ligation [37, 38]. A general observation is that sequencing quality drops towards the maximum read length the current technology is offering. Specific errors of sequencing chemistry and their underlying causes are not scope of this thesis. We will therefore only briefly discuss typical errors for the different sequencing technologies.

PCR artifacts or other amplification errors play a role for the massively parallel sequencing approaches because the rely on a high copy number of the template DNA strand to get clear signals. GC-content or primer selection may interfere with the correct amplification or cluster generation prior to sequencing [37]. This may result in an over- or underrepresentation of specific genomic regions of the sequenced genome [39, 40].

Indels The pyrosequencing and Ion-torrent technologies have problems when sequencing homopolymer-runs. This is due to the fact that with the increasing incorporation rate of the same nucleotide at the same time it gets more difficult to calculate the exact number of nucleotides incorporated. The visual or electrical signal should behave in a predetermined way, but errors may occur if signals do not behave linear. Especially tools using translated nucleotide sequences for analysis may have problem with this type of error, because of the resulting frame shift during DNA to protein translation [41].

Substitutions Dye termination and ligation sequencing do not suffer from problems of homopolymers, because only one nucleotide is incorporated at the same time. Cluster density and other factors may play a role that the color determined, representing a specific base, may be wrong. This leads to substitutions in the DNA sequence [42]. Detecting the difference between *single nucleotide polymorphisms* (SNPs) and sequencing errors is one of the challenges in the analysis of NGS data. SOLID has the advantage of sequencing each base multiple times, so sequencing errors may be detected more easily.

Besides the generalized error types the vendors offer various different protocols which differ in multiple ways: the required minimal DNA amount needed, single-end or paired-end capabilities, read length and insert size. Additionally not only DNA sequencing technologies advanced. Many of the companies offer RNA sequencing solutions based on their technologies. It is also possible not only to target the whole genome with an approach, but to use a targeted procedure to limit the view onto special regions of interest. Typical regions are complete exomes or specific ribosomal RNA sequences, but also user specific regions could be targeted.

Taking the current developments into account it can be said that sequencing technologies overtook the whole process of sequence analysis and are now far more advanced than the up- and down-stream preparation and analysis. Various types of technologies produce similar, but slightly different types of data which need to be processed and analyzed. Also the amount of data changed, challenging data storage, access and analysis tools to its final.

The presented technologies all have their own advantages and drawbacks and may be suitable for specific task. The toolset available for specific types of data also may vary and is potentially better for different questions in mind.

With this in mind the sequencing technology to use in a project has not only to be determined by throughput and error rate but also by suitability for the specific task. Nevertheless most regarded values for sequencer selection are still read length, throughput and costs. The two most commonly used technologies are until now 454 and Illimina sequencing. The two major differences are that 454 is offering longer read length at a higher costs and Illumina a higher throughput costing a fraction per sequences base. Especially Illimina's cost-throughput ratio enabled it be favored in metagenomic projects around the globe.

Chapter 3

Genomics and Metagenomics

The previous chapter introduced current technologies to get sequence information from organisms. In this chapter we will present current approaches and challenges in the field of genomics and metagenomics.

3.1 Approaches in Genomics

One idea of genomics is the understanding of genomic features (e.g. genes). To understand which genes play which role in the organism it is crucial to have a potentially complete representation of the DNA content of the organism. Ideally the complete DNA sequence is resolved. This hold only true for selected model organisms (e.g. human, mouse) where a reference genome exists in a *finished* state. Besides that, a number of genomes exists in various different resolutions and states: draft genomes which maybe miss only repetitive regions, genomes only existing of *contigs* (i.e. longer resolved DNA fragments), or even raw sequencing data.

Sequencing data from today's sequencers comes as small chunks of DNA so called reads of DNA. The length of those chunks range from 50 to 700 base-pairs depending on the method used. Two fundamental steps of the initial analysis in genomics are the alignment or assembly of the reads.The alignment against a known reference sequence ensures that the experiment worked and acts as a base for the following analysis of the sample. If no public reference or at least contigs for the organism in question exist, the first step is creating an assembly using the available data.

In the early days of genomics when Sanger sequencing was the only technique available, an assembly of the reads was done by looking for longer overlaps of the reads and than rebuilding the original DNA sequence from these overlaps. Because of the shorter read length of the modern sequencers the basic overlap approach is not feasible anymore. The basic idea stays the same, but modern methods look for shared subsequences (k-mers) and build a graph which then gets resolved during the assembly process. The first graph for an assembly was a *de Bruijn graph*. Popular assembler for short reads are VELVET [43], EULER-SR [44], SOAPdenovo [45] and ABySS

[46]. Most of the modern assemblers make use of previously described mate-pairs libraries to resolve complicated regions.

Modern alignment programs use one or more reference sequences to place to reads upon. Depending on their capabilities some tools are even referred as *Reference Guided Assemblers*. They are designed to handle the vast amount of sequencing data, and some of them are especially tuned for a special sequencing technology. Those optimizations include file format, read lengths and technology specific error models. Most modern algorithms use either a hash based approach (indexing reads and/or reference) or use Burrows Wheeler transformation (BWT) for indexing of the reference sequences.

In general the hash based approach hashes multiple subsequences (SEEDs) of the input and than compares this against the reference sequences, often referred to as *database*. Allowing some hashes to be unmatched the algorithm is able to find even non perfect matches. Depending on the subsequence length, hash collisions (i.e. the same hash is representing two different sequences) are unlikely to happen, so this approach is extremely fast compared to earlier methods. Tools implementing a hash based approach are MAQ [47], SOAP(2) [48, 49], BFAST [50].

Most hash based methods have been redeemed by BWT based methods. Some of the most popular aligners for short reads using BWT are BOWTIE [51], BOWTIE 2 [52], and BWA [53]. The advantages of the BWT over the classical hash are rapid search capabilities and compression of the index. Burrows-Wheeler is simplified based on a suffix array of shifted positions of the reference sequence. The compression during index creation has the additional advantage of loading the complete reference into memory speeding up the analysis further.

After creating longer stretches of sequences or placing the reads onto a known reference genomics offers a multitude of question and procedures to investigate. One of the main topics is the detection of genomic variance and its induced changes to the phenotype of the organism in question. *Single nucleotide polymorphisms* (SNPs) are the most common genetic source of variance and can be relatively easily detected when comparing multiple sequences.

SNP detection will be later revisited in Chapter 8.

3.2 Challenges for Metagenomics

The complexity of metagenomic samples introduces new challenges which are different than the original genomic approach. A classic way to give an overview is to categorize computational challenges into three basic questions.

The three metagenomic questions

The first computational question in a metagenomic study tries to identify the composition of the sampled community, e.g. figuring out the taxonomic content of a dataset. This is usually achieved by using either database-based methods or composition based approaches. Both ideas - comparing either features of sequences or sequences itself - have various advantages as well as disadvantages which will be discussed later on.

The second classical question is concerned about the functional content of a metagenomic sample. Reconstructing genes and interaction between organisms is a complex task. One way of inspecting the functional content of a metagenomic sample is also by some kind of sequence comparison. This time the sequences need to be translated and compared against known protein databases. Knowledge about the function of the reference proteins can then be transferred to the dataset. It may also be possible to successfully apply gene prediction algorithms from genomics - depending on available fragment length.

The last question deals with the comparison of multiple metagenomic datasets. Examples are experiments where different datasets come from either the same location and different time-points (e.g. prior and past a specific event) or just from different locations. One approach is to analyze samples as previously described and use this information to compare the samples with each other. The second approach compares the sequencing information of the samples directly. This has the advantage that similarities not represented in the database can be detected.

We have seen that a central point in the analysis therefore is the assignment of reads to taxa or functions. This process is commonly referred to as *binning*.

Binning of metagenomic reads

Solving questions in a metagenomic study usually depends on functional and taxonomic assignment of reads. Different approaches have been developed to tackle this specific problem. Comparing the whole metagenomic dataset to all sequences in a public database such as the NCBI-NR or NT [54] database using BLASTX or BLASTN [55] is the most thorough approach, but is computationally very demanding.

As such reference databases continue to grow, and as the datasets to be analyzed continue to grow as well, this type of analysis is becoming increasingly challenging.

In the context of mapping and resequencing, numerous new algorithms and tools have been developed for ultra fast mapping of short reads to a reference genome using for example Burrows-Wheeler transformation or Bloom filters, for example BFAST [50], BLAT [56] or Bowtie [51]. Unfortunately, such tools are not directly applicable to metagenomic data, often failing to map more than one percent of all reads of a typical metagenome dataset, as they require near identity between sequences.

A different approach is to try to predict the affiliation of the single reads to specific species, using machine learning techniques such as HMMs or SVMs. The idea is based on the assumption that related species have correlating GC-content, k-mer frequencies etc. Tools in this categories are for example Phymm, PhymmBL [57], Treephyler [12], Naive Bayesian Classifier (NBC) [11] and PhyloPythia [58].

While such methods require an initial training effort, potentially they are much faster than a brute-force BLAST analysis of all reads. Their main drawback is that they do not provide alignments for the reads, which is a limitation because biologists often resort to inspecting such alignments to decide whether a match is significant. The longer the fragments are the better the tools usually perform. Studies show that large contigs such as >8kb outperform BLAST regarding accuracy and runtime, but usually metagenomic fragments are much shorter. Moreover, an alignment to known functions is often required to perform a functional analysis of a dataset.

Recent methods such as Rapsearch(2) [10, 9] and Pauda [59] combine modern mapping approaches with the use of reduced alphabets. In principal those tool translate the amino acid sequence to a reduced form, allowing faster search patterns to be employed. Afterwards nucleotide or protein alignments are calculated allowing a functional analysis as well. Those tools are able to achieve high speed analysis while keeping sensitivity and specificity reasonably high.

Classification: Taxonomy and Ontologies

Information about sequence content, including sequence origin and protein function, needs to be presented and organized in a feasible manner. Species are often grouped in a taxonomy whereas functions are represented in some form of ontology. Especially for functional classification various schemes exist, but also for taxonomy multiple classifications are available. We will briefly introduce three common classification systems: one taxonomy and two systems for functional content.

NCBI-Taxonomy: One scheme for the classification of organisms is the NCBI taxonomy [60] where different species are grouped and visualized as a rooted tree. Intermediate nodes represent various common categories such as Family, Genus, Order or Kingdom. The NCBI taxonomy currently consists of approximate 1 million entries and is based on all sequences in the public NCBI database. It currently represents about 10% of all known species on the planet. A more detailed look into the taxonomy will be given in Chapter 6.

SEED: The SEED classification [61] maps genes (features) onto functional roles which appear in different subsystems. Subsystems are logical units combining features with similar high-level association (e.g. RNA Metabolism). Right now there are about 13,000 functional roles represented in the system.

KEGG: The Kyoto Encyclopedia of Genes and Genomes (KEGG) [62] maps genes onto KEGG orthology (KO) groups. KO groups are linked to enzymes which are connected to different pathways. The latest KEGG release includes information about approximately 10 million genes from roughly 2,500 different organisms. Genes are organized in roughly 15,000 KOs.

The MEtaGenomeAnalyzer: MEGAN

Classification of reads often yields multiple, sometimes scored, hits for each single read. As one can not assume the best match to be the origin of a metagenomic read various tools exist to postprocess the output. Here we shortly present the MEtaGenomeAnalyzer (MEGAN) [13] as we will use the software throughout this thesis.

The software was originally published in 2007 [63] and has been constantly updated and improved.

For taxonomic classification MEGAN uses a lowest common ancestor (LCA) approach assigning the read to the lowest common ancestor of all matched species in the taxonomy. For the calculation only valid hits above a specific threshold are taken into account. The LCA algorithm is a conservative approach minimizing false positives by decreasing specificity if multiple organisms are matched.

For functional analysis MEGAN uses the SEED and KEGG classification schemes which are represented in the software as trees. To perform the analysis MEGAN identifies the best match to a reference sequence with a functional role or KO group respectively. Technically the functional assignment is based on mapping files where RefSeq accession numbers are linked to SEED functional roles and KEGG KOs respectively. Taxonomic mapping is achieved using text based analysis of reference sequence identifiers.

After the import and mapping process the software allows users to visually inspect and analyze metagenomic datasets.

Computational challenges

Besides the specific challenge of assigning reads to species or functions, metagenomic samples introduce other more generic computational challenges.

Challenge 1: Data Complexity The composition of an environment sample usually consists of a mix of known, only partially known and completely unknown species. The complexity of the samples therefore introduces new challenges during the analysis.

Most tools developed for genomic use cases are only partially applicable in metagenomic projects. The bigger part of tools was designed to handle either resequencing tasks with known reference sequences or the creation of new reference sequences using DNA reads originating from a single organism. Gene prediction and most other algorithms either require or profit from long read fragments to work properly.

Overall the assembly of metagenomes is still experimental and actively discussed in literature. The main concern is that most assembly algorithms are originally designed with the assumption that all reads originate from the same species. Using a de novo assembler without evaluating it on metagenomic data first is therefore not recommended. Problems occur mostly due to cross species repeats, varying coverage depth and required sequence depth [64]. Nevertheless some designated metagenome assembly toolkits [65, 66, 67] already exist.

Varying abundances of species in the sample and database also limit the analyses, as quantitative as well as qualitative conclusions can not be easily done. The binning process can be heavily biased by over- or under-representation of specific genes or taxa. Completely novel or missing sequences in the database additionally may result in an incomplete taxonomic or functional description of the sample.

Challenge 2: Data Volume & Accuracy The mix of DNA originating from various different species in varying abundances requires the use of modern sequencing technologies to capture a reasonable amount of DNA representing the whole community. Those systems already have a throughput of 100 GB within 24 hours, enabling a detailed view into a metagenomic sample. On the other hand this huge amount of data emphasises the need for state of the art analysis methods, when using the whole run for a single sample.

The rapid generation of new data has an additional effect. With the number of known sequences rapidly increasing the data to compare a sample to increases as well, further slowing down succeeding analyses.

Up until 2008 this trend was mostly compensated by the parallel evolvement of computational power. Sequencing throughput basically developed parallel to Moore's Law, which oversimplified describes the trend that reasonable cheap computational power doubles every 16 to 24 months. Since 2009 the development of sequencing technologies has clearly outperformed Moore's Law, which makes it harder and more expensive to analyze all data just by increasing computational power [68]. Also this computation cost is often neglected in the overall costs and time estimation [69]. Frankly said the development has shifted the bottleneck in the study from data generation to data analysis.

Though data generated by modern technologies is relatively good it is not insusceptible to errors. Quality control steps have to be taken to remove potential errors before submitting data and making it publicly available.

Challenge 3: Data Access & Sustainability Increasing data volumes also require novel ideas for data storage and sharing. As with all modern studies the topic has to be investigated by many specialist of different fields. The data generated has to be accessed, stored and ideally archived. Depending on the analysis done the result may be more complex and in a different format than for a genomic project. For metagenomes different public repositories are available often combined with analysis tools [15, 70], but various sites exist providing project specific metagenomes in varying formats to download. Major disadvantages of centralized systems are for example the dependancy on third-parties, data security, transfer volumes and inflexibility of analysis approaches. Most public sites are therefore only suitable for post-publication storage, if at all. Private data sharing can be easily achieved on various ways including offline (physical) data transfers or accessible data staging areas (e.g. ftp), but an own integrated analysis solution may also be an interesting option as they offer multiple advantages.

Regarding sustainability of public web services experience shows a high potential of the service being orphaned or just vanishing after a certain period [71]. The main advantage of private data sharing areas is the full control over the data and of the potential shutdown. As usually the main discussion will come back to funding eventually, so no general solution can be provided.

Additional Challenges: Metadata & Comparative Metagenomics The analysis results need to be stored with as much information about analysis parameters as possible, because a single change can interfere with the reproducibility of the result.

In 2006 Foerstner et al. [72] already pointed out that comparing different metagenomic dataset one "may end up comparing apples and oranges". The keynote of this is that data itself relies on the exact sampling protocol used. A comparison of different samples can only be validated if all interfering factors may be accounted for. Those may include size filters, sampling site, depth, technology, binning and gene calling protocols. The storage of this so called *metadata* is crucial for making reasonable assumptions when comparing two or more datasets. Different formats have been suggested but a common format was yet not found and is therefore not comprehensively put into action.

Comparing multiple metagenomic samples is difficult, because many variables need to be accounted for. Different taxa and functions, varying sample sizes, reference and genome sizes need to be taken into consideration. Already presented metagenomic storage sites include comparison capabilities [15, 70], but other approaches also exist offering a more flexible way to compare multiple samples [13, 73].

Part II

Addressing computational challenges in metagenomics

The previous part introduced modern sequencing technologies, gave an introduction into genomics and metagenomics and described typical computational challenges in metagenomics. This part introduces methods of quality control and than moves on to two metagenomic specific challenges. We first discuss methods for determining the content of a metagenomic sample focusing on the comparison of sequencing reads and reference databases. We then continue by studying current methods for assigning reference sequences to specific species or functions. The last chapter focuses on the computation, the access and sharing of high volume data using a workflow management system.

Chapter 4

Prelude: Quality Control

Before starting with any kind of analysis the most important step is to begin with some sort of quality control (QC). The initial QC step insures that the data used is in usable condition, and if not, it modifies the data to convert it into an usable state. Unfortunately this step is often neglected or only poorly executed. This result is a not optimal or even not utilizable analysis, wasting computational and human resources. In this chapter we will first discuss the types of errors and their detection and later potential strategies how to remove error from the data.

4.1 Detection of Errors and Solutions

The first source of errors, sample extraction, library preparation and sequencing is usually beyond control of bioinformaticians. It has been shown that even processing the same sample at different facilities with the same protocols yields different results. This introduces huge effects in the comparison of multiple samples.

Sequencing Errors

In Section 2.1.3 we already introduced sequencing technology specific errors. Each cycle can produce base calls of non optimal quality. Detecting *bad* bases is therefore a requirement to successfully handle this type of errors. The first successful tool to estimate a base specific quality was the tool Phred. The idea was fast adopted by most vendors and laboratories. This so called Phred quality score (QS) is defined as

$$Q = -10 log_{10} P$$

with P being the probability of an incorrect call.

Broadly speaking the Phred quality score is a measure of the correctness of the corresponding base. Figure 4.1 displays the common range of Phred scores.

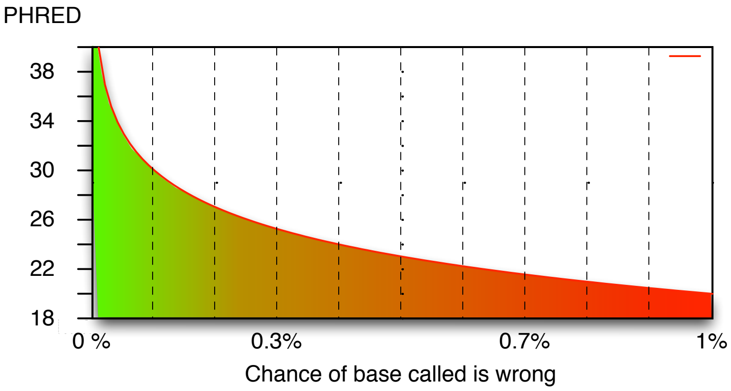


Fig. 4.1: Figure of the relation Phred vs. probability of a wrong base call. As an example a score of 30 represents an accuracy of 99.9% i.e. 1 in 1000 bases is wrong.

QS in FASTQ files are often encoded as one ASCII character. Early ASCII encoding varies between vendors as well as early Solexa models use odds instead of probabilities, so some attention is required when using older file formats.

Depending on the manufacturer, machine and sequencing mode companies claim and assure different minimal qualities. Illumina for example asserts that 80% of the HiSeq 2500 data has a quality score of 30 or above.

It has been shown that the score gives a reliable indicator of sequence quality [74, 75]. Quality scores are mostly used in alignment tools and assembly programs. Also some SNP detection algorithms use QS to distinguish between a true SNP or sequencing error [76].

Unfortunately tools such as BLAST do not make use of this information. Malde [77] suggested such an extension of BLAST but has not gained a broad audience. The extension is based on a position based substitution matrix combining substitution events and QS of the sequencers. Other approaches use only a high coverage to remove potential sequencing errors. With metagenomics having usually a low coverage those approaches are not suitable.

Quality along reads usually tend to drop towards the end. Trimming read ends speeds up the analysis as overall data gets reduced. Different strategies exist including cutting at fixed positions, after a bad base or using a sliding window approach. Using a sliding window is the most feasible way as this ensures that single base errors in the middle of the read do not truncate reads unnecessarily. Windowsize and quality threshold depend on the overall quality of the data and if downstream analysis is capable of using QS.

Besides wrong base calls sequencers may introduce PCR artifacts or artificial sequences from barcodes, linker, primers and adapter. These parts need to be identified and removed from the data as well. For protocol based sequences (linker, primers and adapters) the best way of doing this is comparing it against a list of potential sequences. Filter algorithms used should to be able to detect sequences with sequence errors as well as reverse complements.

Tools such as FastQC [78] are capable of generating QC reports. A graphical report is generated of various quality measures. Widely used tools for sequence manipulation are fastx_toolkit [79] and ea-utils [80]. The tool PrinSEQ [81] provides an integrated approach generating a graphical representation followed by a guided approach how to process the input data.

Merging of reads

When using a paired-end protocol with very short fragments it can happen that the same paired-end fragment gets sequences from both ends. This effect can be used to compute a consensus read eliminating potential sequencing errors. For this the reads have to be joined from both ends and an overlap consensus needs to be calculated. Tools completing this tasks ea-utils [80], FLASH [82] and SHERA [83]. Algorithms, scoring and implementation differ, ea-utils for example use a squared distance for the alignment and is written in C. Also handling of unmerged reads may differ. It is best to keep unmerged reads for further use.

Low complexity and Sequence Duplication

Sequenced DNA can contain so called *low complexity* regions which are defined as "intervals with highly biased distributions of nucleotides" [84]. The problem with those regions is that besides their zero informational value they use resources during analysis and also could produce significant hits with no biological meaning [84, 85]. Such regions should therefore be removed. The most popular algorithms for this are the DUST and SEG. The algorithms are implemented in BLAST and Segmasker respectively, but tools using simpler algorithms, e.g. using a compression index [13], also exist. Depending on the tool used, the result may be only a modified read file using either lower case letters or Ns. Depending on downstream tools those regions or reads need to be removed prior to further processing.

Artificial sequences, such as poly-A/T tails, may also be introduced by sequencing when fragments are too short for example. Those regions may align well with low complexity sequences in sequence databases and therefore create false positives [81]. Low complexity filters and special poly-A/T trimmer can mark and/or remove those sequence parts.

Sequence duplication can occur through either sequencing protocols or amplification steps after DNA extraction and may affect the downstream analysis . This is especially the case if low amounts of DNA is used. If a high number of sequence duplication is suspected it is recommended to scan the data for multiple instances of the same read and to remove those reads. Some methods remove duplicates only after mapping and not during QC (for example samtools [86]). Failure to do so may result in computational overhead and bias during coverage and abundance estimation with resulting bias in the following analysis steps such as variant calling.

FastX-toolkit [79], GATK [87] and PrinSEQ [81] are capable of removing duplicates as well as detect low complexity reads during QC.

Contamination

Contaminating may play a role in some projects, although it is not considered a technical error. Especially host-based projects, where the sampling location is related to another organism (i.e. host), face huge amount of host DNA in the initial sample. It is crucial that contamination is already confined during the wet lab phase. Failure to do so results in the host DNA competing for amplification and sequencing chemistry displacing DNA of interest. The sequenced result will yield a high number read originating from the host and only a fraction of DNA of interest.

In a general (not host based) metagenomic project contamination checks should also be performed as studies indicate that contamination is not as rare as expected [81].

After sequencing, host based DNA can be relatively easily removed by mapping all reads against the hopefully known host genome. This can be done with short read mappers such as Bowtie [51] and BWA [53]. Earlier approaches use BLAST against the human genome [88], which we disencourage from using because of its long runtime. Another approach by Willner et al. [89] use dinucleotide relative abundance to estimate metagenome contamination on a per sample base. This has the disadvantage of not being able to remove contamination reads from the sample.

Schmieder and Edwards [81] point out, that contamination removal is only as good as the used database with the problem that with every resequencing of the human genome novel sequences can be found [90]. They are also concerned about the an introduced bias by BWA during contamination removal. BWA replaces an unknown base N with a random base, which could lead to artificial hits if the reference genome contains longer stretches of N. Introducing the tool DECONSEQ [91] they provide a complete contamination filtering and removal suite, which uses different modified databases for sequence comparison, working reasonable fast.

Overall removing host based DNA and other contamination additionally helps downstream analysis by diminishing data to be processed. We suggest that contamination checks should be performed as last step of the QC.

After the data has been cleaned one can start on the analysis of the sample. The following chapter will discuss the needed steps in detail.

Chapter 5

Metagenomics: Determining the Content of a Sample

The ability of second generation sequencing technologies to produce huge amounts of DNA at a low cost has shifted the bottleneck in metagenomic projects from data generation to data analysis (see Chapter 2). The comparison of environmental DNA reads to known reference sequences using BLAST is now the most expensive step. While taxonomic predictors based on machine learning techniques promise to speed-up the analysis of datasets by avoiding time consuming sequence alignments, alignments are still considered an important part of an analysis (see Section 3.2). The following section introduces a combined approach enabling a faster analysis while still generating full sequence alignments in the last step. We show that the approach offers an up to 10-fold speedup in comparison to a full BLASTX [55] comparison against NCBI-NR [54], while improving the assignment accuracy. The result of the reduced BLAST comparison can be used to perform a functional analysis of the dataset.

5.1 Hybrid Method

We propose to use a hybrid approach to analyze the taxonomic content of a metagenomic dataset. This approach first uses a taxonomic classifiers such as NBC [11] or Phymm [57] to bin reads by taxonomic assignment, and then blasts the reads only to those reference sequences that correspond to the assigned taxon. Finally, we apply the lowest common ancestor approach [13] to all significant matches for a read to obtain a final prediction.

5.1.1 Method

In this approach, we first use a fast taxonomic classifier to assign each read to one taxon in the NCBI taxonomy. Note that reads will be placed at different ranks of the NCBI taxonomy, as they come from different genes and thus have different levels of conservation. In this study we investigate the use of NBC [11] and Phymm [57] as taxonomic classifiers.

In the second part of the approach, we compare each read using BLASTX against the part of NCBI-NR that contains sequences from all taxa that lie below the node to which the read was assigned. For example, if a read was assigned to the node *Alphaproteobacteria* by the taxonomic classifier, then the read will be blasted only against those sequences that belong to the class of *Alphaproteobacteria*.

In the third and final part of the approach, the results of all BLASTX comparisons are concatenated and then provided to the program MEGAN [13] as input. MEGAN uses the LCA algorithm to place each read on to the node in the NCBI taxonomy that is the lowest common ancestor node of all species for which the read has a significant BLASTX hit.

To facilitate the second part of the approach, we split the NCBI-NR database into different sub-databases at the ranks of Superkingdom, Phylum, Class and Order of the NCBI taxonomy resulting in 3, 89, 243 and 1175 nested databases respectively. Then, each read is blasted against the smallest enclosing sub-database. For example, a read that is assigned to the Family of *Anaplasmataceae* will be blasted against the database associated with the Order of *Rhizobiales*.

Although NCBI-BLAST has a feature that allows one to limit the search range to a specific range of organisms, in practice this is much slower than blasting against a reduced database. An additional advantage of working with sub-databases is that it simplifies parallelization.

In slightly more technical detail, after running the taxonomic classifier on a given set of reads, each read is placed in an input FastA file for the appropriate sub-database and then each file is blasted against its sub-database. Figure 5.1 illustrates the process with an example of three different classes of bacteria. All scripts are written in Python and are made available to the public as script package as well as included in our local instance of the workflow management system GALAXY [92].

In summary, the hybrid approach proceeds as follows:

1. Use a taxonomic classifier to assign each read of the input dataset to a node in the NCBI taxonomy.
2. For each read, perform a BLASTX comparison against the sub-database corresponding to all species that lie below the node to which the read was assigned in the previous step.
3. For each read, apply the LCA algorithm to all significant BLAST hits found for the read to perform a final placement of the read in the NCBI taxonomy. Optionally, use the BLAST hits to place the read in the SEED functional classification, as described in [93].

5.1.2 Performance

To investigate the potential speed-up and accuracy of the hybrid approach, we performed a simulation study using six different datasets generated by MetaSim [94], each containing 10000 reads. The taxonomic distribution of species in the simulated datasets was adapted from the FAMES [95] profiles for low, medium and high-complexity metagenomic datasets. For each complexity class we simulated one 454 [31] and one Solexa [96] mate-pair run, see Table 5.1. All datasets were compared against a version of the NCBI-NR database using NCBI BLASTX version 2.2.23+ (default parameters).

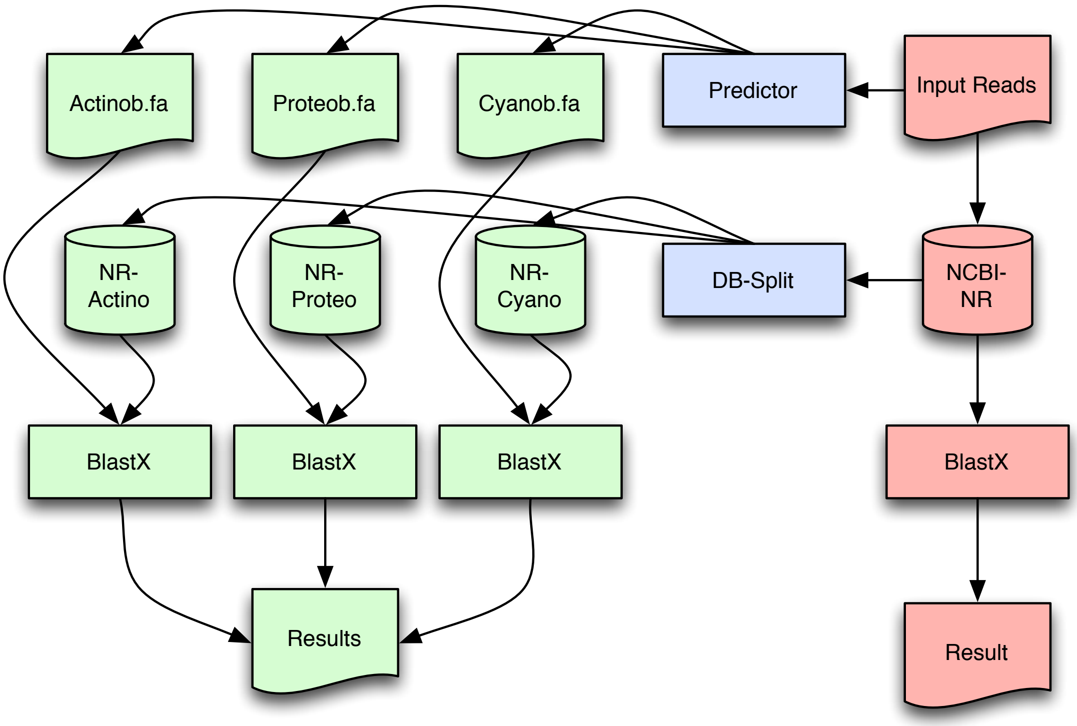


Fig. 5.1: Comparison of the hybrid and brute-force approaches. In the brute-force approach (red), all reads are compared against the whole reference database (in this case, NCBI-NR), using BLASTX. In the hybrid approach, the database is split into smaller sub-databases (green) using a utility called DB-split (blue). The predictor (blue) splits the input reads into sub-input files. Each such file is compared against its corresponding sub-database (green).

We used the Naive Baysian Classifier (NBC) [11] and Phymm(BL) [57] as taxonomic classifiers in the first step of the hybrid approach. We split the BLAST-NR database at the level of Superkingdom, Phylum, Class and Order of the NCBI taxonomy, as mentioned above. To analyze those reads for which the employed taxonomic classifier was not even able to predict a taxon at Superkingdom level, we also kept a copy of the whole NR database.

The output of the taxonomic classifier under consideration was parsed and each read was then placed in a sub-input file according to the predicted taxon. On average about 99% of all reads were placed in a file corresponding to the taxonomic rank of Order, independent of sequencing technology, dataset complexity or prediction method. The sub-input files were then blasted against the corresponding sub-databases and the resulting BLAST files were then merged.

We used the program MEGAN [13] to analyze the BLASTX results obtained from all three approaches (the two hybrid approaches and the brute-force approach), as described above.

Dataset	Read length	Clone length	Standard dev. of clone length	Number of reads
454	250 bp	8,000 bp	800	10,000
Solexa	80 bp	300 bp	30	10,000

Tab. 5.1: Sequencing parameters for simulation study

Accuracy

We analyzed the accuracy of all three processing methods (hybrid approach using NBC, hybrid approach using Phymm and brute-force approach) on all three datasets (low, medium and high complexity), for two different sequencing technologies (454 and Solexa) at the taxonomic ranks of Species, Genus and Order.

For each of these combinations, we assigned every read into one of five bins, depending on whether the read was assigned to (1) a correct or (2) an incorrect node of the considered taxonomic rank, (3) a correct or (4) an incorrect higher node, or (5) if it was not assigned or had no significant hit at all.

Figure 5.2 illustrates the results at the Species rank. For all three datasets (low, medium and high complexity) and both sequencing technologies, the hybrid approach using NBC shows slightly better accuracy than the brute-force BLASTX approach, whereas the hybrid approach using Phymm always performs less well, especially for the Solexa reads.

At the rank of Genus (Figure 5.3), the number of correctly assigned reads is higher than at the Species level. Again, the NBC-based hybrid approach works slightly better than the brute-force approach, while the Phymm-based approach works slightly less well on the 454 data and substantially less well on the Solexa data.

At the rank of Order (Figure 5.4), the NBC-based hybrid approach achieves almost perfect results on the 454 data and better results than all other methods on the Solexa data. While the brute-force approach assigns less reads to the correct Order, the other reads are correctly assigned to nodes at a higher level of the taxonomy. Such assignments are “underpredictions” rather than false positives. In other words, the NBC method is more specific than the brute-force approach (at all three ranks considered).

In summary, in terms of accuracy, the hybrid approach using NBC performs better than the brute-force BLASTX approach, while the hybrid approach using Phymm performs worse, in particular on shorter reads.

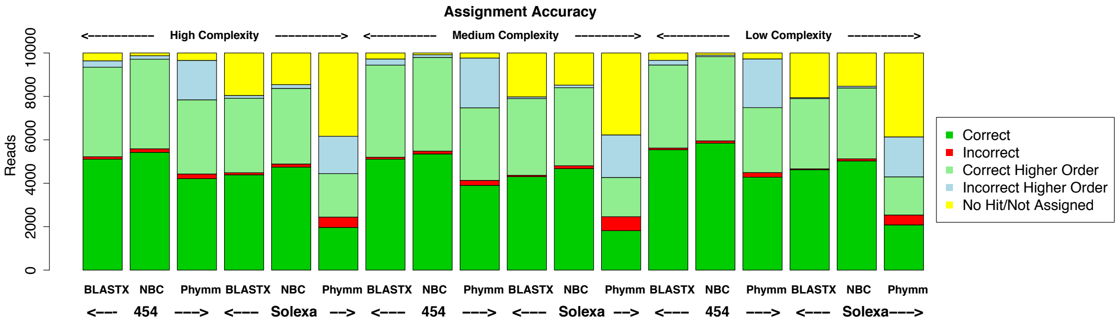


Fig. 5.2: Accuracy at rank of Species. We show the performance of the brute-force BLASTX approach (labeled BLASTX), the hybrid approach using NBC (labeled NBC) and the hybrid approach using Phymm (labeled Phymm), for two different simulated sequencing technologies (454 and Solexa), for three different simulated metagenomes (high complexity, medium complexity and low complexity). For each combination, from bottom to top, we show the number of reads assigned to the correct species (labeled Correct), an incorrect species (labeled Incorrect), to a correct higher taxon (labeled Correct Higher Order), a wrong higher taxon (labeled Incorrect Higher Order) or not assigned (labeled No Hit/Not Assigned).

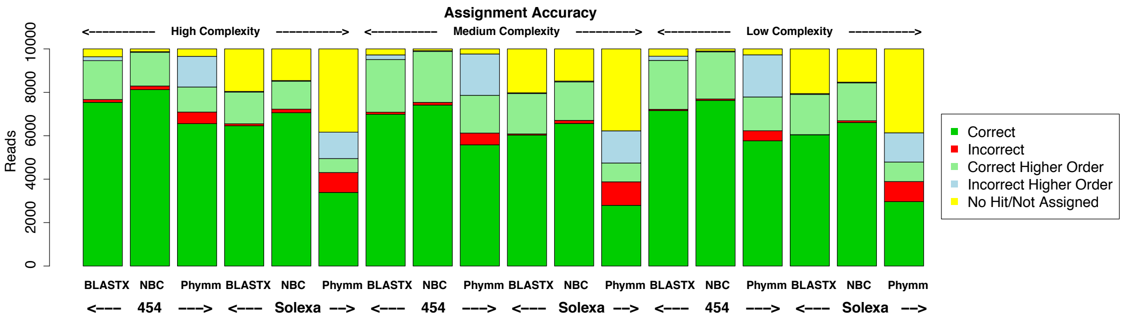


Fig. 5.3: Accuracy at rank of Genus. For details, see previous 5.2.

Processing Times

All programs were run on an AMD Opteron 2 GHz System with 8 GB of memory. The programs were executed as single core applications to keep the results comparable between the different scenarios.

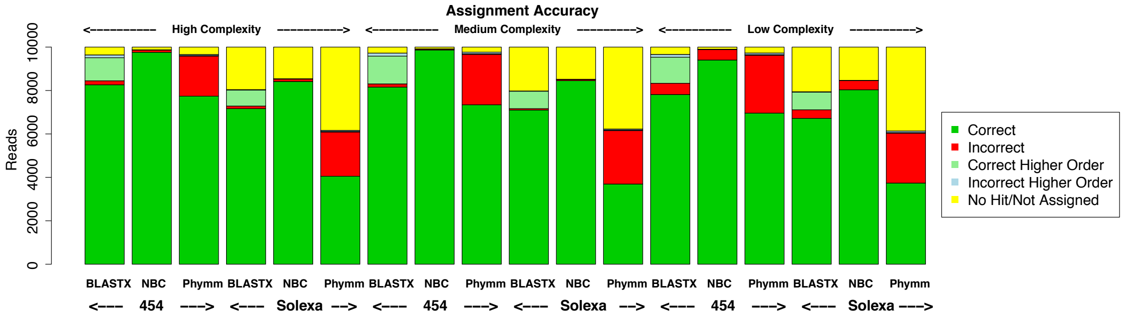


Fig. 5.4: Accuracy at rank of Order. For details, see figure 5.2.

Figure 5.5 displays the overall runtimes for each of the datasets. The two hybrid approaches, using NBC and Phymm, require roughly the same amount of time, and are both significantly faster than the brute-force BLASTX approach. The speed-up for 454 reads is more than ten-fold, whereas for the shorter Solexa reads it is slightly over three-fold. The reason for this discrepancy is that the 454 contain roughly three times as many bases as the Solexa datasets (both have the same number of reads, but the 454 reads are approximately three times as long). This shows that the runtime for the brute-force BLAST approach depends much more on the sequence length than for the hybrid approaches.

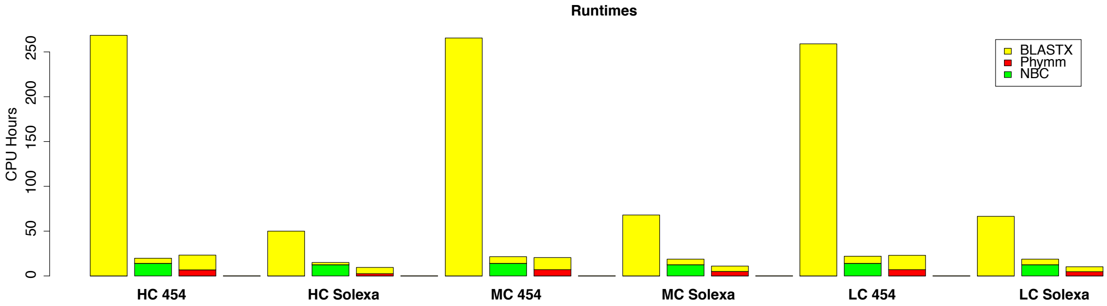


Fig. 5.5: Comparison of runtimes on each of the six datasets high-complexity using 454 (labeled HC 454), high-complexity using Solexa (labeled HC Solexa) and so forth. We plot the run time for the brute-force BLASTX, the hybrid approach using NBC and the hybrid approach using Phymm.

5.1.3 Discussion and conclusions

This simulation study suggests that taxonomic classifiers can indeed be employed to obtain a substantial speed-up of analysis without sacrificing accuracy. In fact, in this study the hybrid approach using the NBC method exhibits a higher accuracy than the brute-force BLASTX approach. While the result of this simulation study is promising, one weakness of any such simulation study is that the effect of “dark DNA”, that is, of DNA coming from species that are not represented in the reference database, is not taken into account.

In Figure 5.6 we show a comparison of the brute-force BLASTX and NBC-based approaches in the medium complexity 454 dataset. Here we see that the performance of the NBC-based approached is supported by the fact that it is not distracted by matches to Eukaryotes, which appears to be a problem for the brute-force approach. Also, we see that many of the reads are assigned to very unspecific nodes in the taxonomy. By the nature of the LCA approach, this can also be explained by distracting matches to eukaryotic model species. In Figure 5.7, we present a high-level comparison of the SEED functional classification of reads, computed as described in [93]. Surprisingly, the NBC-based approach gives rise to a slightly higher number of predictions than the brute-force approach.

The attainable speed-up depends on the read length, the longer the reads, the more significant the speed-up. Because the hybrid approach uses the predicted taxon to choose the sub-database to BLAST against, the employed taxonomic classifier should ideally have high sensitivity rather than high specificity, because, due to the nature of the hybrid approach the BLAST is focused on one part of the taxonomy and thus cannot recover from a erroneous taxonomic assignment by the classifier. This explains the difference in performance between the NBC and Phymm approaches. It remains to be seen whether PhymmBL, the successor of Phymm, can overcome this problem.

The comparatively poor performance of the brute-force BLASTX analysis on the 454 sequences points to a weakness of the LCA algorithm as implemented in MEGAN. The program currently does not take multiple BLASTX matches to the same species into account and this leads to a loss of specificity, which is apparent in Figures 5.2, 5.3 and 5.4. Because the use of a taxonomic classifier restricts the BLASTX analysis to a part of the NCBI taxonomy, this problem does not occur in the hybrid approaches.

An additional speedup of the hybrid approaches could be obtained by additionally splitting the reference database at a lower taxonomic rank such as Family. However, this would lead to an increase of misclassified reads, as the performance of taxonomic classifiers is less reliable at these lower levels.

Out of the box, NBC is trained to classify microbial sequences. Hence, when applied to metagenome datasets that contain eukaryotes or viruses, say, it is important that one trains the NBC classifier on eukaryotic sequences, or viruses, as well.

Based on this study, for metagenomes consisting of microbial sequences, we recommend using the hybrid approach with NBC to obtain more accurate results in less time. Once the BLASTX comparison of the reads against the relevant sub-databases has been completed, the BLAST matches can be used to compute a functional binning of the reads, as well, for example using SEED [61] as described in [93], thus obtaining both a taxonomic and functional analysis in a much shorter time-frame.

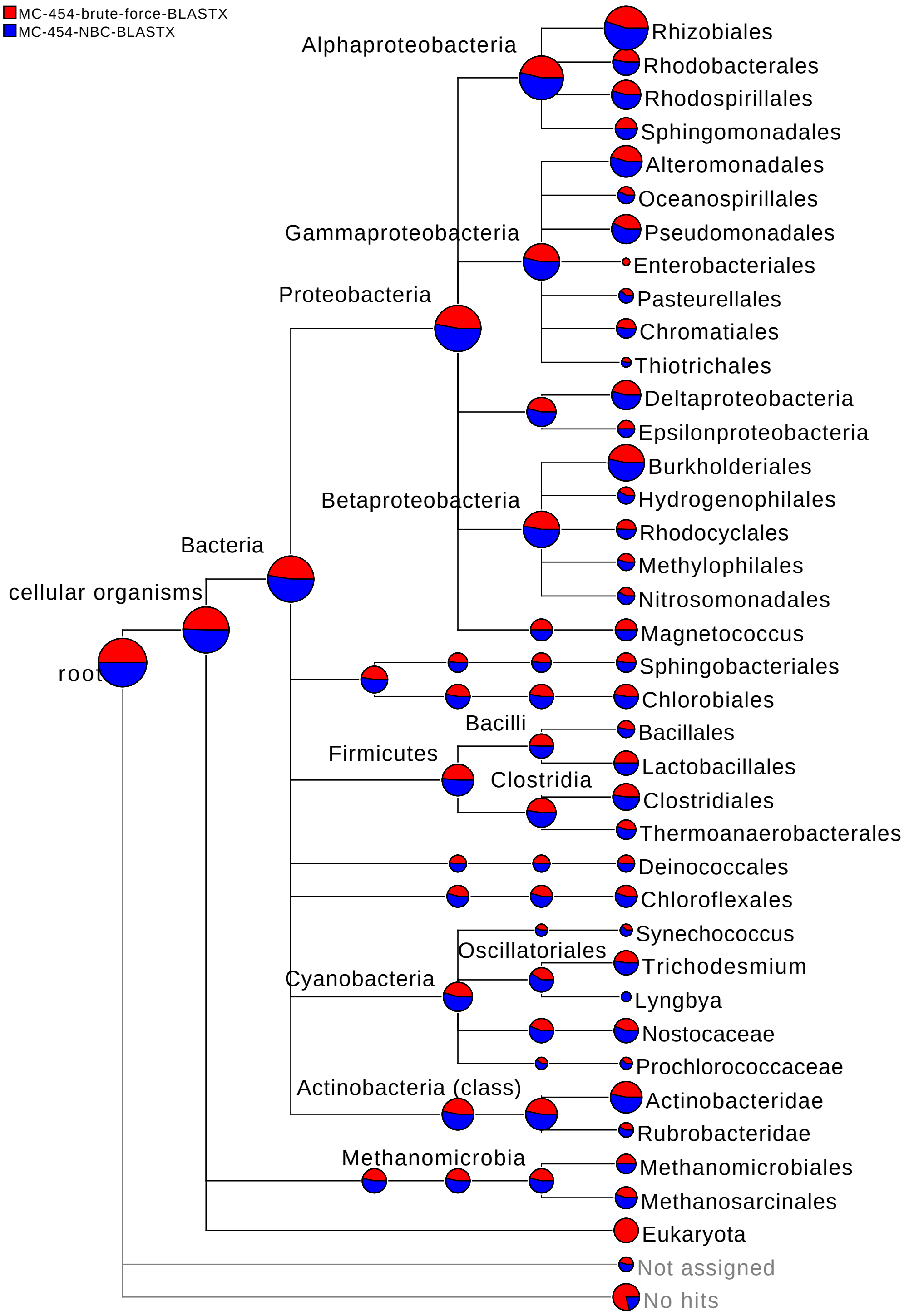


Fig. 5.6: Taxonomic analysis performed by MEGAN on the medium complexity 454 dataset, comparing the results obtained by brute-force BLASTX (red) and the NBC-based hybrid (blue) approach. Nodes are scaled logarithmically by the number of reads assigned to, or below, them.

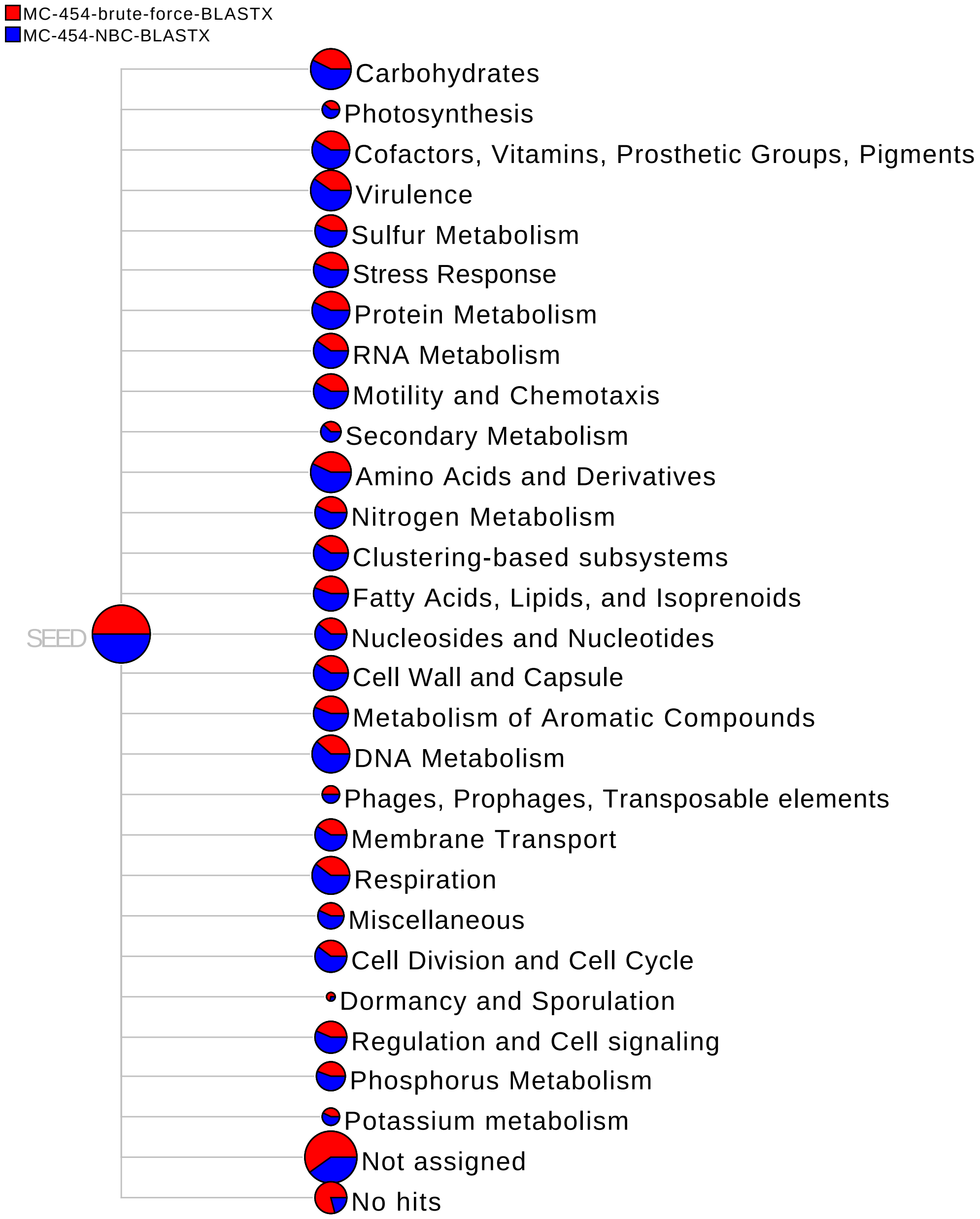


Fig. 5.7: SEED analysis performed using [93] on the medium complexity 454 dataset, comparing the results obtained by brute-force BLASTX and the NBC-based hybrid approach.

Chapter 6

Database and Assignment Accuracy

In the previous chapter we have shown that the result of a metagenomic analysis is highly dependent on the method used for classification of the reads. Generally speaking the main topic of the previous chapter was "*How do we map sequences?*" focusing on the comparison of sequences against a reference database. Here we investigate the question "*What do we map against?*" and "*How do we achieve this?*". This time we focus on the process of going from the sequence to the actual organism or function.

Though not as computational as expensive as sequence comparison this step is equally crucial during downstream analysis. When using database based methods the content of the database itself, as well as the approach for mapping database entries to different species has a high impact on the accuracy of the result.

In this chapter we inspect the coverage of the NCBI-NR database and accuracy of different mapping methods. We will use MEGAN to perform the mapping from database hit to taxon or function. Based on our findings we suggest improvements to currently implemented methods. In the last part of the chapter we evaluate the performance of the new approach with an *in vitro* simulated metagenomic dataset.

6.1 Analysis of the NCBI-NR database

One often neglected factor of a metagenomic analysis is the coverage of the database. The included sequences and their description determine what the analysis (i.e. comparison against the database) can detect. One of the most commonly used databases for metagenomic analyses is the NCBI-NR (non-redundant) database [54] which is updated on a daily basis. The non-redundant definition is hereby that absolutely identical sequences are merged into a single entry, i.e. identical sub-sequences are not removed. The database currently consists of entries from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq databases. Most sequence databases such as the NCBI-NR database offer various alphanumerical or plain text identifiers for each entry, Listing 6.1: Protein sequence with multiple identifiers separated by '>'. Each identifier has a text description as well as different specially coded notations in the beginning. Text descriptions and sequence have been artificially shortened for layout reasons.

```
>gi|332795804|ref|YP_004457304.1| thiosulfate-quinone oxidoreductase [..]
>gi|1729430|emb|CAA69986.1| subunit of the terminal oxidase with [..]
>gi|1742921|emb|CAA70827.1| terminal oxidase subunit [Acidianus ambivalens]
>gi|332693539|gb|AEE93006.1| thiosulphate-quinone oxidoreductase [..]

MSGKQSEEFKRTEKMTRMEYLFPVRFAVGWMFLDGGLRKAVLKPAKLDPNSASFVG[..]
```

explaining the origin of the sequence, see Listing 6.1 for example.

As a direct result the taxonomic and functional assignment quality depends on the identifiers used for mapping. Highly curated sequence identifiers like the RefSeq accession number become more rare as databases continue to grow rapidly. Additionally the composition of NCBI-NR implies that only a part of the sequences may have such a RefSeq entry.

To asses the impact of identifier composition and represented species and function we analyze the performance of three different NCBI-NR databases downloaded on 03.11.09, 15.11.12, and 14.01.13. Databases will later be referred to as their year of download.

Identifier distribution

The NCBI-NR database downloaded 2013 contains a total of 22 million sequences with 63 million sequence identifiers. Those sequence sum up to 7.7 billion bases. Nearly 17 million identifiers are hypothetical entries covering 7.5% of all bases. Details can be found in Table C.3. Within the database the distribution of identifiers is widely spread: About one third of the sequences have only one unique identifier and therefore represent a very specific mapping, 57% of sequences contain two identifiers. The remaining sequences have three or more identifiers with a maximum of nearly 17,000 different identifiers for a single sequence. Exact numbers can be found in Table C.4. Figures 6.1 and 6.2 illustrate the results.

In comparison to previous years the number of sequences and bases in the database more than doubled (2013 vs. 2009), the number of hypothetical entries tripled. The identifier count per sequence composition stayed relatively the same during this period.

The success of NGS is the main contributing factor of database growth. The above-average rise of hypothetical entries can be explained by the increasing usage of gene prediction algorithms and the trend of submitting those predictions to public repositories, as methods are expected to work more reliably. Establishing confident numbers of purely hypothetical entries is difficult, because proteins are not consistently named (e.g. hypothetical, hypo., predicted, pred., theoretical etc.). The number is definitely much higher than presented here. A high number of identifiers for a single sequence indicates that the sequence is highly conserved between various taxa. The more identifiers a single sequence has, the less specific the assignment later on will be.

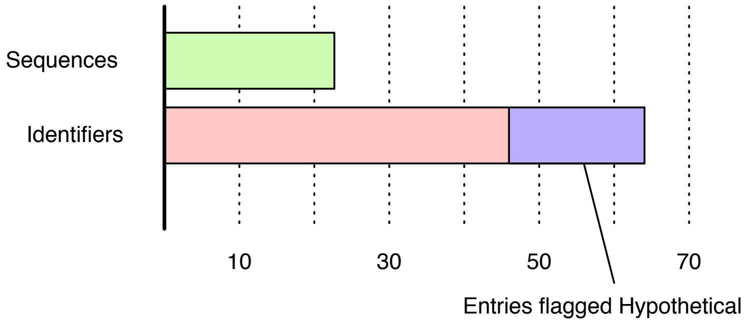


Fig. 6.1: Absolute number of sequences and identifiers in millions for the 2013 NCBI-NR database. Identifiers containing the phrase *hypothetical* were marked as hypothetical entries.

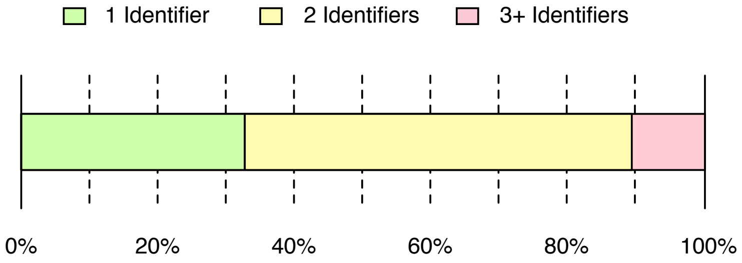


Fig. 6.2: Illustration of identifier distribution for the 2013 NCBI-NR database. *3+ identifiers* describes all sequences with three or more identifiers for a single reference sequence.

Assignment Rate and Taxonomic Coverage

To asses the taxonomic coverage of the databases in combination with an assignment approach we map the sequences in the database onto the taxonomic tree using a specific identifier. The initial method in MEGAN for assigning reads onto the taxonomy was to parse the whole text description of the read to find the maximum match between all taxonomic nodes and the description. This method does not rely on any maintaining of mapping files but is prone to typos and inconsistent naming.

For the representation of different taxa MEGAN uses the NCBI taxonomy. For this analysis we use the snapshot of the taxonomy downloaded on the 30th of January 2013. The NCBI taxonomy website provides an overview of the currently represented nodes [60]. A visual comparison of the taxonomic composition is displayed in Figure 6.3.

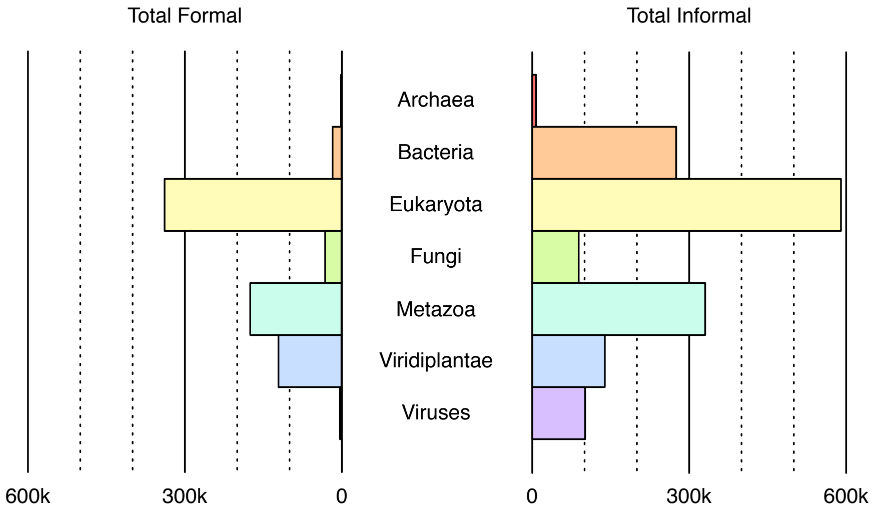


Fig. 6.3: Total number of entries for each subtree in the NCBI taxonomy.

The taxonomy consists of formally and informally named entries. Including informal entries into the taxonomy the number of taxa doubles, but for selected taxa (e.g. *Archaea*) the number increases by a factor of 10. Tables C.5 and C.6 contain exact numbers. General usage of informal names, in the taxonomy or database, bears various problems for computational analysis. Hits for single species may be shared between similar nodes distorting analysis results, as a valid hit originating from a formal species may be contained in multiple informal nodes. It is also possible that reads matching an informally named reference sequences can not be placed correctly, as names differ. Nevertheless if mapping algorithms are using a text based naming approach we recommend to use the taxonomy with informal, uncultured, and unassigned nodes to ensure that the maximum of reads can be (correctly) placed.

All three databases where imported into MEGAN with default options, except *Min-Support=1* was set, to ensure that taxa represented by only a single sequence will also be displayed. As taxonomic coverage indicator we display the exact subtree of the taxon of a specific rank and count the number of leaves below that taxon. We define the ratio of actual leaves to the number of maximum possible leaves as taxonomic coverage for this specific taxon.

In the following we compare the three databases for the big four kingdoms: *Bacteria*, *Eukaryota*, *Archaea* and *Viruses*. Results show an increase in the taxonomic coverage over the years with a focus on Eukaryota and Viruses. With the latest database covering only 12% of the bacteria represented on species level in the taxonomic tree only a small portion can be specifically addresses in a metagenomic approach. This should be taken into account when analyzing complex datasets. It should be noted that sequences of more species have been submitted to the NCBI-NR database, but if sequences are shared between various organisms they will be combined so that a specific classification is not possible anymore. Detailed results can be found in table C.7 and C.8. Figure 6.4 visualizes the results.

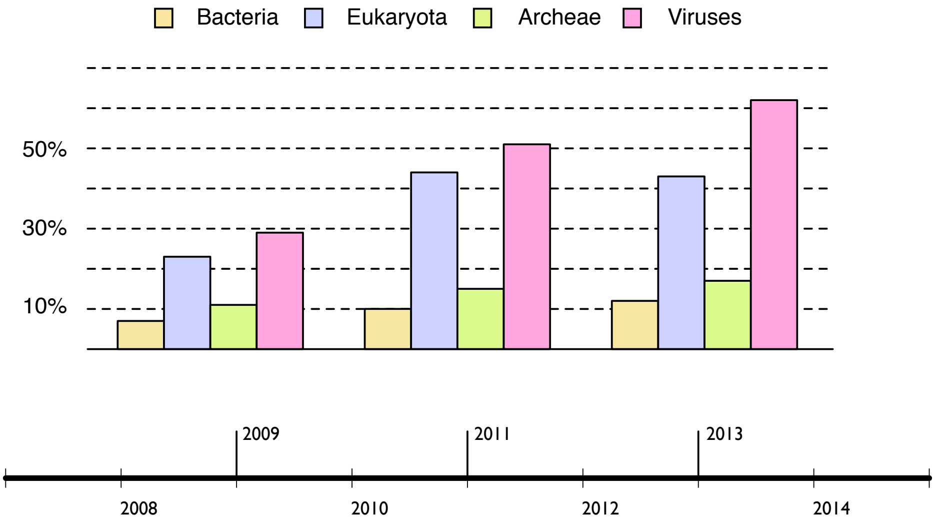


Fig. 6.4: Taxa (leaves) covered by the NCBI-NR databases for the four main subtrees *Bacteria*, *Eukaryota*, *Archaea*, and *Viruses*.

Additionally to taxonomic coverage of a database we define the term *assignment efficiency* as the percentage of sequences which can be placed onto the taxonomy (or function respectively). Table 6.1 shows details about the number of assigned and not assigned sequences for all three databases using the default text parser in MEGAN. Table C.9 lists the total numbers.

Dataset	Assigned	Not Assigned
2013	99.62%	0.38%
2011	99.35%	0.65%
2009	98.37%	1.63%

Tab. 6.1: Table of assigned and not assigned sequences to the taxonomy using the default name parser in MEGAN.

Overall the text parser is performing reasonably well with an assignment efficiency of 99.62% for the latest database. Compared to the previous years, assignment efficiency increases by more than one percent even in the context of the total number of sequences rising. This may indicate that maybe some measures of quality control have been taken and naming is checked accordingly. Taking the high number of sequences into account still approximately 85,000 sequences can't be mapped to any taxonomic node. Most of those unassigned sequences are the result of nonsense and incomplete descriptions in the sequence identifier and IDs containing typos.

Functional mapping using SEED and KEGG

Functional analysis in MEGAN originally used RefSeq accession numbers for mapping reads to the different functions. The drawback is that only a portion of the sequences of the NCBI-NR database have such identifiers. From all IDs of the database 2013 only 44% contain a RefSeq-ID. Additionally a mapping file is required to map a RefSeq-ID to a SEED functional role or KEGG KO.

SEED The currently used mapping file contains 1.3 million entries from RefSeq to SEED identifiers. The SEED mapping is based on a freely available file from http://theseed.org which is now deprecated and was not updated by the maintainers since 2011.

KEGG The latest mapping file contains approximately 2.1 million entries from RefSeq to KEGG KO. The KEGG consortium introduced a payed subscription model for direct access to files starting 01.07.2011. The original mapping file is based on information representing the latest freely available version.

Both systems, SEED and KEGG, are represented as an individual tree in MEGAN. To adjust to the potential incompleteness of the tree we created a dummy tree only containing the nodes *Assigned* and *Unassigned* as intermediate nodes. This was done to minimize the effect of having a valid mapping and missing it because it is not represented in the currently used tree. We then imported the three NCBI-NR databases accordingly using the RefSeq accession number as mapping identifier. Results are shown in Table 6.2. Total numbers can be found in Table C.10.

	SEED		KEGG	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
2013	2.70%	97.30%	6.76%	93.24%
2011	4.31%	95.69%	10.78%	89.22%
2009	6.93%	93.07%	15.32%	84.68%

Tab. 6.2: Percentage of sequences in the NCBI-NR database mapped to SEED and KEGG functions using the default RefSeq parser in MEGAN.

In general only a small percentage of sequences can be mapped to SEED or KEGG functions. With increasing database size the total numbers (C.10) only vary slightly which results in less efficient mapping performance. Mapping only 2.5% in the case of SEED and 6.7% for KEEG is far from the optimum. Outdated mapping files and only partial RefSeq-ID coverage of the database restricts functional analyses in the current implementation in its efficiency. Another restriction for functional analysis is that a high number of sequences have not been assigned to any known SEED or KEGG function. KEGG assignments, for example, only exist for approximately 2,300 different organisms.

6.2 Improved Database Mapping

As shown in the previous section mapping of sequences to functions or species can be further improved. Especially functional assignment is far from optimal. Database annotation, curation and optimization are possible solutions, but are not feasible in our case. We will focus on improving the mapping process.

6.2.1 GenBank Identifier Mapping

Text based and RefSeq ID mapping approaches have both substantial drawbacks. Name consistency errors, missing identifiers and also the low coverage of RefSeq-IDs within the database requires a new way of mapping. We suggest to use *GenBank Identifiers* (GI) for mapping database matches to the taxonomy and functional content. Nearly all reference sequences in the NCBI-NR database contain a GI number in their identifier. GI information needs to mapped to taxa, SEED and KEGG respectively to achieve a result. Mapping file generation and evaluation will be discussed in the following sections.

Taxonomic Mapping

The NCBI already provides a list of GI to taxonomic ID mapping file which is available at `ftp://ftp.ncbi.nih.gov/pub/taxonomy/`. Database, taxonomy and mapping file originating from the same resource (NCBI) is optimal as this should minimize potential inconsistencies between the files. The mapping file is updated once a week providing a consistent source to new database and taxonomy versions. MEGAN was modified to load a GI to taxon-ID mapping file and to use the GI in the sequence identifier for mapping instead of the previously introduced text based approach. The provided mapping file is a modified binary version of the NCBI file, to enable memory efficient access.

The mapping file downloaded on the 30th of January 2013 contains about 73 million entries. After file generation and modification of MEGAN we reanalyzed mapping efficiency as previously described. For the 2013 database the number of not assigned reads dropped from 85,235 to 19,629, leaving only 0.09% of sequences which could not be assigned to any node. Most of the sequences not assigned don't have a valid GI number or descriptive text in their identifier.

Detailed results can be found in Table 6.3 and C.11.

For the previous databases a decrease in assignment performance can be detected (e.g. from 98,37% to 91,68% for the 2009 database). A possible explanation of this effect is the usage of a database and a mapping file from two different time-points. GI numbers and taxonomic identifiers are updated if new versions of sequences or taxa are available, invalidating older

Dataset	Assigned	Not Assigned
2013	99.91%	0.09%
2011	97.51%	2.49%
2009	91.68%	8.32%

Tab. 6.3: Table of assigned and not assigned sequences of the NCBI-NR database using the new GI parser in MEGAN.

mapping information. This has to be kept in mind when redoing an analysis: Reanalyzing a BLAST result using solely an updated mapping file may not increase the assignment efficiency, it may eventually decrease significantly.

We also calculated the taxonomic coverage as previously defined to estimate the quality of our new mapping method combined with the database. The result shows an overall decrease in taxonomic coverage. Up to 5% less of the leaves or taxa are represented in the new assignment. Table C.12 and C.12 show detailed results. Figure 6.5 visualizes this result.

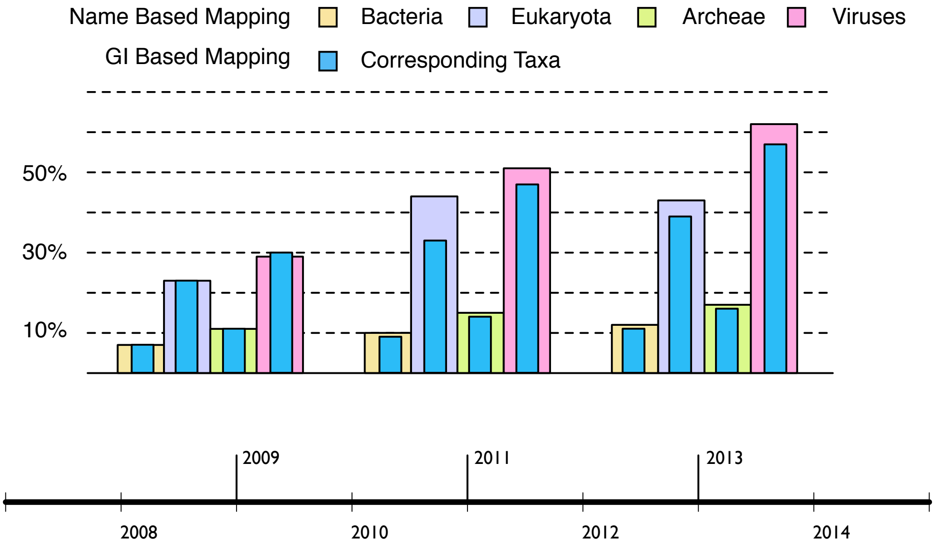


Fig. 6.5: Taxa (leaves) covered by the NCBI-NR databases for the four main subtrees *Bacteria*, *Eukaryota*, *Archaea* and *Viruses* using the old name based method (big bars) and new GI based approach (small, blue bar).

With the total assignment efficiency increasing, the decrease in taxonomic coverage can be explained by previously incorrectly placed matches of the text parser algorithm - compared to the mapping file. The second explanation is that the mapping file provided by the NCBI is faulty. We will later analyze an *in vitro* simulated metagenomic dataset to further assess performance of both methods. A simple way to determine whether text description or the GI number provide the correct mapping entry does not exist.

KEGG Mapping

The underlying KEGG organism:gene to KO number mapping is based on the last freely available KEGG version dated 30.06.2011. This mapping includes functional information of 1,526 different organisms. This is a limitation as only functions similar or related to the already annotated ones can be classified during the analysis.

The latest KEGG release list 2,440 organisms to be included, but paid service strategy of the KEGG consortium denies access to all relevant information via FTP and does not permit bulk load transfers via HTTP. For this study we decided to use the web API nonetheless to access the necessary data for evaluating KEGG for further use.

We constructed a new mapping file using different information accessible via the HTTP REST-API available at `http://rest.kegg.jp`.

`http://rest.kegg.jp/list/ko` Provides a list of valid KO numbers.

`http://rest.kegg.jp/list/organism` Provides a list of valid organism identifiers.

`http://rest.kegg.jp/conv/<organism>/ncbi-gi` Conversion tables from KEGG-GeneID to GI per organism.

`http://rest.kegg.jp/link/genes/<KO>` KEGG-Geneid to KO number translation tables

The GI mapping information provided by the KEGG maintainers through the API is only limited, as only few GI mappings were generated. To enhance the mapping we therefore used additional mapping information provided by the Uniprot consortium [97]. The Uniprot database is highly cross referenced and provides a daily up to date mapping file between different known identifiers. The *idmapping.dat* file downloaded on the 6th of February 2013 contains a total of 380 million entries. Out of the total approximately 52 million entries are GenBank identifiers and 8 million are KEGG entries. Technically the Uniprot mapping offers a basic Gi to KEGG organism:gene mapping which can be combined with data generated from the web API.

All information was merged to generate a new direct mapping file from GI to KEGG KO. The whole process is displayed in Figure 6.6.

The resulting mapping file contains about 6 million entries, tripling the number of potential valid mapping entries. MEGAN was adapted accordingly to use this file instead of the RefSeq identifiers for functional KEGG mapping. The new mapping file was used during the import of the databases and the assignment efficiency was calculated as in the previous section.

Final results show an increase of assigned sequences to KEGG function ranging from 2% to 7% of total reads, doubling the number of assigned sequences for the latest database. Results can be found in Table 6.4 or C.14 respectively.

Though we were able to double the number of assigned reads still a huge number of reads miss a functional role. One reason for this may be that only a fraction of proteins in the database have a known function and if they have it is unclear if they are represented in one KEGG pathway and have a valid KO number. Additionally the high number of new mappings may result from the same protein being active in more than one pathway. That's why the number of valid mappings triples and the number of assigned sequences only nearly doubles.

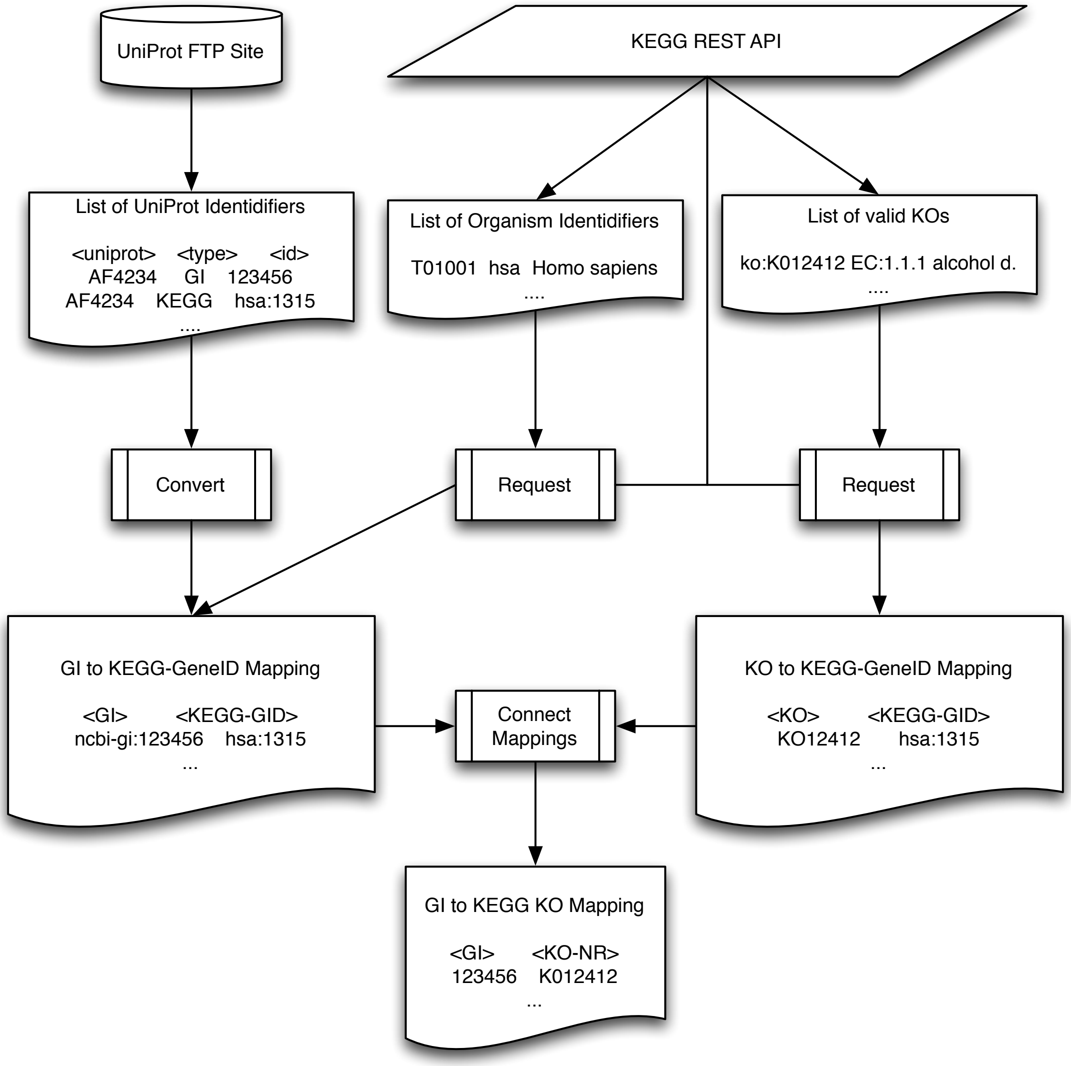


Fig. 6.6: Visualization of the KEGG mapping file generation. The KEGG REST API is available at http://rest.kegg.jp. Additional mapping information using UniProt was generated and merged with information from the KEGG API.

Dataset	Assigned Sequences	Not Assigned	Change (Assigned Sequences)
13	13.97%	86.03%	+7.21
11	15.55%	84.45%	+4.77
9	17.26%	82.74%	+1.94

Tab. 6.4: Table of sequences from the NCBI-NR database mapped to KEGG KOs with the new GI based mapping method.

SEED Mapping

The original RefSeq to SEED mapping file is based on a, now outdated, freely available file from theseed.org providing GI to SEED Subsystem identifiers. Maintainers of the SEED do not provide any updated (mapping) information on their websites. Additionally the Uniprot database does not offer any SEED based mapping entries. The only way to access information about the SEED functional classification system is via a perl API. Compared to the KEGG REST API the SEED API is not well documented and function of the servers is interrupted on a random basis. The server does not provide a success or fail status, so there is no way to differentiate between complete or incomplete data transfer while accessing the information.

The API package includes a variety of example servers offering different functionality with the already mentioned restrictions.

svr_all_features Returns a list of all features.

svr_all_subsystems Returns a list of SEED-subsystems.

svr_subsystem_roles Generates a list of all valid roles.

svr_aliases_of Returns known aliases of the features given as input.

svr_function_of Additionally return the functions of a given input set.

All information combined results in a GI to SEED mapping file. Figure 6.7 visualizes the mapping file generation.

As with KEGG a dummy tree containing only the node *Assigned* and *Not Assigned* as intermediate nodes and all functional roles as leaves was generated to minimize the effect of a potential incomplete tree.

The resulting file contains 6.1 million entries. We also converted the old RefSeq-ID to SEED mapping file to an GI to SEED mapping file using Uniprot RefSeq to GI mappings. Merging both files we created a final mapping file consisting of 7.2 million entries.

MEGAN was modified to use the new mapping file and all three databases were imported. Table 6.5 shows the final results. Total numbers can be found in Table C.15.

Dataset	Assigned Sequences	Not Assigned	Change (Assigned Sequences)
13	2.67%	97.33%	-0.03%
11	3.76%	96.24%	-0.55%
9	6.09%	93.91%	-0.84%

Tab. 6.5: Table of sequences in the NCBI-NR database mapped to SEED IDS with the new GI based mapping method.

Unfortunately the new mapping method is outperformed by the old existing RefSeq based method. One reason may be that the quality of the alias data the SEED servers provide is of very high, or file transfer was interrupted at some point during transmission. The decrease in performance with newer databases may indicate that the mapping data SEED provided is not up

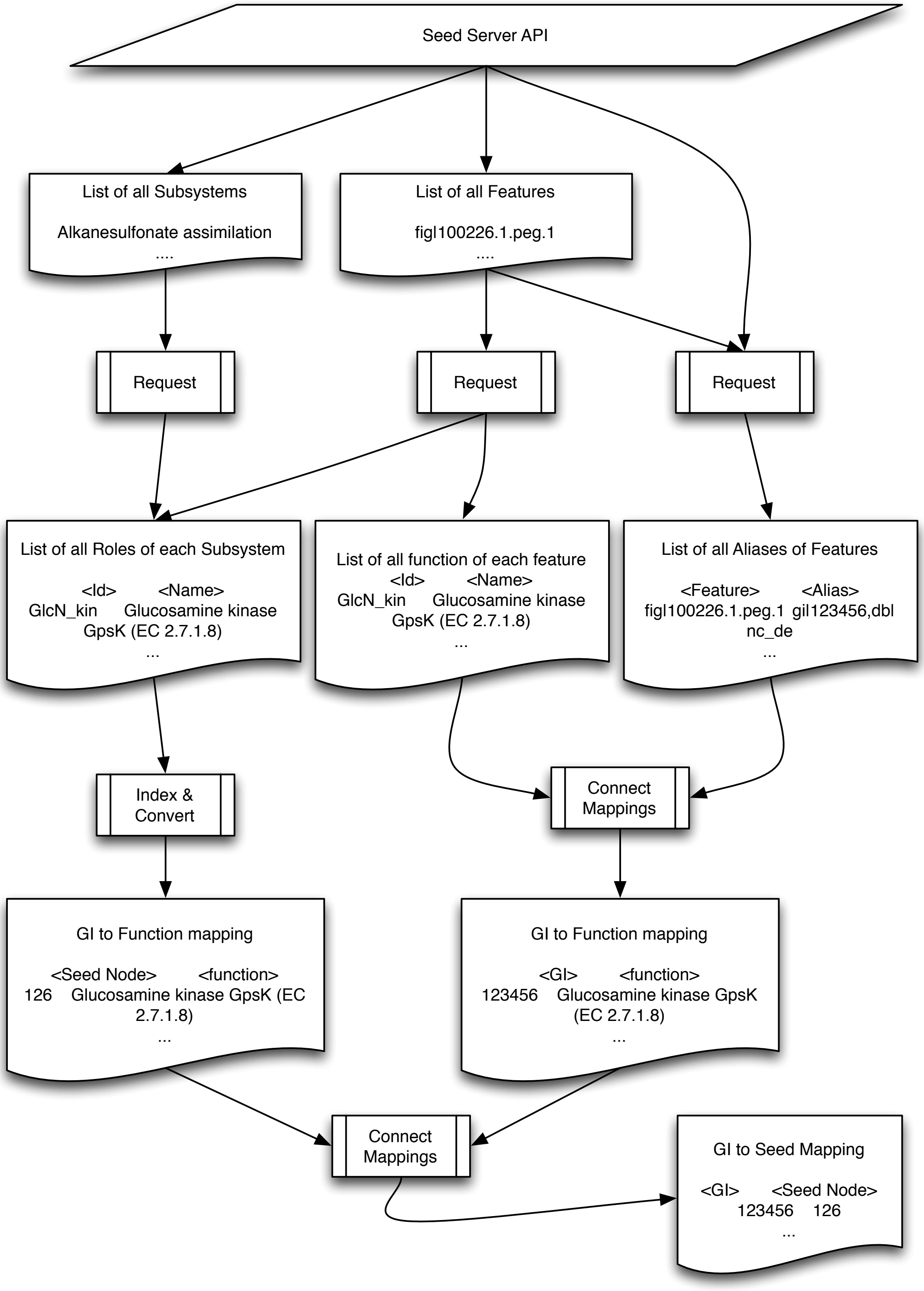

Seed Server API
List of all Subsystems
Alkanesulfonate assimilation
....
List of all Features
figl100226.1.peg.1
....
Request
Request
Request
List of all Roles of each Subsystem
<Id> <Name>
GlcN_kin Glucosamine kinase GpsK (EC 2.7.1.8)
...
List of all function of each feature
<Id> <Name>
GlcN_kin Glucosamine kinase GpsK (EC 2.7.1.8)
...
List of all Aliases of Features
<Feature> <Alias>
figl100226.1.peg.1 gil123456,dbl nc_de
...
Index & Convert
Connect Mappings
GI to Function mapping
<Seed Node> <function>
126 Glucosamine kinase GpsK (EC 2.7.1.8)
...
GI to Function mapping
<GI> <function>
123456 Glucosamine kinase GpsK (EC 2.7.1.8)
...
Connect Mappings
GI to Seed Mapping
<GI> <Seed Node>
123456 126
...

Fig. 6.7: Visualization of the SEED Mapping generation. The SEED Server API provides information about available features and subsystems. With this a list of roles, functions of features and aliases of features can be queried. All information is connected for the final mapping.

to date. The same phenomena was encountered when using an updated mapping file with an older database.

6.3 In vitro simulated Metagenomic Dataset

To asses the assignment performance and quality of the GI based mapping approach we reanalyzed an *in vitro* simulated dataset from Morgan et al. [98]. In the study the authors selected ten different organism for which the whole genomes are known. The species are from all three domains of life and represent highly related organisms as well as only distantly related ones. The original composition as well as additional information can be found in the original study. A known number of cells of each organism were selected and combined. Thereby a synthetic metagenome with a known sequence content was created. The cells were then sequenced using different established metagenomic protocols. An overview of the result of the original study can be found in Figure 6.8.

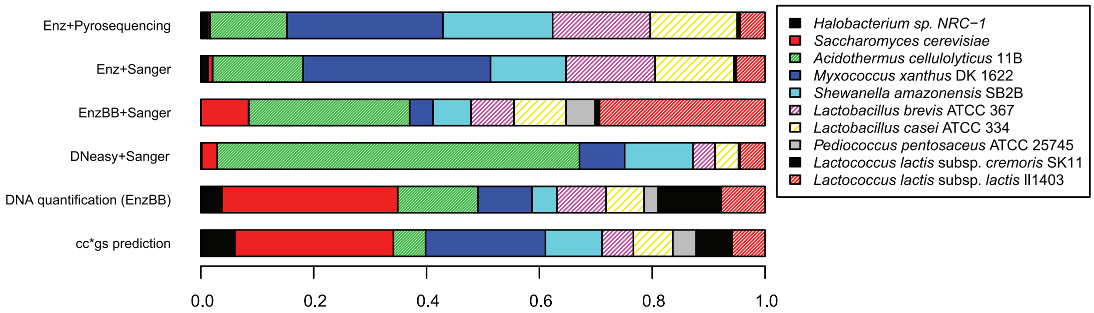


Fig. 6.8: Predicted and observed frequencies of sequence reads from each organism using different sequencing technologies. Figure from [98](open access CCAL, no special permissions required).

Besides the known biases the authors suggest to use this metagenomic dataset to asses the binning accuracy of already established and newly developed tools. Here we will use the reads from this study generated by a 454 GS 20 (pyrosequencing) to compare both mapping approaches.

The sequences were downloaded from the NCBI Short Read Archive with accession numbers SRR033547, SRR033548 and SRR033549. Reads were extracted using the sratoolkit with the "-W" option activated to trim the files accordingly to the submitters' information. The archives hold a number of 112, 19,837 and 505,962 reads respectively. Datasets will be referred to as their relative size *small*, *medium*, and *big*. Reads were blasted against the database downloaded in January 2013 using $BLAST + $2.2.27 with default options except *-show_gis* was set to ensure that GI numbers are included in the results. Resulting files were imported into MEGAN using the default options and in a second run using the new optimized mapping methods.

6.3.1 Results

Based on the number of assigned reads both methods, text-based and GI based, perform nearly identical. Table 6.6 displays the result. Table C.16 contains detailed numbers.

Analyzing the big dataset the new method can assign slightly more reads to taxonomic nodes (0.01%). The assignment efficiency is identical for the other datasets.

	Text Based		GI Based	
Dataset	Reads Assigned	Reads Not Assigned	Reads Assigned	Reads Not Assigned
Big	84.84%	15.15%	84.85%	15.15%
Medium	85.15%	14.85%	85.15%	14.85%
Small	12.28%	87.72%	12.28%	87.72%

Tab. 6.6: Assignment efficiency to the NCBI taxonomy of the text based method and the GI based mapping approach.

To further verify both methods we created a synthetic MEGAN dataset containing the expected number of reads according to the ratios published in the original study. Using MEGAN a comparison file was generated for each of the datasets in comparison to the expected values. Only the comparison of the big dataset will be discussed as the same results were observed in all three files.

Figure 6.9 shows the comparison of the synthetic set and both mapping methods on the *Kingdom* rank. Both mapping methods underestimate the abundance of Eukaryota and Archaea, whereas the Bacteria are well covered. The text based method also overestimates the number of Viruses more clearly than the new method. Also the old method is slightly less specific in its assignments.

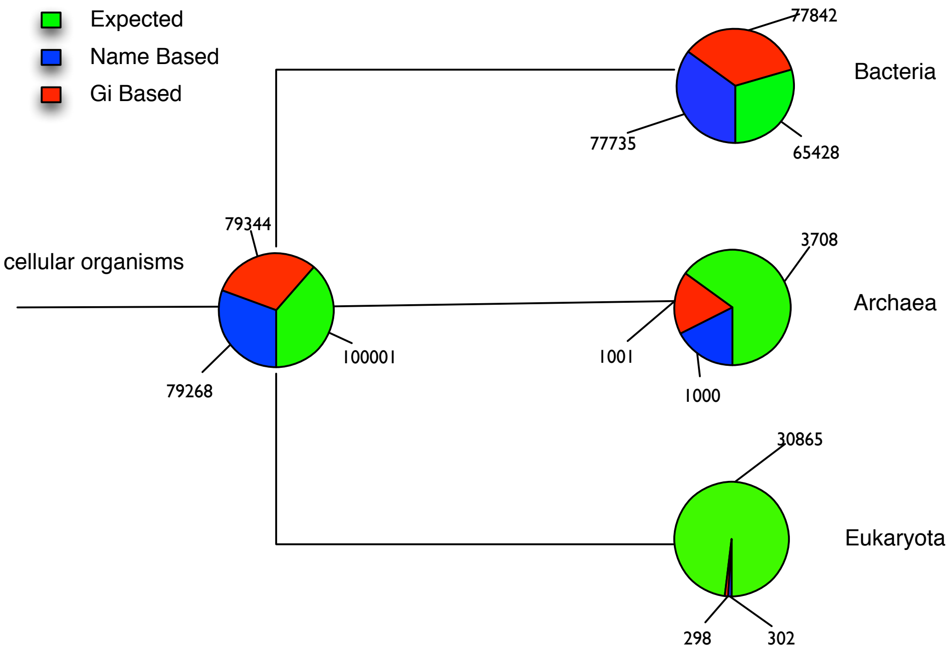


Fig. 6.9: Comparison of read assignments to the taxonomy of both mapping methods and the number of expected reads at *Kingdom* level.

When exemplarily focusing on the only species of Archaea present in the sample the classic method is able to assign more reads to the specific strain. The newer method classifies the reads to the corresponding species, so the mapping information provided by the NCBI is introducing this effect. On species level the GI method assigns two reads more in total, see Figure 6.10.

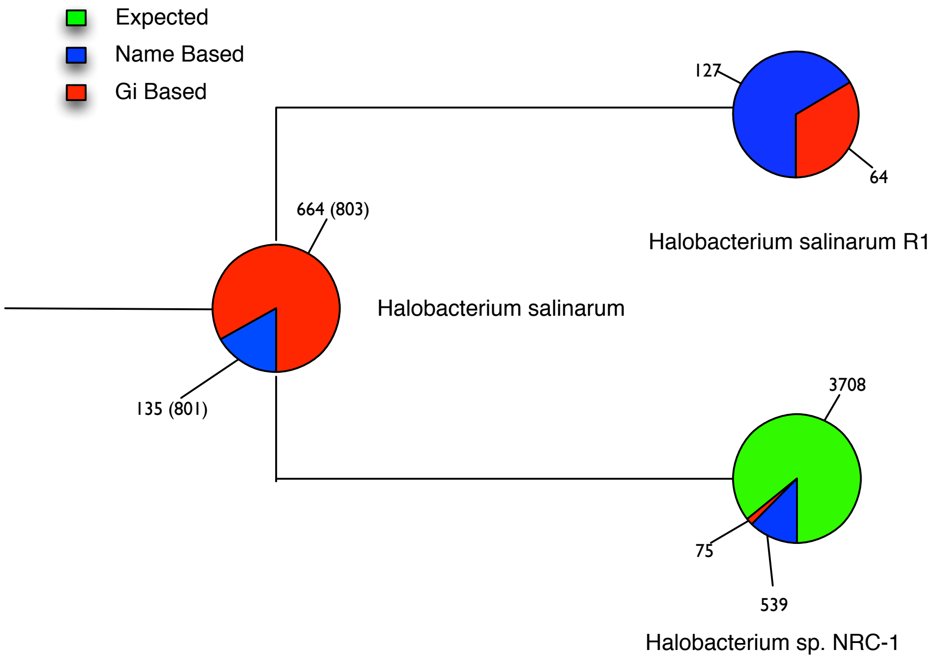


Fig. 6.10: Read distribution for halobacterium. The chart displays the number of reads assigned (sum of all reads). The name based method assigns more reads to the specific strain. Overall assignment is only marginally better with the GI based method (802 vs. 801 reads)

On the species level under the Bacteria subtree the new method is able to perform better and manages to match slightly more reads to the specific species in the sample. As an example Figure 6.11 displays read distribution for the Lactobacillus group.

On species level the GI based method is able to assign 50% more reads to the correct species than the the text based method. This trend can be seen throughout the whole dataset. With the total read efficiency being the same, the new method provides in total a higher number of true positive and more specific hits. This is a major improvement: The more specific assignments of reads enable a more detailed look into the sample.

Functional Analysis

The authors of the original study do not take the functional content of the in vitro simulated metagenomic dataset into account. We therefore can only compare old and new mapping methods based on the number of assigned reads. Results for the small dataset are listed in the tables, but not further discussed because of the extremely low read count.

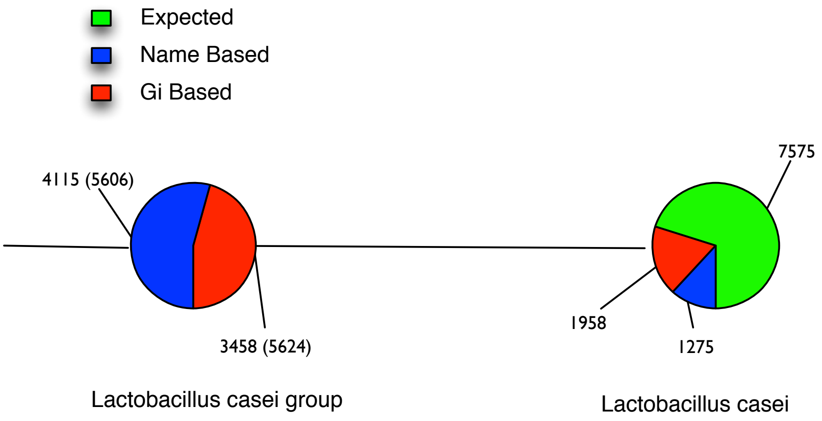


Fig. 6.11: Read distribution for the Lactobacillus group. The chart displays the number of reads assigned (sum of all reads). The GI based method is more specific as more reads are assigned to lactobacillus casei. The text based method assigns more reads to the higher group but is still slightly outperformed by the GI based method.

The mapping performance to different KEGG KOs is shown in Table 6.7, see Table C.17 for exact numbers.

	Text Based		GI Based	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
Big	44.75%	55.25%	46.71%	53.29%
Medium	45.39%	54.61%	47.26%	52.74%
Small	7.02%	92.98%	8.77%	91.23%

Tab. 6.7: Assignment efficiency to KEGG KOs of the RefSeq based method and the new GI based mapping approach.

Overall assignment rate is good compared to the results of the database analysis in the previous section. For all three datasets the GI based method performs better with 1-2% of more matches assigned to a KEGG function. For this relatively small datasets this means about 10,000 more reads can be classified in case of the biggest dataset. We expect the performance to be the same on bigger datasets which will easily result in 100-250k more reads assigned to a known function which is a good result.

The functional SEED analysis is shown in Tables 6.8 and C.18.

The overall assignment rate is with 33-37% not as high as the KEEG assignment rate but still higher than expected. The new method again outperforms the RefSeq based method by up to 4.5%. This is surprisingly good because the maximum mapping efficiency as shown in the previous section is lower with the new method. This number could be artificially low because

	Text Based		GI Based	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
Big	32.87%	67.14%	37.11%	62.89%
Medium	33.17%	66.83%	37.46%	62.54%
Small	3.51%	96.49%	3.51%	96.49%

Tab. 6.8: Assignment efficiency to SEED functions of the RefSeq based method and the GI based mapping approach.

only a low complexity dataset was used for evaluation with high chances that reference sequences are mostly covered by RefSeq identifiers as well established genomes were selected. Real life metagenomic datasets will introduce new species with only GI number mapping (if at all).

6.4 Conclusion

General problems occurring within metagenomic analyses are database coverage and functional classification of sequences. The number of high quality and well annotated sequences added to reference database is comparable small, although databases get bigger. The functional classifications of bacterial sequences are only known for a fraction of organisms. Especially SEED seems to be outdated and not well maintained. Replacing SEED with an updated system such as Clusters of Orthologous Groups of proteins (COGs) [99] and derivated systems such as KOGs [100], and eggNOGs [101, 102] may be an interesting alternative depending on further development in the next years. Overall the mapping accuracy is highly dependent on the freely available data supplied by the authors and maintainers of databases and functional classification systems.

The method for assigning reference sequences to function or taxa is a key element of the analysis and responsible for the quality of whole study.

We have shown that it is possible to generate a GI based mapping approach which outperforms the currently used text based and RefSeq based mapping approaches. On the taxonomic level our method was able to assign the same number of reads to a more specific level. Both results for KEGG and SEED show that the GI based mapping improves the functional analysis. Results were confirmed using a synthetic dataset and showed that the more specific hits are indeed true positive assignments.

With the GI based method we are still able to use the new tabulator based BLAST format for reduced file size. MEGAN will in that case compute the alignment from this output and will present it to the user. With the limitation that taxonomy, mapping file, and database have to be kept in sync to avoid artifacts seen during the analysis, the new method outperforms the older one. The new assignment method has been integrated in MEGAN and is available as option during import of BLAST files. We therefore suggest to use the new method in new ongoing analysis.

Chapter 7

Processing, Accessing and Sharing High Volume Data

Analyzing a single sample requires various different steps to get a reliable result as we have seen in the previous sections. The used tools offer, most of the time, a multitude of options and are often only accessible via the command line. The combination of different tools sometimes requires an intermediate conversion step transforming one file format into another. The output produced by each step of the analysis leads to a high number of files which need to be managed efficiently.

Additionally the user has to wait for a single step to complete before he can continue with the next step. The sighted user can automate this process to some point, but still has to invest some time into the process. When processing multiple samples the same process has to be repeated all over again, sometimes with small adaptions. The analysis results and sometimes the analysis as whole needs to be made available to a multitude of people, often not coming from the same background.

Besides this, the installation of the tools itself and more often of the required dependencies take some time and are often difficult to complete for people without training in a computer related discipline.

Running the analysis within a centralized environment has multiple advantages. Tools and data have to be only maintained in a single place enabling easy management and sharing. Especially web-based solutions enable flexible access to a broader audience. Various public instances of services promising easy analysis of metagenomic data already exist as introduced in Part I. Transfer volume, data safety and security as well as sharing of computational resources may be disadvantages when using those platforms as sole source for the complete analysis.

7.1 Galaxy-Server

To match the different requirements for processing and accessing high volume data we suggest to use a local instance of the *Galaxy* platform. Galaxy is an open, web-based platform for the analysis of high-throughput genomic data. It was originally developed by Goecks et al. [92, 103] as "a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". A public instance of the website is available via `http://usegalaxy.org`, but this introduces previously mentioned disadvantages as not allowing users to implement their own tools and sharing of computational power between all users. Also the public galaxy is more focused on genomics than metagenomics only offering a small toolset for metagenomic analyses. With galaxy being open-source it is possible to set up a local server which than can be modified to specific needs. This includes management of tools, data access, and computational resources in an efficient way.

7.1.1 Technical Design

The local Galaxy-Server is accessible via `https://galaxy.informatik.uni-tuebingen.de`. To manage computational resources and sensitive experimental data access is restricted to selected user accounts of the department and extra local user accounts to offer external collaboration partners easy access. Web access is limited to the secure HTTPS protocol to ensure data privacy and integrity. The server manages multiple instances offering a *stable*, *legacy version*, and *local* instance which is a development snapshot of the tools. This ensures that the productive instance will continue working while new tools are implemented and updates of the base system are imported. To increase responsiveness the different instances serve more than one web-server thread per instance. Each entity is accessible through its own URL and is proxied through Apache to load balance access to the server. The Apache web-server in combination with multiple PAM modules is also handling user authentication and authorization. Jobs are run by the job-runners locally or computational expensive jobs will be submitted to the local compute cluster for computation. Scheduling of cluster jobs is done by the SunGridEngine (SGE) of the cluster-master. Metadata is stored in a local PostgreSQL database for newer instances, the read-only legacy version has metadata stored in a MySQL database. The file system is connected through NFS with the cluster and the departments share. Local workspace is excluded to ensure high throughput for local computations. Full and incremental backups are done through a separate *sshfs* connection to third computer via *duplicity*. The layout is visualized in Figure 7.1.

The instances are based on a single mercurial repository cloned from the public official galaxy repository on Bitbucket `https://bitbucket.org/galaxy/galaxy-central/src`. Base system code and own tools can therefore be easily shared between the instances. Development process and documentation is provided by a content management system (*Redmine*) running on the server (Figure 7.2). Source code, tickets, changelog, news, and documentation can be viewed through the website. This allows easy interaction between involved developers and users.

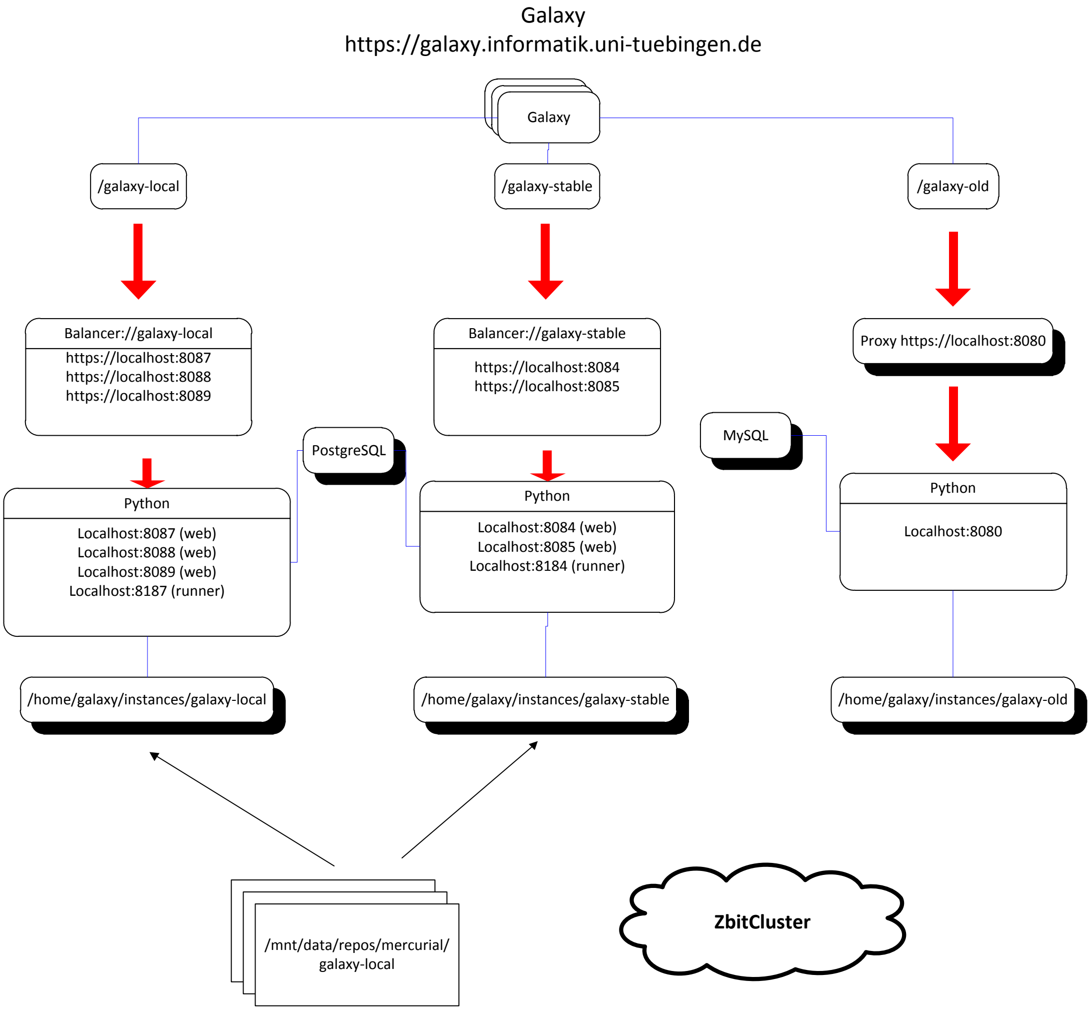


Fig. 7.1: Layout of the local Galaxy Server hosting three different instances. Instances offer multiple web-server which are load-balanced through the systems Apache webserver to ensure responsiveness when accessed by multiple users at the same time.

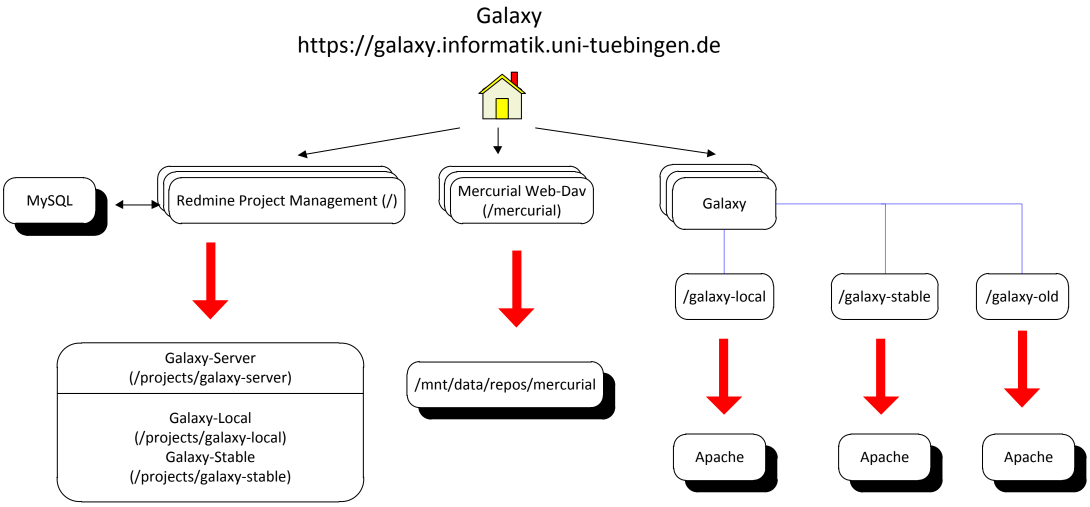

Galaxy
https://galaxy.informatik.uni-tuebingen.de
MySQL
Redmine Project Management (/)
Mercurial Web-Dav (/mercurial)
Galaxy
/galaxy-local
/galaxy-stable
/galaxy-old
Galaxy-Server (/projects/galaxy-server)
Galaxy-Local (/projects/galaxy-local)
Galaxy-Stable (/projects/galaxy-stable)
/mnt/data/repos/mercurial
Apache
Apache
Apache

Fig. 7.2: Beside the galaxy instances the server offers a mercurial repository and redmine environment. Both additional environments are also shared through the main Apache webserver allowing easy interaction for developers and users.

7.1.2 Tools and Pipelines

The graphical user interface makes it easy to use different tools and combine them into whole workflows. The basic galaxy instance already supplies a variety of tool description files allowing the executing of such tools (if installed on the server). Tool description files are XML files containing information about possible parameters, input/output file formats, descriptions and information about job execution. The file is automatically used for generating the user interface during interaction (Figure 7.3). New tools can be easily integrated by writing a new specific XML file.

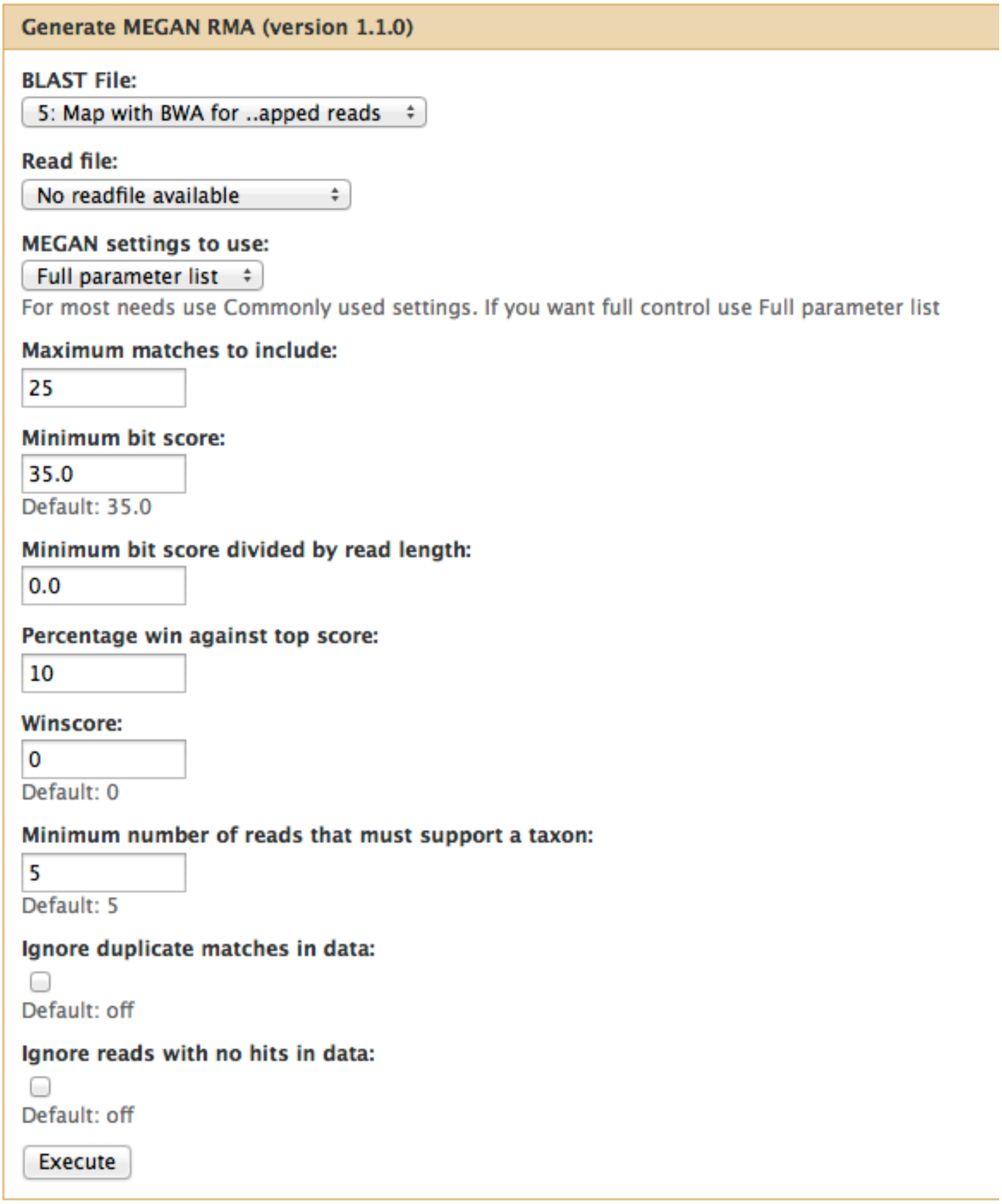
Depending on the project task we integrated already existing tools into galaxy. The following list gives an overview of external tools.

EA-Utils: ea-utils is capable of filtering adapters and trim reads accordingly. It is also able to merge paired end reads into longer fragments [80].

Bowtie: Version 1 and 2 of the short read mapper was included for filtering metagenomic reads which originate from a specific host. Matching reads were removed from the input file [51, 52].

GATK: The Genome Analysis Toolkit (GATK) was integrated to use the *UnifiedGenotyper* for SNP calling, *VariantFiltration* for SNP filtering and *FastaAlternateReferenceMaker* for consensus generation [87].

Picard: The following tools from Picard were included: *addOrReplaceReadGroups* for preparing BAM files for further processing and *MarkDuplicates* for duplication flagging and removal [104].



Generate MEGAN RMA (version 1.1.0)

BLAST File:
5: Map with BWA for ..apped reads

Read file:
No readfile available

MEGAN settings to use:
Full parameter list
For most needs use Commonly used settings. If you want full control use Full parameter list

Maximum matches to include:
25

Minimum bit score:
35.0
Default: 35.0

Minimum bit score divided by read length:
0.0

Percentage win against top score:
10

Winscore:
0
Default: 0

Minimum number of reads that must support a taxon:
5
Default: 5

Ignore duplicate matches in data:
□
Default: off

Ignore reads with no hits in data:
□
Default: off

Fig. 7.3: MEGAN user interface automatically generated by GALAXY. The user can input various parameters or can use the program's default options.

BLAST: Different versions of BLAST and the BLAST+ suite were integrated [55].

samtools: Samtools-wrapper were extended to efficiently convert SAM-files and sort and merge BAM-files [86].

MEGAN: The MEGAN tool-wrapper creates a rma-file from any valid input-format. An optional read file can be supplied as well as options for the LCA algorithm can be set. GI mapping files to use can also be defined [13].

mapDamage : For ancient DNA projects the mapDamage scripts were included. The tool calculates and plots possible DNA damage patterns using a mapping file. Damage patterns can later be used for authentication of the sample [105].

Some tools did not offer the needed capabilities so own tools were developed and integrated as well. The following list gives an overview. If an external tool is required as dependancy a reference is given.

Trimmer: Trims FASTQ reads for bad quality using a sliding window approach. Thresholds can be set accordingly.

Adapter removal: Removes known adapters from the sequence. The tool takes a FASTQ read file and FASTA adapter file as input.

Linker Removal 454: Removes known 454 linker from the reads. Similar to the previous tool, this tool removes 454 bases linker sequences.

Convert Solid: Converts solid color space reads into base space. The tool generates all valid reads where the CS read may originate from.

Plot Coverage: Using Bedtools [106] and GnuPlot the tool plots the coverage graphically over the whole genome.

Calculate Coverage Histogram : Using Bedtools the tool generates a coverage histogram.

Create Consensus: This tool creates a consensus sequence based on a *variant calling file* from GATK. Unconfident or filtered sites are marked as 'N'.

Get Mapping Stats: Extracts information about a mapping file, such as number of mapped reads.

In addition to the tools multiple databases, indexes (NCBI-NR, NCBI-NT, RefSeq, Environmental, viral etc.) and reference genomes (hg19, various model organisms and bacterial strains) have been integrated.

A combination of single tools have been deployed in various smaller projects by multiple users. For more complex tasks the tools have been combined to different workflows. For example SNP calling and consensus creation workflows for ancient DNA research contain up to 46 different steps. A simple exemplarily metagenomic workflow is displayed in Figure 7.4.

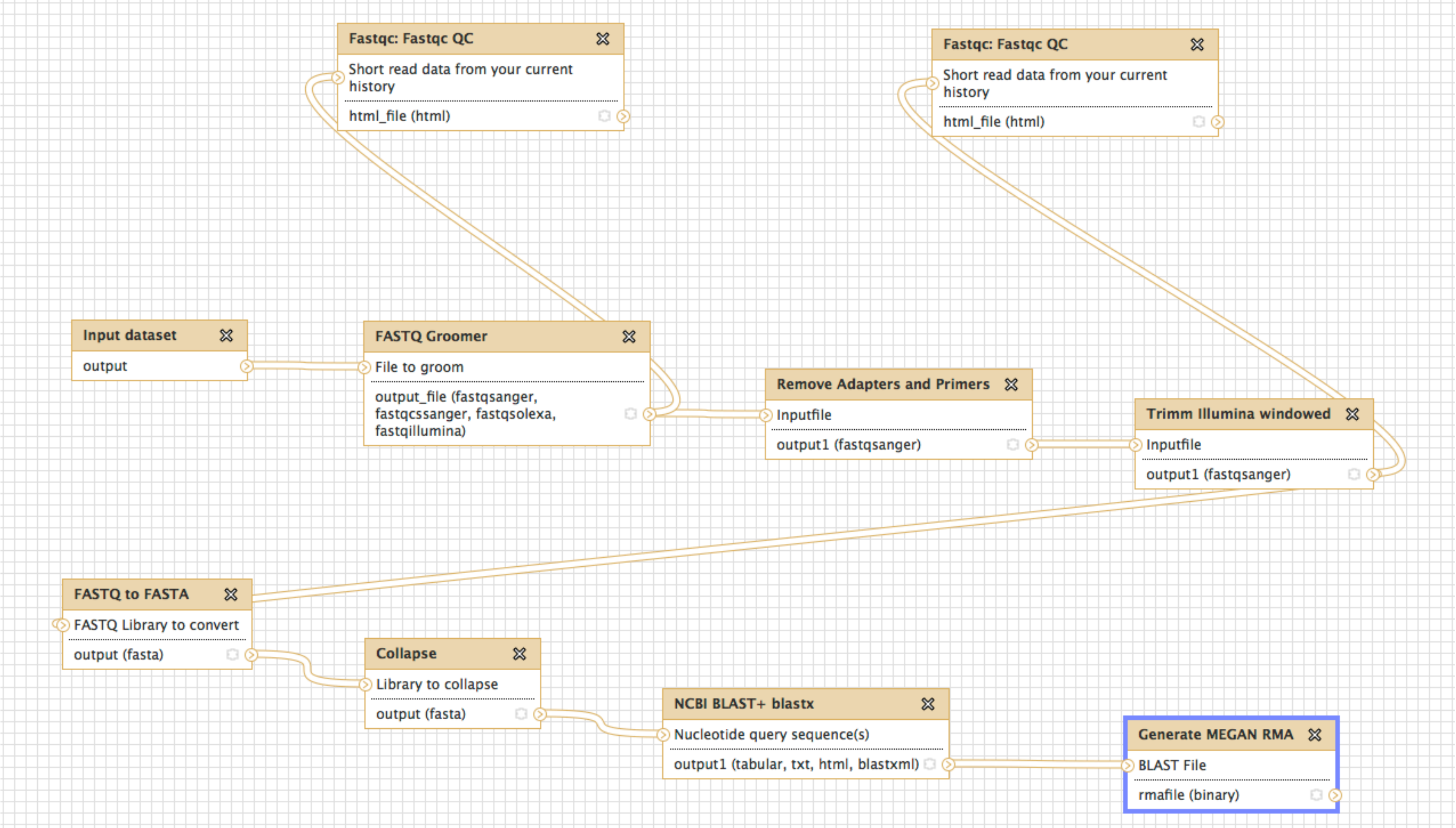


Fig. 7.4: Sample Metagenomic workflow for Illumina data. The workflow starts with sequencing data in the fastq format. After import an initial quality check is done (*FastQC*). The data is quality processed and checked again. After that the file is converted into plain fasta format and blasted using BlastX from NCBI. In the end a MEGAN rma file is generated for the user to download and inspect the data.

7.1.3 Discussion

Aside from work for metagenomics and ancient DNA research the server has been extensively used and extended by Magdalena Feldhahn from Oliver Kohlbachers group, University of Tuebingen. For her project the instance was extended for building multiple immunoinformatics workflows [107]. Tools include additional mapping, SNV detection, SNV epitope prediction and database export. Design and implementation allowed parallel development of tools as well as workflow building, testing and execution. We have demonstrated that galaxy provides a flexible way to use tools already in the basic version, integrate external tools as well as easy integration of own tools. Sharing of results or complete workflows with external collaborators allowed short feedback times while discussing results. The initial investment in time for setting up the server and implementing new tools is easily paid of by savings later on.

Part III

Ancient DNA Research

Chapter 8

The Ancient Mycobacterium Tuberculosis

This chapter is about the analysis of ancient Mycobacterium tuberculosis samples. We first give a short introduction and explain typical challenges in ancient DNA research. Additionally needed techniques are explained. After the analysis a short summary is given.

8.1 Motivation and Background

The disease tuberculosis (TB) remains until today a serious health problem around the world. The *Global tuberculosis report 2012* [108] states that in 2011 approximately 9 million new infections and 1.4 millions TB induced deaths took place. Especially human immunodefficiency virus (HIV) positive patients infected with TB develop the disease. Leaving the disease untreated 70% of the patients die within 10 years. Focus of TB infections are mostly developing countries, but numbers are eventually rising all over the planet. Especially the increased encounter of multidrug-resistant TB strains (MDR-TB), hindering the effective treatment, is a worrying development [108]. A study suggest that TB's over average capabilities of adapting to drugs and effective transmission between hosts is based on the specific genetics and therefore the evolution of M. tuberculosis [109].

8.1.1 Theories on the Evolution of Mycobacterium tuberculosis (complex)

In principal there are two major theories about the evolution of today's M. tuberculosis. A long supported theory was the descent of M. tuberculosis out of M. bovis as a result of a zoonotic event - that is the infection of humans by an animal pathogen. A newer theory suggest the evolution of today's strains out of a ancestral group named *M. prototuberculosis*.

TB: A result of zoonosis: The long time established theory suggest that evolution of TB started with the beginning of agriculture and domestication of cattle approximately 13,000 years ago [110, 111]. M. bovis infected animals are suspected to infect human and as a result M. bovis adapted to the new host and evolved eventually into today's M. tuberculosis [112]. Until now zoonotic events with M. bovis infecting humans and causing TB, are estimated at rates of 1.4-2.5% of all TB cases [113].

M. prototuberculosis: Wirth et al. [114] suggest that the origin of the modern Mycobacterium tuberculosis complex (MTBC) was lied out approximately 40,000 years ago at the Horn of Africa. Their study support the idea of a pool of mycobacteria where the ancestral MTBC emerged from a pool of microorganisms named *M. prototuberculosis* and co-migrated with humans from Africa [115]. They further suggest that 20-30,000 years later the ancestral strain formed two main lineages with one spreading in human population and one being the source of animal tuberculosis. Other studies also suggest that TB spread from human to animal and not vice versa [116, 117]. Nevertheless the theory of *M. prototuberculosis*, is controversially discussed [118, 119].

8.1.2 Ancient DNA: complex, fragmented and damaged

In contrast to living organisms where damage to the DNA is quickly repaired, post-mortem damage is accumulated over time. Different types of damage happen to the DNA molecules over time including oxidative and hydrolysis damage as well as DNA degradation and crosslinks [22]. Typical results of oxidative damage and DNA degradation are strand breaks and blunt ends explaining the short average read length of 70bp typically encountered when handling aDNA [120]. Hydrolysis mainly leads to misconding lesions which introduces a bias of C -> T and G -> A transitions and to a lower amount of A/T -> G/C transitions [121]. This is a direct result of hydrolytic deamination of cytosine to uracil or hypoxanthine respectively. The damage patterns are not evenly distributed over the reads, but tend to pile up at the ends of each strand where blunt ends are more likely to be exposed to the environment [122]. It has been shown that the use of Uracil-DNA glycosylase (UDG) can repair deaminated sites and after further processing those reads can be sequenced [123].

Besides the challenges of fragmentation and damage of aDNA the concentration within the sample itself is often a problem. Usually the sample is a complex mix of past and present microbial and fungal, host and contaminating organisms as well as the organism of interest. Quality and concentration of retrieved aDNA are highly variable depending on the sample and usually in favor of cold environments such as permafrost soil (e.g. [19]).

Removing contamination and authenticating aDNA remain an evolving challenge. Cautious steps need to be taken to ensure the integrity of the analysis. On the bioinformatics side the introduced damage patterns can be used to authenticate aDNA [120].

8.1.3 Sample Enrichment

Especially the low amounts of target DNA in a sample is sometimes problematic as not enough material can be generated for a genome wide analysis. For this reason early studies mostly focused only on mitochondrial DNA (mtDNA) [22]. By contrast nuclear DNA has the advantage of being less accessible for damaging reactions, resulting in higher preservation despite its relatively low abundance compared to mtDNA. Targeted enrichments offer a solution to this problem. Besides SNP arrays the use of custom made *bait* capture arrays are a reasonable option. This technique uses a chip similar to a microarray where target DNA hybridizes to previously synthesized probes on a chip and unbound DNA gets subsequently removed. Remaining DNA is then sequenced [120]. The drawback of targeted approaches is by definition that one can only enrich the sample for known sequences, so a minimal knowledge of the target organisms is required. One can also extend the capture width by taking related species into account when designing the array. Unfortunately completely novel sequences of the organisms will be missed anyway because no capture bait will be available. Recent studies [23, 24] demonstrated that genome wide reconstruction of ancient strains is feasible using this technique.

8.2 Analysis of Ancient M. Tuberculosis Samples

This section describes the analysis of multiple ancient samples from human remains which are suspected to be infected with M. tuberculosis (TB). Samples of ancient remains were taken from different sites mostly from Southern and Northern America. Skeletal markers, such as specific damage to the spine, ribs or phalanx, suggest an infection with TB. The project is done in collaboration with Anne Stone, Arizona State University. The aim of the experiment is to study the evolutionary relationship between known and the ancient strains. For this, suitable samples need to be identified, consensus sequences need to be created and single nucleotide polymorphisms compared to known strains need to be determined. The analysis presented in this work consists of three different steps: A whole genome shotgun approach, an additional screening process by sequencing samples which were enrichment using four specific genes and finally the sequencing of samples which were genome wide enriched using a custom build bait capture array. We will present each approach in detail with a separate section of methods, results and a short discussion. A final conclusion is given in the end of the section.

8.2.1 Whole Genome Shotgun Analysis

Depending on the environment the sample was found in, the storing condition and the sampling location DNA of the sample may display different levels of DNA degradation. A whole genome shotgun (WGS) approach may yield enough endogenous DNA for the complete analysis, depending on DNA preservation. This WGS analysis was designed as screening for potentially interesting samples.

Methods: A total of 102 samples including blanks as controls were prepared and sequenced using a paired-end protocol on an Illumina MiSeq sequencer. Initial quality control (QC), adapter trimming and merging of paired-end reads was done as described in Chapter 4 using ea-utils. QC identified approximately 5-10 bps in the beginning of the reads forming some sort of artificial k-mers which could not be definitely described and reliably identified. Reads were therefore trimmed 5 bps in the beginning to minimize effect on the downstream analysis. Samples were mapped against all available full genomes of the strains M. avium (2 strains), M. tuberculosis (14 strains) and M. kansasii (1 strain). Samples were also mapped against the human genome (hg19) and additionally the human mitochondrial chromosome as a control. Duplication removal was performed and damage pattern calculated using the mapDamage tool [105]. Average and maximum coverage for each reference was calculated. Mapped reads were individually visually inspected if necessary.

Results: After QC average read length was around 60-80 bp with a standard deviation of 10-30 depending on the sample. For most samples the number of mapped reads to the different complexes was below 0.5%, different spikes were observed with a maximum of 1.5% of mappable reads. Coverage analysis for the individual sequences indicated that on average more than 99% of the reference sequence was not covered. Reads for most samples often clustered in short positions with a very high coverage. Calculated damage plots showed ancient DNA specific damage patterns, but seem artificially flattened on both ends. See Figure 8.1 for an example.

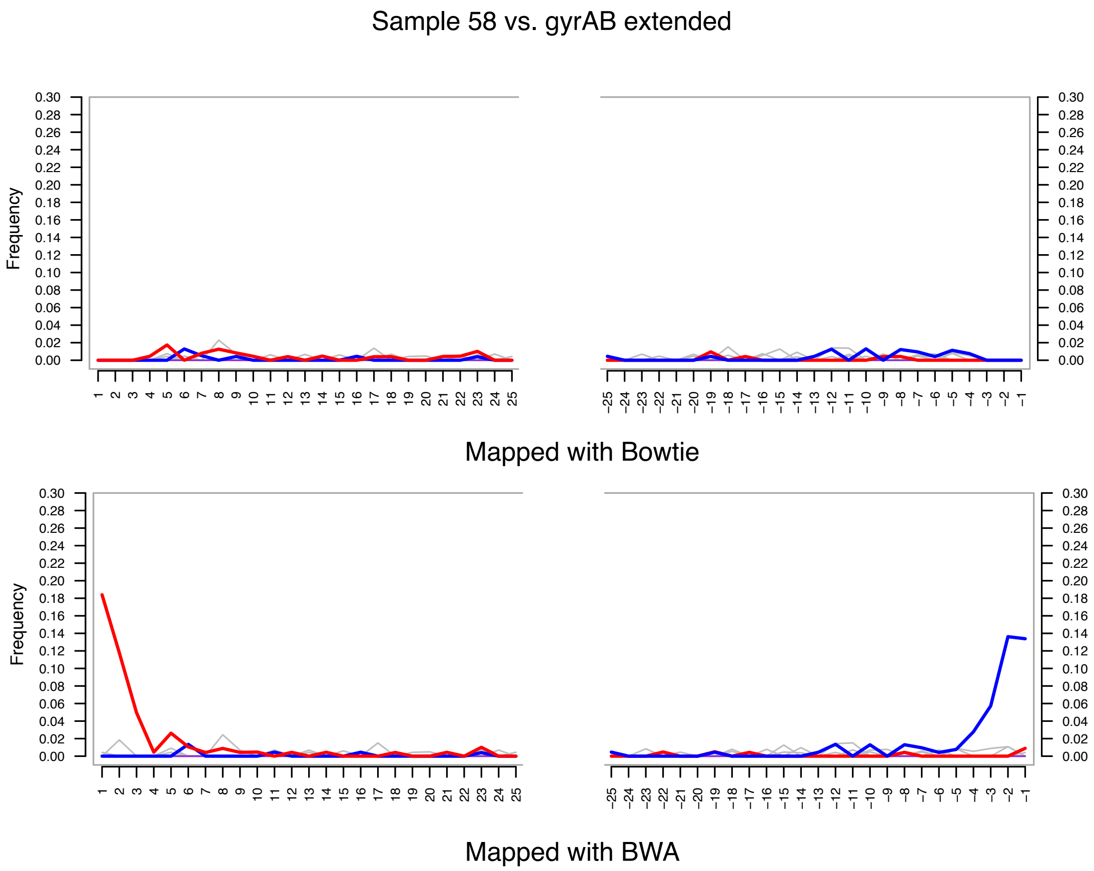


Fig. 8.1: Damage pattern of sample 58 mapped against gene gyrAB (extended). The lower plot is based on the mapping done with BWA. The upper plot is based on mapping done with Bowtie2. One can clearly see the artificial flat ends induced by soft-clipping of Bowtie2. Data taken from the analysis with gene enriched samples for demonstration purposes.

Regarding pathogen coverage Sample 58 has the highest percentage mapping to the tuberculosis complex with 1.58% of reads mapping against the reference. Those reads only span 0.01% of the genome with 100 fold coverage.

A few samples show an exceptionally high number of reads mapped to the human genome with up to 77% of mappable reads within a single sample.

Sample STG380 mapped exceptionally well to the human hg19 genome (77%) and a low number of reads mapping to the human mitochondrial DNA (mtDNA) with only 0.02% of all reads. For the mtDNA coverage calculations showed a 30% coverage with 1-2x fold and maximum coverage of 3. Reads from this sample are well distributed along the chromosome and not clustered in few positions.

Discussion: Preliminary authentication of aDNA was done using average read length and DNA damage pattern. The overall short average read length can be seen as indicator that samples are actually containing ancient DNA. Damage plots support aDNA characteristics except of the flat ends encountered during the analysis. The low percentage of mappable reads to reference sequences has been already described by previous publications (e.g. [23]) and is no indicator for potential problems.

The analysis itself has three major concerns which can be only partly explained during this discussion: flat ends, clustering and k-mers. Although the flat ends could not be explained originally at this point, a later analysis discovered the source of the flat ends. The reason for this phenomena is the combination of the mapDamage tool and the mapping with Bowtie2. During setup we decided to use Bowtie2 in *local-mapping* mode, because we expect the damage pattern to interfere with the default end-to-end mapping. This was successful as more reads got mapped to the reference sequences. Bowtie2 soft-clips the end parts of the reads not in the alignment (e.g. marking bases as not used). Technically the mapping process is well within its defined range. The problem is that the mapDamage script does not consider those clipped positions for damage calculation. In detail the script is considering the soft-clipped ends for length calculation, but not for damage pattern. We assume that this is mainly because the script was designed to work with BWA which does soft-clipping of positions only in cases of bad quality. Masking bases using soft clipping is reasonable when low quality bases would interfere with damage pattern calculation.

Using BWA for mapping can quickly circumvent the problem, also BWA has been already established in different pipelines. The drawback is that BWA does not support local alignments and thus decrease mapping performance because of the damage of aDNA interfering with mapping efficiency.

The second problem, clustering of reads at specific positions could be potentially problematic during a full WGS analyses. Clustering of reads is often some sort of PCR or amplification error. Usually reads can be observed outside clusters as well which is not the case here. One explanation may be the relatively low fragment count per sample which may increase competition during library preparation. Highly amplified fragments will then be preferentially sequenced resulting in clusters. It is also unclear why some samples do not show this behavior. Again the amount of initial DNA during sample preparation may be one reason as samples with high amounts of DNA may potentially get more evenly amplified.

Artificial k-mers in the beginning of the reads could be a result of an unknown and uncontrolled process during the wet lab phase of the project. As the k-mers could not clearly be identified a removal is only partly possible. Other sequencing runs done in the same time-frame seem to encounter similar artifacts [124], so the problem may be sequencing chemistry or the sequencer itself. Artificial parts of the sequence may interfere with the mapping process and can in theory also hinder effective duplicate removal. Incomplete duplicate removal will also support generation of clusters during mapping.

Potential problems during wet lab processes and especially sequencing can be hard to track down and usually can not be solved in silico as multiple preparations and sequencing runs should be performed to pinpoint the exact problem.

Screening using the WGS did not clearly identify potentially interesting samples for the ancient M. tuberculosis project. Sample STG380 will be further investigated in another project because of the exceptionally high number of reads mapping to human hg19. For this study the sample has been excluded as pathogen coverage was not as high as expected. Results show that using WGS for these samples is not practicable as low amounts of DNA preserved within the single samples may cause different unresolved problems during the analysis. An alternative to WGS is the enrichment of the different samples prior to sequencing as already introduced.

8.2.2 Gene Enrichment Analysis

As a preparation to an enrichment using a bait capture array a number of samples were screened to identify suitable samples. Screening was done using four different genes gyrAB, rpoB, katG and mpt40. Three genes rpoB, katG and mpt40 are specific for the M. tuberculosis complex while the the combination of gyrA and gyrB is also present in Mycobateria in general. Using the genes one should be able to estimate the potential M. tuberculosis content of samples and make a final selection for a detailed genome wide analysis.

Methods: Samples were enriched for the four genes prior to sequencing, because of the low amount of DNA of the samples. Sequence information about the four genes for enriching were extracted from the H37Rv strain of M. tuberculosis. Sequencing was done using a paired-end protocol on an Illumina MiSeq. After sequencing QC was done accordingly. All samples were mapped against the four specific gene regions. Additionally mapping was done using extended versions of the genes using extended fragments of the genes including 150bp up and downstream of the region. This was done to ensure that reads which originated from the ends and are overlapping were also mapped correctly. As control the reads were also mapped against human hg19 and the human mitochondrial chromosome to spot potential contamination issues. Average coverage was calculated and coverage distribution was plotted for the extended and not extended fragments. Damage plots were calculated for all completed mappings. Alignments for samples 64.U, 65, 54, 58, and 162 against the extended gene sequences were further manually inspected for final decision of sample selection. For selected samples all reads were compared to a reference database using blast to ensure that no cross contamination occurred. Blast results were analyzed using MEGAN.

Results: Average read length after QC was between 50-70 bps with a standard deviation of 15-25 depending on the sample. Mapping results showed up to 2% of mapped reads for all samples, but was usually in the range below 0.5%. Especially the number of mapped reads for katG and mpt40 usually was below 0.01%. As contamination control the mapping to hg19 and mtDNA was usually negative with 0% of mapped reads. Initial damage plots showed flat ends, but the issue was resolved and final damage plots showed typical aDNA damage patterns. Coverage for the specific regions varied widely: Depending on sample and gene 9% to 100% of the reference sequence was not covered. Selected samples showed average coverage between 1.3x and 27.6x, see Figures 8.2 and 8.3.

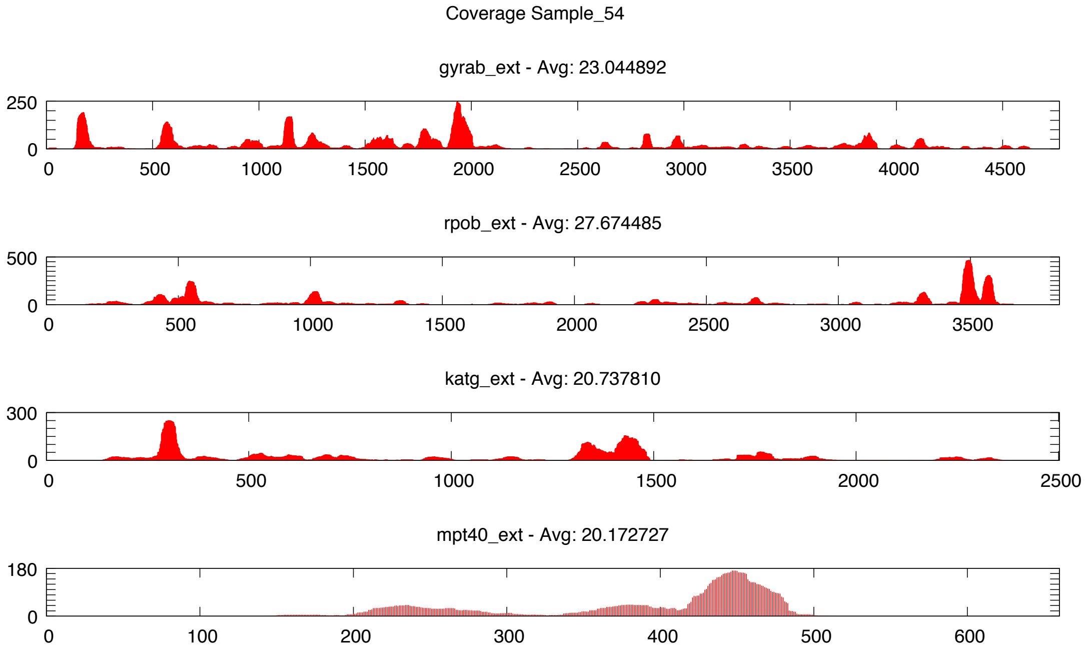


Fig. 8.2: Coverage plot of sample 54 mapped against the four extended gene fragments. The ends of the fragments are not covered as for this regions the sample was not enriched.

Sequence comparison against the NCBI-NT collection using blast for the final three datasets 54, 58, 64 showed significant matches for the M. tuberculosis complex only.

Discussion: Average read length and DNA damage patterns support preliminary authentication of ancient DNA. Sequence coverage do not suggest high read clustering as previously observed. Coverage of gyrAB is not exceptionally huge, which provisionally suggests that reads originating from another Mycobacteria complex did not contaminated the samples. The additional sequence comparison of all mapped reads using blast showed no cross contamination, see Figure 8.4.

Based on this result and final manual inspection the samples 54, 58, 64 where selected for the genome wide enrichment and analysis. All selected samples originate from Peru with a

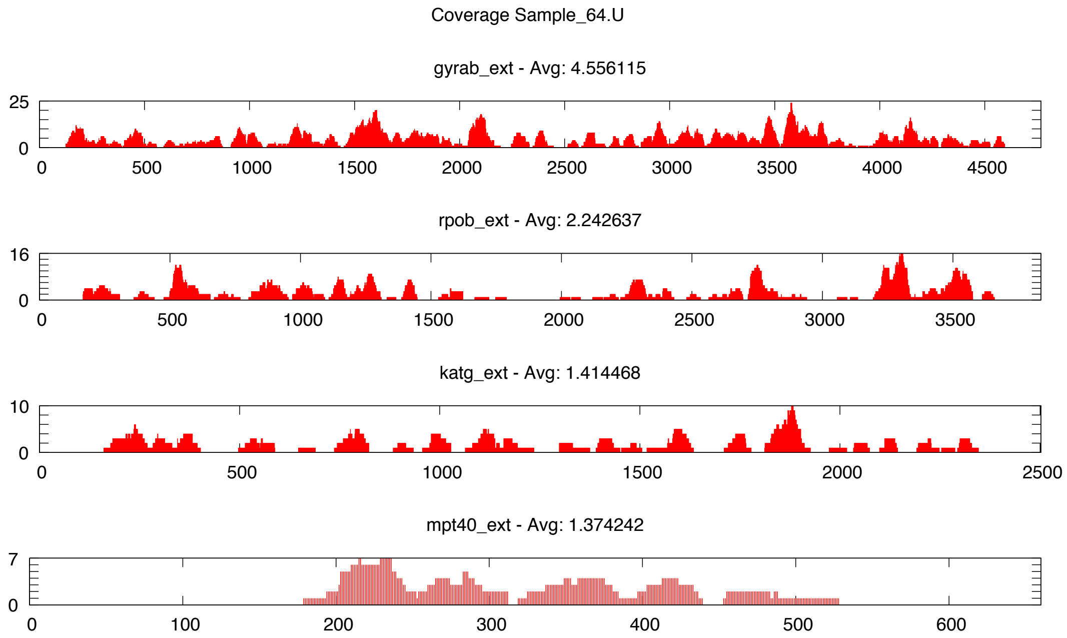


Fig. 8.3: Coverage plot of sample 64.U mapped against the four extended gene fragments. The ends of the fragments are not covered as for this regions the sample was not enriched.

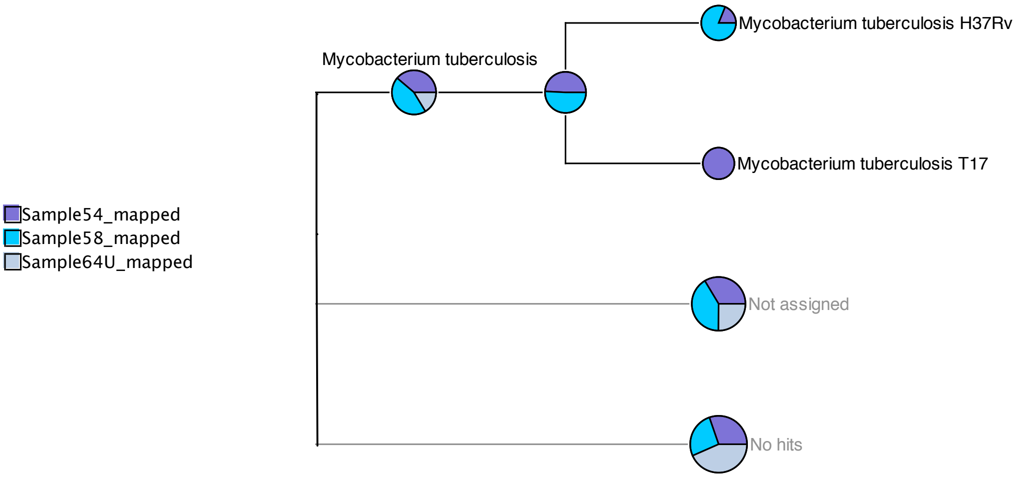

Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis
Mycobacterium tuberculosis T17
Sample54_mapped
Sample58_mapped
Sample64U_mapped
Not assigned
No hits

Fig. 8.4: Taxonomic assignment of samples 54, 58, and 64 using the NCBI-NT database. Visualization only include previously mapped reads to reduce noise.

find spot at high altitude. The dominating colder climate at these heights may have been favorable for DNA preservation.

8.2.3 Genome Wide Capture Array

Combining the results of the gene enrichment and whole genome shotgun analysis, we focused on the Samples 54, 58, and 64 for the following genome wide capture array analysis. The selected samples all originate from Peru and are dated 1285 to 750 AD, see Table 8.1 for an overview.

Name	Location	Dated
54	Peru	AD 750 - 1150
58	Peru	AD 900 - 1285
64	Peru	AD 900 - 1200

Tab. 8.1: Site location and approximate age of the selected samples.

Before sequencing a partial amount of the samples have been UDG (U) treated. Enrichment for partially unknown aDNA strains requires a specially designed capture array. The design and the subsequent analysis will be described in the following section.

Design

The design of the array needs to mirror most of the specificity and variation of multiple strains at the same time. We selected M. africanum (1 strains), M. bovis (4 strains), M. canettii (1 strains), and M. tuberculosis (14 strains) to be included in the final array design. The detailed list including accession numbers is represented in Table C.19. The design of the chip is based on a *degenerated* consensus sequence of the included strains representing each observed variation. Consensus sequence creation is based on a multiple sequence alignment (MSA) which was done using Mauve [125]. The design target was to stay below 2 million probes (i.e. 2 arrays).

Methods: All strains which were selected to be included in the final design (Table C.19) were aligned using Mauve. Genomes of the included KZN strains contain a major rearrangement compared to H37Rv. To optimize the consensus we rearranged the order of the blocks to match the order of the reference H37Rv strain. For this the individual blocks were identified using SPRING [126]. The sequences were then rearranged based on this information. From the MSA a degenerated consensus was created to represent the complete variation at all positions. Each position was encoded using a one letter code (e.g. M: A+C) depending on all nucleotides in the MSA at the specific location. Artificial breaks, such as gaps in the alignment spanning more than 4 bps or brakes introduced during genome rearrangement, were included using 60 bp up and downstream on the original sequence. Overlapping fragments were merged to minimize introduced redundancy. Highly degenerated regions with more than 10 degenerated sites per 60 bps were identified and replaced by the original H37Rv sequence, as those may be artificially induced by the alignment process.

In addition to the main degenerated consensus a separate consensus was created using two M. avium strains: M. avium 104 (NC_008595.1) and M. avium subsp. paratuberculosis K-10 (NC_002944.2). The K-10 strain was rearranged to match the block order of the longer 104 strain.

Array probes with length 60 were designed from the main consensus sequences. Duplicate probes were removed and remaining free space on the array was filled with M. avium and M. kansasii probes. For this, unique regions in M. avium and M. kansasii (NZ_CM000636.3) not represented in the other genomes were calculated.

Results: The initial degenerated main consensus is 5,037,910 bps long with an average coverage of 19.89 during consensus creation. Additional 299 blocks, induced by only partially shared sequence fragments, span 196,111 bps with 2.86 fold coverage. Gaps introduced additionally 23,738 blocks. The final main consensus covers 7,887,472 bps including all flanking positions from blocks and rearrangements. Table C.20 gives an overview of the composition of the degenerated consensus sequence. The degenerated M. avium consensus is 5,922,143 bps long. The consensus contains 883 gaps spanning more than 4 bps, see Table C.21 for base composition.

Discussion During the design phase various obstacles were encountered especially when creating a MSA of the sequences. The initial design considered more reference strains to be included (e.g. M. avium and M. kansasii) on the array. In this early phase we observed that Mauve has potential problems when aligning strains which are too distantly related. As a result we used SPRING for detecting rearrangements in the sequences and reordered them prior to aligning.

Both additional strains showed heavy structural variance compared to H37Rv. For M. avium SPRING calculation showed 78 blocks spanning about 91% of the genome which needed to be rearranged. The genome of M. kansasii (NZ_CM000636.3) contains 138 blocks which were identified. Adding the additional diversity by including those strains we observed that the length of the MSA and therefore the created consensus sequence was to large to fit onto two arrays.

Up to some part the diversity of the strains may be responsible for the observed increase in complexity, but we assume that even with the rearranged strains the MSA is not optimal, as it suddenly gets stretched artificially. Removing M. avium and M. kansasii from the alignment enabled Mauve to create a MSA which was considered valid.

The replacement of highly degenerated regions in the consensus sequence was also based on the assumption that those regions are more likely an artifact of the MSA than representing real genomic diversity.

Including 120 bps of the original sequence if the alignment contains a longer (>4 bps) gap was done to ensure that the gaped region of the sequence is represented as continuous stretch on the array to enable capturing of it. The additional probes on the array covering partially M. avium and M. kansasii ensures that potential diversity represented only in those strains should be potentially captured as well. The final array design represents most of the complexity of the M. tuberculosis complex, while it tries to minimize enrichment bias towards single specific strains.

8.2.4 Capture Array Analysis

Using the previously designed enrichment array the selected samples can be further processed. We again included line and extraction blanks as controls. As an additional check we also included an ancient sample which pathologically showed no signs of a TB infection.

Methods Samples were enriched using the custom build bait array and sequenced on an Illumina HiSeq 2000 using a paired-end protocol. QC was performed accordingly to the previous chapters. Samples were initially mapped against M. avium, M. kansasii and M. tuberculosis strains. Potential PCR duplicates were removed using Picard [104]. Coverage plots and coverage histogram were calculated using BEDTools [106]. DNA damage plots were calculated for all samples.

The three UDG treated samples (54.U, 58.U, 64.U) were additionally mapped against M. canettii, M. africanum, and M. bovis. Duplicate removal was performed, coverage information and damage plots were generated accordingly.

Based on the mapping against M. canettii, M. africanum, M. bovis, and M. tuberculosis SNP calling was performed using the UnifiedGenotyper [87]. SNPs were filtered using a minimum coverage of 5 and a minimum quality of 30.

Using the SNP information we calculated a total of 24 consensus sequences: two consensus sequences for each combination of samples (3) and references (4). The first consensus contains Ns at each position which is not covered by reads or did fail quality checks, this sequence will be referred to as *consensus sequence*. The second consensus contains the original base of the reference sequence at positions previously containing Ns, this will be referred to as *overlay-consensus sequence*. The whole pipeline up to this point was integrated into galaxy as single workflow.

Based on the overlay-consensus sequences a MSA was created and SNP calling performed using Mauve. SNP positions of the overlay MSA were replaced by the initial consensus sequences to ensure that originally not covered SNPs will be marked by Ns and filtered.

To evaluate the MSA results different pairwise alignments were calculated: For each reference sequence we aligned the reference and the corresponding overlay-consensus sequence. This was only performed for sample 54.U. For each reference two alignments were generated: one using the default options and a second time we instructed Mauve to assume collinear genomes. These options were also used to generate additional MSAs using the four TB complex reference sequences.

The three samples were additionally assembled using SOAPdenovo [48, 49] using different k-mer sizes. The longest contigs found for each sample were exemplarily compared against the NCBI-NT database using BLASTN.

Results Table C.22 gives an overview of the total number of reads per sample. Average read length is between 50-65 bps with a standard deviation of 10-25 depending on the sample. The maximum number of mapped reads to M. avium and M. kansasii is 3% whereas the mapping to H37Rv achieved 20-30%. An exception to this are extraction and line blanks as well as the TB negative control: Mapping was usually below 1% regardless of the reference sequence.

Coverage calculations show that samples do not cover 85% to 99% of the M. avium and M. kansasii genomes. In comparison to that only 2% to 16% of H37Rv is not covered by the samples. Overall the not UDG treated samples have a lower count of mappable reads and less overall genome coverage. Again negative controls fail to cover most of the reference sequences.

For the second mapping against the four TB complex references about 18% (64.U), 26% (58.U), and 30% (54.U) of reads were mapped regardless of the reference sequence used. All reference genomes were covered between 97% and 99%. Table 8.2 displays the average fold coverage.

Sample	H37Rv	M. africanum	M. bovis	M. canettii
54.U	34.8	35.2	34.9	33.2
58.U	22.3	22.6	22.4	21.3
64.U	26.5	26.7	26.5	25.3

Tab. 8.2: Average fold coverage of samples against different references.

After consensus generation SNP calling was performed: Depending on the sample approximately 16,000-17,000 raw SNPs were called for M. canettii as reference and 1,200-2,200 raw SNPs for the other three reference sequences. Filtering SNPs removed 10% of all SNPs on average.

The initial overlay-consensus MSA was generated using Mauve. Mauve called 17,122 SNPs and approximately 4 million gapped positions, see Table 8.3 for details. After replacing originally not covered positions a total of 3,913 SNPs were observed.

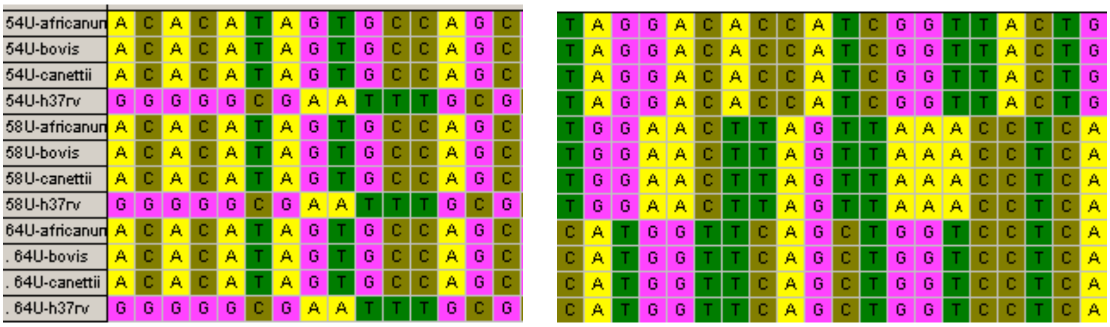
	Default		Collinear	
Sequences	Gaps	SNPs	Gaps	SNPs
Overlay-consensus sequences	4,792,067	17,122	10,827,919	14,761
TB complex references	1,374,252	29,249	1,537,104	28,401

Tab. 8.3: Results for the two MSAs using the overlay-consensus sequences and TB complex reference sequences. Gap positions and SNPs in the alignment created by Mauve using either default options and assuming collinear genomes.

Visual analysis of SNP positions indicated potential problems with the MSA: According to the alignment different bases get called for the same sample at the same position if a different reference sequence is used, see Figure 8.5 for an example.

Table 8.3 also displays the results for the MSA using the TB complex reference sequences. The alignment is overall less gapped and more SNPs are identified.

For the pairwise alignments gaps and SNPs were determined. Depending on the reference and the options used, huge gaps are introduced into the alignment. It is interesting that only for M. bovis and M. tuberculosis the alignment and SNPs stayed constant or only varied sightly. Compared to M. bovis the alignment using M. tuberculosis contains a huge number of gapped positions which may indicate problems with the sequence alignment. Table 8.4 displays the results.



```
54U-africanun A C A C A T A G T G C C A G C
54U-bovis     A C A C A T A G T G C C A G C
54U-canettii  A C A C A T A G T G C C A G C
54U-h37rv     G G G G G C G A A T T T G C G
58U-africanun A C A C A T A G T G C C A G C
58U-bovis     A C A C A T A G T G C C A G C
58U-canettii  A C A C A T A G T G C C A G C
58U-h37rv     G G G G G C G A A T T T G C G
64U-africanun A C A C A T A G T G C C A G C
. 64U-bovis   A C A C A T A G T G C C A G C
. 64U-canettii A C A C A T A G T G C C A G C
. 64U-h37rv   G G G G G C G A A T T T G C G
```

```
T A G G A C A C C A T C G G T T A C T G
T A G G A C A C C A T C G G T T A C T G
T A G G A C A C C A T C G G T T A C T G
T A G G A C A C C A T C G G T T A C T G
T G G A A C T T A G T T A A A C C T C A
T G G A A C T T A G T T A A A C C T C A
T G G A A C T T A G T T A A A C C T C A
T G G A A C T T A G T T A A A C C T C A
C A T G G T T C A G C T G G T C C T C A
C A T G G T T C A G C T G G T C C T C A
C A T G G T T C A G C T G G T C C T C A
C A T G G T T C A G C T G G T C C T C A
```

Fig. 8.5: SNPs based on the MSA of the consensus sequences. Left: Potential misalignment as the reference H37Rv has a different base compared to the other sequences. Right: Assumed correct alignment. The samples are conclusive within themselves regardless of the used reference sequence.

	Default		Collinear	
Reference	Gaps	SNPs	Gaps	SNPs
M. africanum	2,453,232	1,391	0	1,854
M. bovis	0	1,911	0	1,911
M. canettii	53,428	15,531	359,993	14,744
M. tuberculosis	1,277,589	1,928	1,277,577	1,926

Tab. 8.4: Result of the pairwise sequence alignments for sample 54.U. Gap positions and SNPs in the alignment created by Mauve using either default options or assuming collinear genomes.

Additional to the number of gapped positions, the pairwise alignments indicate potential problems with Mauve: Mauve identifies a region as specific for each sequence (see white block in Figure 8.6) although manual inspection shows that both sequences are nearly identical. Additionally Mauve identifies variation in the genomic structure (differently colored blocks in Figure 8.6) for the Overlay-H37Rv vs. Reference-H37Rv alignment which is unlikely.

For the assembly different results were observed depending on k-mer size and sample. The best N50 is 1,043 (sample 54U, k-mer size 83) and the longest contig 23,750 bps long (sample 58U, k-mer size 49). The database search using BLASTN did identify the longest contigs to be only partly covered by the database result, see Table 8.5.

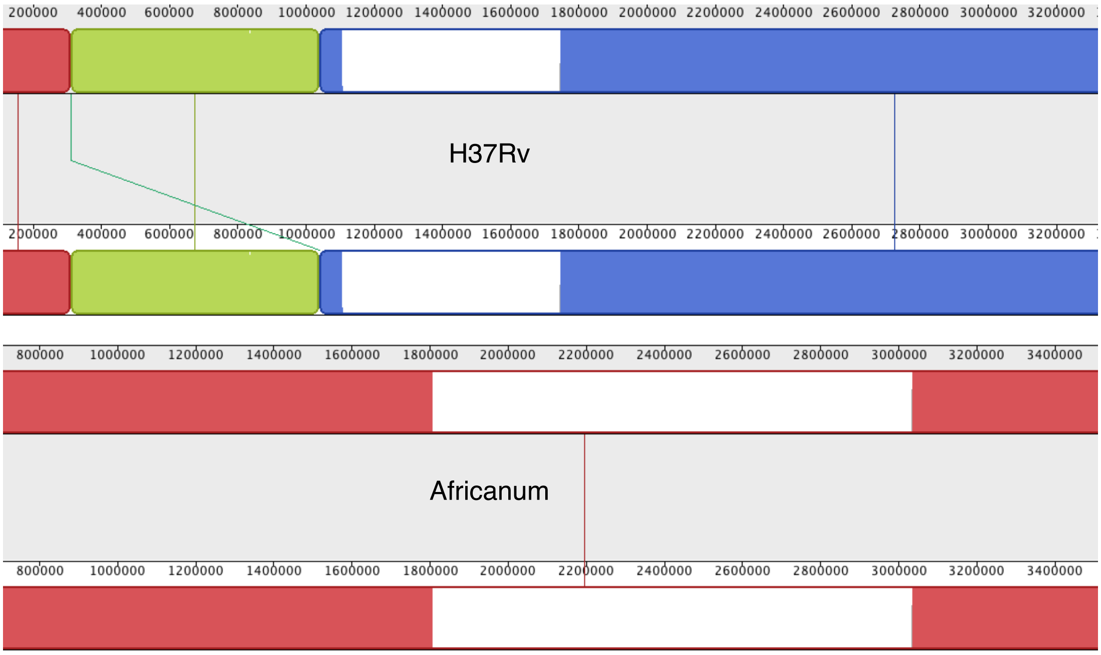


Fig. 8.6: Mauve alignment for two different reference sequences: Overlay-H37Rv vs. Reference-H37Rv and Overlay-M. africanum vs. Reference-M. africanum. In H37Rv Mauve indicates a different genomic structure (different colors). In both alignments Mauve assumes that a section is completely different and unique for the specific genome (white block). However the regions identified by Mauve are nearly similar which indicates a potential problem with the alignments.

Sample	Hit Description	% of Query Covered	E-Value
54.U	Clostridium difficile BI9	94%	0.0
58.U	Alkaliphilus metalliredigens QYMF	73%	0.0
64.U	Acinetobacter baumannii AB307-0294	6%	2e-56

Tab. 8.5: Best BLAST Hit for the longest contig generated during assembly.

Discussion Average read length and DNA damage pattern are again used for preliminary authentication of the aDNA. Additionally the damage pattern of UDG treated samples are not as distinctive as of the not UDG treated samples. This also supports the authenticity of the samples, as DNA damage is partially repaired by the UDG treatment.

The initial mapping versus the M. avium, M. kansasii, and M. tuberculosis strains strains ensured that no complete outgroup was captured. The higher number of mappable reads for the UDG treated samples is due to the decreased DNA damage. The negative controls do not suggest that there are major problems during extraction and preparation of the samples.

The second mapping against the four TB complex reference sequences was done to ensure that the analysis is not substantially biased by mapping exclusively to H37Rv. Based on the number of mapped reads, total coverage and average coverage we are not able to identify an exclusive reference sequence for the subsequent analysis.

The SNP calling for the TB complex mappings resulted in a relatively high amount of SNPs for M. canettii in comparison to the other three reference sequences. A main problem for interpreting the numbers is that SNP numbers vary widely between publications as only few publications perform genome wide SNP analysis. Most studies perform the phylogeny analysis based on SNPs found in single genes (e.g. katG, gyrA, gyrB, hsp65, rpoB, and sodA). Based on this the estimated numbers of SNPs for the full genome are much lower as encountered here.

Two publications [127, 128] investigating full genome SNPs describe more than 14,000 SNPs between two modern strains of M. canettii and H37rv, as well as 2,437 SNPs between the genomes of H37Rv and M. bovis strain AF2122/97. The published numbers suggests that our mapping of the ancient strains to the modern references and the SNP calling is within the possible ranges and not obviously flawed.

During creation of the MSAs we encountered various challenges which we assume are mainly introduced by Mauve. The initial minor problem that Mauve is unable to handle Ns in sequences correctly was solved by using the overlay-consensus sequences for the MSA generation. After the initial MSA generation we replaced not covered positions in the MSA with Ns.

Unfortunately we further discovered potential serious problems with the MSA of the overlay-consensus sequences which is generated by Mauve: The MSA resulted in SNP positions which differ depending on which reference sequences used during mapping (Figure 8.5 left). This should not happen as we expect reads to create a consistent consensus regardless of the reference sequence used. Positions not covered by any read are substituted by N and therefore identified. To clarify: We expect to only see variation between the different samples and not between the different reference sequences (Figure 8.5 right).

The comparison of the alignment using the TB complex reference indicates that the MSA is artificially stretched (see Table 8.3). The pairwise alignment confirmed that Mauve does not align the sequences correctly as it marks whole regions as separate blocks, although sequences are identical. For the pairwise alignment the region is usually half the size of the gapped positions.

For the pairwise alignment of Overlay-H37Rv vs. Reference-H37Rv Mauve indicates structural differences which are highly unlikely based on the process we generated the consensus sequences. It would be interesting to assemble the aligned reads to see if structural changes occurred.

We assume that the observed behavior of Mauve to identify similar blocks as specific for the separate genomes by mistake introduces misalignments when using a higher number of sequences. To pinpoint the problem it is necessary to validate the genome mapping and consensus creation and later inspect multiple alignment positions by hand. At this point the exact trigger is unknown and it is therefore hard to circumvent the problem.

We have to expect that the MSA is invalid which puts the following analysis on hold. Depending on the mechanics one has to consider to replace Mauve with an alternative approach to align the genomes.

The results for the assembly are only preliminary as we have only performed initial tests for different k-mer sizes. Without further tests it remains unclear if the reads were assembled correctly. The still existing DNA damage can interfere with the assembly process which results in the low N50 numbers. The single BLAST results for the longest contig should not be seen representative for the whole dataset. Especially the result for sample 64.U with only 6% of the query covered rises the question if the contig is an artificial construct. Determining the validity of the assembly and the for the match to Clostridium difficile in Sample 54.U. is also of interest. Spores of Clostridium difficile have been detected in hospital air [129] which could have theoretically contaminated the sample during processing depending on the wet-lab's location.

8.2.5 Final Conclusion

In this part we introduced methods for ancient DNA research which enables researchers unique views into the past of evolution. In comparison to modern DNA, aDNA puts unique challenges on the studies: Samples usually only yield low amounts of DNA which is additionally damaged and can be easily contaminated with modern DNA.

Here we described a project using samples from subjects which are suspected to have been infected with M. tuberculosis. The study was done in three different parts in which we encountered various challenges. During the initial screening using a whole genome shotgun approach we observed artifacts which may have been the result of the low amount of DNA and problems with the sequencing chemistry. Screening of samples which have been previously enriched for four different gene regions successfully identified three samples for the continuing analysis. All three samples originated from sites with high altitude. The probably colder climate may have been beneficial for DNA preservation compared to other samples. To overcome the limitation of low amounts of starting DNA we designed a special genome wide enrichment chip. This was done by creating a degenerated consensus of twenty known TB complex strains. During the final analysis of the enriched and sequenced samples various problems were observed with the MSA which could not be explained or solved until now. The key element of the analysis, the multiple sequence alignment generated by Mauve, is suspected to be flawed. In pairwise alignments Mauve identifies clearly identical regions as specific for the separate genomes, creating huge gaps.

We assume that this behavior leads to misalignments when aligning more sequences. Continuing with the downstream analysis does not make sense at this point, as a valid MSA is crucial for identifying SNPs and the subsequent analysis. Overall this part expresses the full spectrum of what can happen during an analysis, including potential wet-lab artifacts and various software flaws.

The continuing analysis outside this work has to establish a way to circumvent the occuring problems, before being able to establish a phylogentic analysis of the samples.

Part IV

Conclusion and perspectives

Chapter 9

Conclusion and Perspectives

Today's modern genomics highly benefits from new sequencing methods, allowing scientist to sequence the genomic content faster, more efficiently and cheaper than ever before. Fields such as metagenomics and ancient DNA research highly benefit from this development. This work introduced and discussed mainly computational challenges for both disciplines.

As experiments and lab equipment do not produce perfect data, it is important to validate and preprocess the data in a quality control step prior to the initial analysis. Chapter 4 presented and discussed various established methods.

The mapping of reads to species or functions is the most important step during a metagenomic analysis. In this thesis we have divided this step into two separate parts: the classification of reads using a reference database and assigning reads based on the matched reference sequences. For the initial classification we have shown that current methods are often slow and / or do not provide any kind of sequence alignment, which is required to establish a thorough analysis. In this context we presented a new hybrid method enabling faster analysis of metagenomic datasets. The main idea is that the search space for the BLASTX run is reduced by classifying reads on a higher level in a previous step. A simulation study shows that the approach is faster and achieves a higher sensitivity than a BLASTX run.

The part of assigning reads to the taxonomy or functional content is highly dependent on the used database. This includes database coverage as well as identifiers provided by the database which is used for mapping. To asses the impact of database accuracy on the assignment we analyzed overall mapping coverage of various database using the NCBI taxonomy. Functional mapping was evaluated by the number of reads which could be mapped to valid SEED functional roles or KEGG KOs.

Results show that mapping approaches using a text based description of sequences are sometimes unable to map the reference correctly because of missing descriptions or typos. For functional analyses we have seen that highly curated identifiers like RefSeq can be found less in newer versions of the database, which results in not mappable references.

We therefore described an improved and generalized mapping approach using a GI based mapping. Results for taxonomic and functional mapping improved using the new approach. Evaluation with an in vitro simulated dataset showed that the GI based approach is able to reduce false positive matches of the text based mapper which were not detected in the prior analysis. Functional mapping accuracy was also improved, but compared to taxonomic assignment the assignment rate was low. The main reason for this is that still only few genomes are functionally annotated. All new mapping methods are already implemented in MEGAN for general use.

Interdisciplinary approaches with many scientists involved strengthens the need for an integrated approach to analyze and share data easily. A few solutions already exist for the integrated analysis of metagenomes. Those solutions are mostly off-site and don't allow a local deployment which would be beneficial to ensure data security, safety and the use of local resources. To ease and automate analysis in our projects the Galaxy workflow management system was set up, modified and evaluated. The system was used in different projects and performed well in data analysis, management and sharing. Exemplary workflows as well as the implemented tools have been described throughout this thesis.

The last part of this work described a current project using ancient DNA. Extinct specimen yield high value information concerning evolutionary and distribution patterns. Especially knowledge of the evolution of pathogens may be beneficial for understanding diseases and the development of new treatments and vaccines. Here were analyzed multiple ancient Mycobacterium tuberculosis samples. The samples were initially screened using a whole genome shotgun approach. Selected samples were specifically enriched using a custom designed hybridization array and then sequenced again. Consensus sequences were generated for three different ancient strains and an analysis for single nucleotide polymorphisms was performed. Quality of the consensus and downstream analysis need to be evaluated further as there are discrepancies depending on the reference sequence used.

Current developments in metagenomics show promising results in the area of fast assignment methods, but further work needs to be done as data volumes tend to increase. Also the development of integrated analysis pipelines enabling interdisciplinary groups to easily share, access and analyze data should be further pursued.

Ancient DNA research will clearly benefit from technical progress. Literally uncountable samples may prove or confute current theories of evolution. From the technical point of view the authentication of aDNA is a topic which needs to be addressed in the future. Additionally untargeted enrichment of aDNA samples is a high value ambition as this will enable to capture the full spectrum of an ancient sample.

Part V

Appendix

Appendix A

Contributions

Hybrid Approach

Nico Weber (NW), Dominik Damerow (DD) and Daniel Huson (DH) contributed to this project. NW conceived the project. DD implemented and tested the scripts. DD and NW performed the analysis. NW, DD and DH contributed to the discussion. Parts of this section were presented at the *1st Thünen Symposium on Soil Metagenomics in Braunschweig 2010*

Database and Assignment Accuracy

Nico Weber (NW), Daniel Huson (DH), and Sonja Hägele (SH) contributed to this project.
NW designed the original study. SH and NW performed the initial experiments to asses the usability of the in vitro simulated metagenomic dataset. NW performed the database analysis and generated the new mapping files. NW and DH suggested improvements to MEGAN and DH implemented them. Final analysis of the in vitro simulated metagenomic dataset was done by NW.

Workflows in Galaxy

Nico Weber (NW) and Magdalena Feldhahn (MF) contributed to this project. NW designed and implemented the Galaxy server. MF integrated, tested the tools and designed and implemented the immoninformatic workflows. NW implemented and tested tools and workflows presented in this thesis. NW and MF contributed to the testing of the server.

aDNA Research

Nico Weber (NW), Johannes Krause (JK), Kirsten Bos (KB), Günther Jäger (GJ), Alexander Herbig (AH), and Kay Nieselt (KN) contributed to this project. JK and KB conceived the project. NW and KB designed the capture array. NW performed the bioinformatic analysis. JK performed the visual inspection of alignments and the phylogenetic analysis. NW, JK, KB, GJ, AH and KN contributed to the discussion.

Appendix B

Publications

Published Manuscripts

- Huson DH, Mitra S, Ruscheweyh HJ, **Weber N**, Schuster SC. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 2011 Sep;21(9):1552-60. [13]
- Daniel Huson and **Nico Weber**. Analysis of Soil Metagenomes using the Metagenome Analyzer Megan. In: Nannipieri P (Editor), Pietramellara G (Editor), Renella G (Editor), *Omics in Soil Science*. In press.
- Daniel Huson and **Nico Weber**. Microbial community analysis using Megan. In: Delong EF (Editor), *Methods in Enzymology. Volume 531: Microbial Metagenomics, Metatranscriptomics, and Metaproteomics, 1st Edition*. In press.

Appendix C

Supplements

C.1 Hybrid Approach

Dataset	Technology	NBC	BLASTX	BLASTX (NBC)	Speedup (Combined)
HC	454	14	268.5	5.75	12.9
HC	Solexa	12.5	50	2.5	3.3
MC	454	14	265.5	7.5	112.3
MC	Solexa	12.5	68	6.25	3.6
LC	454	14	259	8	10.36
LC	Solexa	12.5	66.5	6.25	3.5

Tab. C.1: Runtimes (in CPU hours) for the NBC classifier and BLASTX on simulated data. *HC*, *MC*, and *LC* stand for high, medium, and low complexity respectively. *Technology* describes the error model of the simulated data.

Dataset	Technology	Phymm	BLASTX	BLASTX (Phymm)	Speedup (Combined)
HC	454	6.75	268.5	16.5	11.5
HC	Solexa	2.5	50	7	5.2
MC	454	7	265.5	13.5	12.9
MC	Solexa	5	68	6	6.1
LC	454	7	259	16	11.2
LC	Solexa	4.75	66.5	5.5	6.4

Tab. C.2: Runtimes (in CPU hours) for the Phymm classifier and BLASTX on simulated data. *HC*, *MC*, and *LC* stand for high, medium, and low complexity respectively. *Technology* describes the error model of the simulated data.

C.2 Database Accuracy

Database	Sequences	Bases	IDs	Hypothetical Entries	Hypo. Bases
2013	22,540,640	7,750,666,634	63,671,296	17,970,078	7.6%
2011	16,118,048	5,541,108,926	36,272,663	10,196,207	8.0%
2009	9,987,577	3,407,368,633	22,263,912	6,308,198	5.02%

Tab. C.3: Overview of the composition of NCBI-NR. Columns describe the number of *Sequences*, *Bases*, *Identifiers (IDs)*, and *Hypothetical Entries* with the percentage of *Hypothetical bases*.

	Identifier			Bases covered	
Database	Max	One	Two	by one ID	by two IDs
2013	1,6342	32.7%	56.6%	2,534,829,532	4,387,131,907
2011	9,504	36.8%	53.1%	2,041,652,088	2,944,924,392
2009	6,511	31.0%	58.4%	1,056,622,585	1,992,533,027

Tab. C.4: Overview of the NCBI-NR identifier spread. Columns describe the maximum number of IDs a single entry has (*Max*), the percent of sequences having one or two ID(s) (*One*, *Two*) as well as the number of bases with one or two IDs (*by one ID*),*by two IDs*).

Rank	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	265	131	6,616	287	7,299
Bacteria	3,386	2,304	248,124	21,615	275,429
Eukaryota	20,141	60,460	487,018	23,194	590,809
Fungi	1,756	4,132	80,562	2407	88,853
Metazoa	14,162	39,570	266,838	10,333	330,903
Viridiplantae	2,441	14,183	113,577	8,638	138,839
Viruses	610	387	10,336	89,873	101,206
All taxa	24,432	63,290	759,350	135,004	982,072

Tab. C.5: NCBI Taxonomy statistics downloaded on the 30.1.13. The columns display the number of taxa at specific taxonomic ranks. Informal names are included in this listing, This overview also shows the statistics of the taxonomy used by MEGAN for this study.

Rank	Higher taxa	Genus	Species	Lower taxa	Total
Archaea	265	131	467	0	863
Bacteria	3,386	2,304	11,348	774	17812
Eukaryota	20,141	60,460	254,734	19,342	354,677
Fungi	1,756	4,132	25,765	995	32,648
Metazoa	14,162	39,570	119,452	9,693	182,877
Viridiplantae	2,441	14,183	101,254	8,422	126,300
Viruses	610	387	1,985	0	2,982
All taxa	24,432	63,290	268,567	20,116	376,405

Tab. C.6: NCBI Taxonomy statistics downloaded on the 30.1.13. The columns display the number of taxa at specific taxonomic ranks. Informal names are excluded from this listing for comparison.

Database	Bacteria	Eukaryota	Archaea	Viruses
2013	34,527 (12,80%)	219,996 (43,12%)	1,214 (17,59%)	62,159 (62,03%)
2010	27,592 (10,23%)	225,089 (44,12%)	1,060 (15,36%)	51,843 (51,73%)
2009	19,531 (7,24%)	121,291 (23,77%)	793 (11,49%)	30,056 (29,99%)

Tab. C.7: Number of leaves covered by the NCBI-NR database using the text-based parser. Results are displayed for kingdoms *Bacteria*, *Eukaryota*, *Archaea* and *Viruses*.

Database	Bacteria	Eukaryota	Archaea	Viruses
2013	39,913 (14.49%)	280,487 (47.48%)	1,628 (22.30%)	64,309 (63.54%)
2010	32,071 (11.64%)	287,885 (48.73%)	1,424 (19.51%)	49,797 (49.20%)
2009	22,859 (8.30%)	172,799 (29.25%)	1,065 (14.59%)	31,831 (31.45%)

Tab. C.8: Number of total taxa covered by the NCBI-NR database using the text-based parser. The selected subtrees are *Bacteria*, *Eukaryota*, *Archaea* and *Viruses*.

Database	Total	Not Assigned	Assigned	Assignment Efficiency
2013	22,540,640	85,235 (0.38%)	22,455,405	99.62%
2011	16,118,048	104,608 (0.65%)	16,013,440	99.35%
2009	9,987,577	162,453 (1.63%)	9,825,124	98.37%

Tab. C.9: Number of total, assigned and not assigned sequences of the NCBI-NR database using the default name parser in MEGAN.

	SEED		KEGG	
Database	Assigned	Not Assigned	Assigned	Not Assigned
2013	689,557 (2.70%)	24,851,083 (97.30%)	1,726,223 (6.76%)	20,814,417 (81.50%)
2011	694,447 (4.31%)	15,423,601 (95.69%)	1,737,791 (10.78%)	14,380,257 (89.22%)
2009	691,824 (6.93%)	9,295,753 (93.07%)	1,530,234 (15.32%)	8,457,343 (84.68%)

Tab. C.10: Total and relative number of *Assigned* and *Not Assigned* sequences of the NCBI-NR database to valid SEED and KEGG mappings. Mapping was performed using the RefSeq based approach.

Database	Total	Not Assigned	Assigned
2013	22,540,640	19,629 (0.09%)	22,521,011 (99.91%)
2011	16,118,048	402,126 (2.49%)	15,715,922 (97.51%)
2009	9,987,577	830,728 (8.32%)	9,156,849 (91.68%)

Tab. C.11: Number of total, assigned and not assigned sequences from the NCBI-NR database using the GI parser in MEGAN.

Database	Bacteria	Eukaryota	Archaea	Viruses
2013	32,010 (11.87%)	200,469 (39.29%)	1,169 (16.93%)	57,877 (57.76%)
2010	25,925 (9.61%)	168,421 (33.01%)	1,026 (14.86%)	47,431 (47.33%)
2009	18,934 (7.02%)	121,395 (23.79%)	792 (11.47%)	30,778 (30.71%)

Tab. C.12: Number of total leaves of a specific taxonomic rank *Bacteria*, *Eukaryota*, *Archaea* and *Viruses* and percentage covered by the NCBI-NR Database using the GI based approach

Database	Bacteria	Eukaryota	Archaea	Viruses
2013	37,416 (13.58%)	269,148 (45.56%)	1,579 (21.63%)	60,520 (59.80%)
2010	30,788 (11.18%)	231,039 (39.11%)	,1418 (19.43%)	49,991 (49.40%)
2009	22,926 (8.32%)	173,803 (29.42%)	1,117 (15.30%)	33,084 (32.69%)

Tab. C.13: Number of total taxa of the subtree of a specific taxonomic rank *Bacteria*, *Eukaryota*, *Archaea* and *Viruses* and percentage covered by the NCBI-NR Database using GI based approach.

Database	Assigned Sequences	Not Assigned
2013	3,148,802 (13.97%)	19,391,838 (86.03%)
2011	2,506168 (15.55%)	13,611,880 (84.45%)
2009	1,723,434 (17.26%)	8,264,143 (82.74%)

Tab. C.14: Number of sequences from the NCBI-NR database mapped to KEGG KOs using the GI based mapping method.

Database	Assigned Sequences	Not Assigned
2013	602,036 (2.67%)	21,938,604 (97.33%)
2011	605,383 (3.76%)	15,512,665 (96.24%)
2009	607,837 (6.09%)	9,379,740 (93.91%)

Tab. C.15: Number of sequences from the NCBI-NR database mapped to SEED functional roles using the GI based mapping method.

	Text Based		GI Based	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
Big	402,736 (84.84%)	71,933 (15.15%)	402,771 (84.85%)	71,925 (15.15%)
Medium	15,874 (85.15%)	2,768 (14.85%)	15,874 (85.15%)	2,768 (14.85%)
Small	7 (12.28%)	50 (87.72%)	7 (12.28%)	50 (87.72%)

Tab. C.16: Comparison of assignment efficiency to the NCBI taxonomy of the text-based method and the GI-based mapping approach.*Assigned* is the number of assigned reads whereas *Not Assigned* represents the number of not assigned reads.

	RefSeq Based		GI Based	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
Big	212,407 (44.75%)	262,289 (55.25%)	221,741 (46.71%)	252,955 (53.29%)
Medium	8,461 (45.39%)	10,181 (54.61%)	8,811 (47.26%)	9,831 (52.74%)
Small	4 (7.02%)	53 (92.98%)	5 (8.77%)	52 (91.23%)

Tab. C.17: Comparison of assignment efficiency to KEGG KOs of the RefSeq based method and the GI based mapping approach. *Assigned* is the number of assigned reads whereas *Not Assigned* represents the number of not assigned reads.

	RefSeq Based		GI Based	
Dataset	Assigned	Not Assigned	Assigned	Not Assigned
Big	156,012 (32.87%)	318,684 (67.14%)	176,125 (37.11%)	298,517 (62.89%)
Medium	6,183 (33.17%)	12,459 (66.83%)	6,983 (37.46%)	11,659 (62.54%)
Small	2 (3.51%)	55 (96.49%)	2 (3.51%)	55 (96.49%)

Tab. C.18: Comparison of assignment efficiency to SEED functional roles of the RefSeq-based method and the GI-based mapping approach.*Assigned* is the number of assigned reads whereas *Not Assigned* represents the number of not assigned reads.

C.3 aDNA: Array Design

Tab. C.19: Reference genomes included in the main array design.

GI	RefSeq	Description
339630083	NC_015758.1	Mycobacterium africanum GM041182 chromosome, complete genome
31791177	NC_002945.3	Mycobacterium bovis AF2122/97 chromosome, complete genome
121635883	NC_008769.1	Mycobacterium bovis BCG str. Pasteur 1173P2 chromosome, complete genome
224988383	NC_012207.1	Mycobacterium bovis BCG str. Tokyo 172, complete genome
378769743	NC_016804.1	Mycobacterium bovis BCG str. Mexico chromosome, complete genome
340625033	NC_015848.1	Mycobacterium canettii CIPT 140010059, complete genome
375294201	NC_016768.1	Mycobacterium tuberculosis KZN 4207 chromosome, complete genome
253796915	NC_012943.1	Mycobacterium tuberculosis KZN 1435, complete genome
148821191	NC_009565.1	Mycobacterium tuberculosis F11, complete genome
297749916	NZ_CM000789.2	Mycobacterium tuberculosis KZN R506 chromosome, whole genome shotgun
306478687	NZ_CM000788.2	Mycobacterium tuberculosis KZN V2475 chromosome, whole genome shotgun
383305933	NC_017026.1	Mycobacterium tuberculosis RGTB327 chromosome, complete genome
386003090	NC_017528.1	Mycobacterium tuberculosis RGTB423 chromosome, complete genome
385993125	NC_017523.1	Mycobacterium tuberculosis CCDC5079 chromosome, complete genome
385989534	NC_017522.1	Mycobacterium tuberculosis CCDC5180 chromosome, complete genome
50953765	NC_002755.2	Mycobacterium tuberculosis CDC1551 chromosome, complete genome
385996772	NC_017524.1	Mycobacterium tuberculosis CTRI-2 chromosome, complete genome
148659757	NC_009525.1	Mycobacterium tuberculosis H37Ra, complete genome
57116681	NC_000962.2	Mycobacterium tuberculosis H37Rv chromosome, complete genome
297595741	NZ_CM000787.2	Mycobacterium tuberculosis KZN 4207 chromosome, whole genome shotgun

Code	Original	Count	Percent
A	A	852,060	16.9
C	C	1,642,074	32.5
G	G	1,631,011	32.3
T	T	850,750	16.8
M	A+C	4,650	00.0
R	A+G	13,176	00.2
W	A+T	1,647	00.0
Y	C+T	12,959	00.0
S	G+C	8,696	00.1
K	G+T	4,622	00.0
H	A+T+C	87	00.0
B	G+T+C	139	00.0
D	G+A+T	76	00.0
V	G+A+C	168	00.0
N	A+C+G+T	15,795	00.3

Tab. C.20: Composition statistics of the main degenerated consensus sequence used for building the enrichment arrays. *Code* and *Original* represent the code in the consensus or the nucleotides encountered in the multiple sequence alignment. *Count* and *Percent* reflect the absolute and relative occurrence.

Code	Original	Count	Percent
A	A	894,233	15.4
C	C	1,975,501	34.0
G	G	1,981,726	34.1
T	T	899,551	15.5
M	A+C	3,580	00.0
R	A+G	14,271	00.2
W	A+T	990	00.0
Y	C+T	14,456	00.2
S	G+C	13,490	00.2
K	G+T	3,684	00.0
H	A+T+C	0	00.0
B	G+T+C	0	00.0
D	G+A+T	0	00.0
V	G+A+C	0	00.0
N	A+C+G+T	0	00.0

Tab. C.21: Composition statistics of the degenerated M. avium consensus sequence. *Code* and *Original* represent the code in the consensus or the nucleotides encountered in the multiple sequence alignment. *Count* and *Percent* reflect the absolute and relative occurrence.

C.4 Analysis

Sample Name	Right Reads	Left Reads	Total Number of Reads
54U	103,846,833	103,091,354	206,938,187
54N	9,640,140	9,559,346	19,199,486
58U	25,727,133	25,615,876	51,343,009
58N	5,149,171	5,058,027	10,207,198
64U	66,215,214	65,803,949	132,019,163
64N	3,112,331	3,042,352	6,154,683
LSD16U	6,878,939	6,835,964	13,714,903
EB1.20U	3,074,674	2,901,840	5,976,514
EB1.31	6,373,853	5,896,757	12,270,610
EB2.20	4,796,914	4456957	9,253,871
EB2.31	10,621,110	9933779	20,554,889
LB1U	8,398,013	7,811,861	16,209,874
LB2U	7,268,581	6,853,502	14,122,083
Total			517,964,470

Tab. C.22: Number of raw reads per sample after paired-end sequencing using a Illimina HiSeq 2000. U indicate UDG treated samples, N are non-treated samples. EB are extraction blanks, whereas LB are line blanks. Sample LSD16 is a suspected non-TB associated sample as an additional negative control.

Bibliography

[1] U.S. Dep. of Energy, Office of Energy Research, and Office of Biological and Environmental Research. Human Genome Program. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

[2] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[3] X-Prize Foundation. Archon Genomics X-Prize. http://genomics.xprize.org/, 1997.

[4] Elaine R. Mardis. The $1,000 genome, the $100,000 analysis? *Genome Med*, 2(11):84, 2010.

[5] J. Handelsman, M.R. Rondon, S.G. Brady, J. Clardy, and R.M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology*, 5:245–249, 1998.

[6] Autoimmunity Research Foundation. Detecting bacteria: Data from the Marshal Protocol Knowledge Base. Available at: http://mpkb.org/home/pathogenesis/microbiota/detecting Accessed 17.04.2013.

[7] J Craig Venter, Karin Remington, John F. Heidelberg, Aaron L. Halpern, Doug Rusch, Jonathan A. Eisen, Dongying Wu, Ian Paulsen, Karen E. Nelson, William Nelson, Derrick E. Fouts, Samuel Levy, Anthony H. Knap, Michael W. Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O. Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, Apr 2004.

[8] National Institute of Health. Human Microbiome Project. http://commonfund.nih.gov/hmp/, 1997.

[9] Yongan Zhao, Haixu Tang, and Yuzhen Ye. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, 28(1):125–126, Jan 2012.

[10] Yuzhen Ye, Jeong-Hyeon Choi, and Haixu Tang. RAPSearch: a fast protein similarity search tool for short reads. *BMC Bioinformatics*, 12:159, 2011.

[11] Gail Rosen, Elaine Garbarine, Diamantino Caseiro, Robi Polikar, and Bahrad Sokhansanj. Metagenome fragment classification using N-mer frequency profiles. *Adv Bioinformatics*, 2008:205969, 2008.

[12] Fabian Schreiber, Peter Gumrich, Rolf Daniel, and Peter Meinicke. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*, 26(7):960–961, Apr 2010.

[13] Daniel H. Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res*, 21(9):1552–1560, Sep 2011.

[14] Wolfgang Gerlach, Sebastian Jünemann, Felix Tille, Alexander Goesmann, and Jens Stoye. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics*, 10:430, 2009.

[15] Elizabeth M. Glass, Jared Wilkening, Andreas Wilke, Dionysios Antonopoulos, and Folker Meyer. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, 2010(1):pdb.prot5368, Jan 2010.

[16] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp*, 2(1):3, 2012.

[17] R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson. Dna sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–284, 1984.

[18] S. Pääbo. Molecular cloning of ancient egyptian mummy dna. *Nature*, 314(6012):644–645, 1985.

[19] Webb Miller, Daniela I. Drautz, Aakrosh Ratan, Barbara Pusey, Ji Qi, Arthur M. Lesk, Lynn P. Tomsho, Michael D. Packard, Fangqing Zhao, Andrei Sher, Alexei Tikhonov, Brian Raney, Nick Patterson, Kerstin Lindblad-Toh, Eric S. Lander, James R. Knight, Gerard P. Irzyk, Karin M. Fredrikson, Timothy T. Harkins, Sharon Sheridan, Tom Pringle, and Stephan C. Schuster. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220):387–390, Nov 2008.

[20] Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Zeljko Kucan, Ivan Gusic, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, May 2010.

[21] Craig D. Millar, Leon Huynen, Sankar Subramanian, Elmira Mohandesan, and David M. Lambert. New developments in ancient genomics. *Trends Ecol Evol*, 23(7):386–393, Jul 2008.

[22] Ermanno Rizzi, Martina Lari, Elena Gigli, Gianluca De Bellis, and David Caramelli. Ancient DNA studies: new perspectives on old samples. *Genet Sel Evol*, 44:21, 2012.

[23] Kirsten I. Bos, Verena J. Schuenemann, G Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, Sharon N. DeWitte, Matthias Meyer, Sarah Schmedes, James Wood, David J D. Earn, D Ann Herring, Peter Bauer, Hendrik N. Poinar, and Johannes Krause. A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, 478(7370):506–510, Oct 2011.

[24] Verena J. Schuenemann, Pushpendra Singh, Thomas A. Mendum, Ben Krause-Kyora, Günter Jäger, Kirsten I. Bos, Alexander Herbig, Christos Economou, Andrej Benjak, Philippe Busso, Almut Nebel, Jesper L. Boldsen, Anna Kjellström, Huihai Wu, Graham R. Stewart, G Michael Taylor, Peter Bauer, Oona Y-C. Lee, Houdini H T. Wu, David E. Minnikin, Gurdyal S. Besra, Katie Tucker, Simon Roffey, Samba O. Sow, Stewart T. Cole, Kay Nieselt, and Johannes Krause. Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. *Science*, Jun 2013.

[25] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Natl. Acad. Sci*, 74:560–564, 1977.

[26] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[27] Life Technologies. Life Technologies Receives FDA 510(k) Clearance for Diagnostic Use of Sanger Sequencing Platform and HLA Typing Kits. Available at: http://www.reuters.com/article/2013/02/11/ca-life-fda-idUSnPnLA57279+160+PRN20130211 Accessed 20.06.2013.

[28] Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W. Kinzler, and Bert Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A*, 100(15):8817–8822, Jul 2003.

[29] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, and E. Kawashima. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res*, 28(20):E87, Oct 2000.

[30] Pål Nyrén. The history of pyrosequencing. *Methods Mol Biol*, 373:1–14, 2007.

[31] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P

Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[32] Gerardo Turcatti, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res*, 36(4):e25, Mar 2008.

[33] Milan Fedurco, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res*, 34(3):e22, 2006.

[34] Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, Sep 2005.

[35] Nader Pourmand, Miloslav Karhanek, Henrik H J. Persson, Chris D. Webb, Thomas H. Lee, Alexandra Zahradníková, and Ronald W. Davis. Direct electrical detection of DNA synthesis. *Proc Natl Acad Sci U S A*, 103(17):6466–6470, Apr 2006.

[36] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner, Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G Thomas Roth, and James Bustillo. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, Jul 2011.

[37] Daniel Aird, Michael G. Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, 12(2):R18, 2011.

[38] Douglas R. Smith, Aaron R. Quinlan, Heather E. Peckham, Kathryn Makowsky, Wei Tao, Betty Woolf, Lei Shen, William F. Donahue, Nadeem Tusneem, Michael P. Stromberg, Donald A. Stewart, Lu Zhang, Swati S. Ranade, Jason B. Warner, Clarence C. Lee, Brittney E. Coleman, Zheng Zhang, Stephen F. McLaughlin, Joel A. Malek, Jon M. Sorenson, Alan P. Blanchard, Jarrod Chapman, David Hillman, Feng Chen, Daniel S. Rokhsar, Kevin J. McKernan, Thomas W. Jeffries, Gabor T. Marth, and Paul M. Richardson. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res*, 18(10):1638–1642, Oct 2008.

[39] Dmitry A. Bolotin, Ilgar Z. Mamedov, Olga V. Britanova, Ivan V. Zvyagin, Dmitriy Shagin, Svetlana V. Ustyugova, Maria A. Turchaninova, Sergey Lukyanov, Yury B. Lebedev, and Dmitriy M. Chudakov. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol*, 42(11):3073–3083, Nov 2012.

[40] Wei Shao, Valerie F. Boltz, Jonathan E. Spindler, Mary F. Kearney, Frank Maldarelli, John W. Mellors, Claudia Stewart, Natalia Volfovsky, Alexander Levitsky, Robert M. Stephens, and John M. Coffin. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology*, 10:18, 2013.

[41] Fredrik Lysholm. Highly improved homopolymer aware nucleotide-protein alignments with 454 data. *BMC Bioinformatics*, 13:230, 2012.

[42] Irina Abnizova, Steven Leonard, Tom Skelly, Andy Brown, David Jackson, Marina Gourtovaia, Guoying Qi, Rene Te Boekhorst, Nadeem Faruque, Kevin Lewis, and Tony Cox. Analysis of context-dependent errors for illumina sequencing. *J Bioinform Comput Biol*, 10(2):1241005, Apr 2012.

[43] Daniel R. Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829, May 2008.

[44] Mark J. Chaisson and Pavel A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Res*, 18(2):324–330, Feb 2008.

[45] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272, Feb 2010.

[46] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J M. Jones, and Inanç Birol. ABySS: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–1123, Jun 2009.

[47] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[48] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, Mar 2008.

[49] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, Aug 2009.

[50] Nils Homer, Barry Merriman, and Stanley F Nelson. BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, 4(11):e7767, 2009.

[51] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[52] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, Apr 2012.

[53] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, Mar 2010.

[54] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Res*, 37(Database issue):D26–D31, Jan 2009.

[55] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[56] Henrik Stranneheim, Max Käller, Tobias Allander, Björn Andersson, Lars Arvestad, and Joakim Lundeberg. Classification of DNA sequences using Bloom filters. *Bioinformatics*, 26(13):1595–1600, Jul 2010.

[57] Arthur Brady and Steven L Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*, 6(9):673–676, Sep 2009.

[58] Alice Carolyn McHardy, Héctor García Martín, Aristotelis Tsirigos, Philip Hugenholtz, and Isidore Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4(1):63–72, Jan 2007.

[59] Daniel H. Huson and Chao Xie. A poor man's BLASTX–high-throughput metagenomic protein database search using PAUDA. *Bioinformatics*, Jun 2013.

[60] National Institute of Health. The NCBI taxonomy. http://www.ncbi.nlm.nih.gov/taxonomy, 1997.

[61] Ross Overbeek, Tadhg Begley, Ralph M Butler, Jomuna V Choudhuri, Han-Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, Naryttza Diaz, Terry Disz, Robert Edwards, Michael Fonstein, Ed D Frank, Svetlana Gerdes, Elizabeth M Glass, Alexander Goesmann, Andrew Hanson, Dirk Iwata-Reuyl, Roy Jensen, Neema Jamshidi, Lutz Krause, Michael Kubal, Niels Larsen, Burkhard Linke, Alice C McHardy, Folker Meyer, Heiko Neuweger, Gary Olsen, Robert Olson, Andrei Osterman, Vasiliy Portnoy, Gordon D Pusch, Dmitry A Rodionov, Christian Rückert, Jason Steiner, Rick Stevens, Ines Thiele, Olga Vassieva, Yuzhen Ye, Olga Zagnitko, and Veronika Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–5702, 2005.

[62] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[63] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome Res*, 17(3):377–386, Mar 2007.

[64] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nat Rev Genet*, 14(3):157–167, Mar 2013.

[65] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*, 40(20):e155, Nov 2012.

[66] Jonathan Laserson, Vladimir Jojic, and Daphne Koller. Genovo: de novo assembly for metagenomes. *J Comput Biol*, 18(3):429–443, Mar 2011.

[67] Yu Peng, Henry C M. Leung, S. M. Yiu, and Francis Y L. Chin. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*, 27(13):i94–101, Jul 2011.

[68] Clyde A Hutchison, 3rd. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*, 35(18):6227–6237, 2007.

[69] Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: http://www.genome.gov/sequencingcosts. Accessed 17.04.2013.

[70] Victor M. Markowitz, I-Min A. Chen, Ken Chu, Ernest Szeto, Krishna Palaniappan, Yuri Grechkin, Anna Ratner, Biju Jacob, Amrita Pati, Marcel Huntemann, Konstantinos Liolios, Ioanna Pagani, Iain Anderson, Konstantinos Mavromatis, Natalia N. Ivanova, and Nikos C. Kyrpides. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*, 40(Database issue):D123–D129, Jan 2012.

[71] Sebastian J. Schultheiss, Marc-Christian Münch, Gergana D. Andreeva, and Gunnar Rätsch. Persistence and availability of Web services in computational biology. *PLoS One*, 6(9):e24914, 2011.

[72] Konrad U. Foerstner, Christian von Mering, and Peer Bork. Comparative analysis of environmental sequences: potential and challenges. *Philos Trans R Soc Lond B Biol Sci*, 361(1467):519–523, Mar 2006.

[73] Elizabeth A. Dinsdale, Robert A. Edwards, Barbara A. Bailey, Imre Tuba, Sajia Akhter, Katelyn McNair, Robert Schmieder, Naneh Apkarian, Michelle Creek, Eric Guan, Mayra Hernandez, Katherine Isaacs, Chris Peterson, Todd Regh, and Vadim Ponomarenko. Multivariate analysis of functional metagenomes. *Front Genet*, 4:41, 2013.

[74] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–194, Mar 1998.

[75] D. Walther, G. Bartha, and M. Morris. Basecalling with LifeTrace. *Genome Res*, 11(5):875–888, May 2001.

[76] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, 23(4):452–456, Dec 1999.

[77] Ketil Malde. The effect of sequence quality on sequence alignment. *Bioinformatics*, 24(7):897–900, Apr 2008.

[78] Bioinformatics Group at the Babraham Institute. FastQC Website. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2011.

[79] Hannon Lab. Fastx-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/, 2013.

[80] Erik Aronesty. ea-utils : "Command-line tools for processing biological sequencing data". http://code.google.com/p/ea-utils, 2011.

[81] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, Mar 2011.

[82] Tanja Magoc and Steven L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, Nov 2011.

[83] Sëbastien Rodrigue, Arne C. Materna, Sonia C. Timberlake, Matthew C. Blackburn, Rex R. Malmstrom, Eric J. Alm, and Sallie W. Chisholm. Unlocking short read sequencing for metagenomics. *PLoS One*, 5(7):e11840, 2010.

[84] Aleksandr Morgulis, E Michael Gertz, Alejandro A. Schäffer, and Richa Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol*, 13(5):1028–1040, Jun 2006.

[85] J. C. Wootton and S. Federhen. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*, 266:554–571, 1996.

[86] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[87] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010.

[88] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, Jan 2009.

[89] Dana Willner, Rebecca Vega Thurber, and Forest Rohwer. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol*, 11(7):1752–1766, Jul 2009.

[90] Ruiqiang Li, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, Yuanyuan Ren, Geng Tian, Jinxiang Li, Guangyu Zhou, Xuan Zhu, Honglong Wu, Junjie Qin, Xin Jin, Dongfang Li, Hongzhi Cao, Xueda Hu, Hélène Blanche, Howard Cann, Xiuqing Zhang, Songgang Li, Lars Bolund, Karsten Kristiansen, Huanming Yang, Jun Wang, and Jian Wang. Building the sequence map of the human pan-genome. *Nat Biotechnol*, 28(1):57–63, Jan 2010.

[91] Robert Schmieder and Robert Edwards. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*, 6(3):e17288, 2011.

[92] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–Unit 19.1021, Jan 2010.

[93] Suparna Mitra, Paul Rupek, Daniel C Richter, Tim Urich, Jack A Gilbert, Folker Meyer, Andreas Wilke, and Daniel H Huson. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 Suppl 1:S21, 2011.

[94] Daniel C Richter, Felix Ott, Alexander F Auch, Ramona Schmid, and Daniel H Huson. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, 3(10):e3373, 2008.

[95] Konstantinos Mavromatis, Natalia Ivanova, Kerrie Barry, Harris Shapiro, Eugene Goltsman, Alice C McHardy, Isidore Rigoutsos, Asaf Salamov, Frank Korzeniewski, Miriam Land, Alla Lapidus, Igor Grigoriev, Paul Richardson, Philip Hugenholtz, and Nikos C Kyrpides. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, Jun 2007.

[96] David R Bentley. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16(6):545–552, Dec 2006.

[97] UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, 40(Database issue):D71–D75, Jan 2012.

[98] Jenna L. Morgan, Aaron E. Darling, and Jonathan A. Eisen. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One*, 5(4):e10209, 2010.

[99] R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, Oct 1997.

[100] Roman L. Tatusov, Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, Raja Mazumder, Sergei L. Mekhedov, Anastasia N. Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V. Sverdlov, Sona Vasudevan, Yuri I. Wolf, Jodie J. Yin, and Darren A. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

[101] J. Muller, D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L. J. Jensen, and P. Bork. eggnog v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*, 38(Database issue):D190–D195, Jan 2010.

[102] Sean Powell, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic, Tobias Doerks, Lars J. Jensen, Christian von Mering, and Peer Bork. eggnog v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res*, 40(Database issue):D284–D289, Jan 2012.

[103] Jeremy Goecks, Carl Eberhard, Tomithy Too, Anton Nekrutenko, and James Taylor. Web-based visual analysis for high-throughput genomics. *BMC Genomics*, 14(1):397, Jun 2013.

[104] Unknown. Picard tools. `http://picard.sourceforge.net`, 2011.

[105] Aurelien Ginolhac, Morten Rasmussen, M Thomas P. Gilbert, Eske Willerslev, and Ludovic Orlando. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15):2153–2155, Aug 2011.

[106] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.

[107] Magdalena Feldhahn. Computational methods for personalized cancer therapy based on genomics data. Dissertation - Uni-Tuebingen, 2013.

[108] World Health Organisation. Global tuberculosis report 2012. `http://www.who.int/tb/publications/global_report/en/index.html`, 2012.

[109] Sebastian Gagneux, Marcos V. Burgos, Kathryn DeRiemer, Antonio Encisco, Samira Munoz, Phillip C. Hopewell, Peter M. Small, and Alexander S. Pym. Impact of bacterial genetics on the transmission of isoniazid-resistant Mycobacterium tuberculosis. *PLoS Pathog*, 2(6):e61, Jun 2006.

[110] Stewart T. Cole. Comparative and functional genomics of the mycobacterium tuberculosis complex. *Microbiology*, 148(Pt 10):2919–2928, Oct 2002.

[111] S. Sreevatsan, X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser. Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A*, 94(18):9869–9874, Sep 1997.

[112] M. D. Iseman. Evolution of drug-resistant tuberculosis: a tale of two species. *Proc Natl Acad Sci U S A*, 91(7):2428–2429, Mar 1994.

[113] Borna Müller, Salome Dürr, Silvia Alonso, Jan Hattendorf, Claudio J M. Laisse, Sven D C. Parsons, Paul D. van Helden, and Jakob Zinsstag. Zoonotic Mycobacterium bovis-induced Tuberculosis in Humans. *Emerg Infect Dis*, 19(6):899–908, Jun 2013.

[114] Thierry Wirth, Falk Hildebrand, Caroline Allix-Béguec, Florian Wölbeling, Tanja Kubica, Kristin Kremer, Dick van Soolingen, Sabine Rüsch-Gerdes, Camille Locht, Sylvain Brisse, Axel Meyer, Philip Supply, and Stefan Niemann. Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog*, 4(9):e1000160, 2008.

[115] M Cristina Gutierrez, Sylvain Brisse, Roland Brosch, Michel Fabre, Bahia Omaïs, Magali Marmiesse, Philip Supply, and Veronique Vincent. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. *PLoS Pathog*, 1(1):e5, Sep 2005.

[116] R. Brosch, S. V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L. M. Parsons, A. S. Pym, S. Samper, D. van Soolingen, and S. T. Cole. A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proc Natl Acad Sci U S A*, 99(6):3684–3689, Mar 2002.

[117] Thierry Garnier, Karin Eiglmeier, Jean-Christophe Camus, Nadine Medina, Huma Mansoor, Melinda Pryor, Stephanie Duthoy, Sophie Grondin, Celine Lacroix, Christel Monsempe, Sylvie Simon, Barbara Harris, Rebecca Atkin, Jon Doggett, Rebecca Mayes, Lisa Keating, Paul R. Wheeler, Julian Parkhill, Bart G. Barrell, Stewart T. Cole, Stephen V. Gordon, and R Glyn Hewinson. The complete genome sequence of Mycobacterium bovis. *Proc Natl Acad Sci U S A*, 100(13):7877–7882, Jun 2003.

[118] Noel H. Smith. A re-evaluation of M. prototuberculosis. *PLoS Pathog*, 2(9):e98, Sep 2006.

[119] Sylvain Brisse, Philip Supply, Roland Brosch, Veronique Vincent, and M Cristina Gutierrez. "A re-evaluation of M. prototuberculosis": continuing the debate. *PLoS Pathog*, 2(9):e95, Sep 2006.

[120] Mark Stoneking and Johannes Krause. Learning about human population history from ancient and modern genomes. *Nat Rev Genet*, 12(9):603–614, Sep 2011.

[121] Eske Willerslev and Alan Cooper. Ancient DNA. *Proc Biol Sci*, 272(1558):3–16, Jan 2005.

[122] Adrian W. Briggs, Udo Stenzel, Philip L F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, Johannes Krause, Michael T. Ronan, Michael Lachmann, and Svante Pääbo. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*, 104(37):14616–14621, Sep 2007.

[123] Adrian W. Briggs, Udo Stenzel, Matthias Meyer, Johannes Krause, Martin Kircher, and Svante Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*, 38(6):e87, Apr 2010.

[124] Günther Jäger. Personal Communication. Unpublished, 2013.

[125] Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, 5(6):e11147, 2010.

[126] Ying-Chuan Liu Ying Chih Lin, Chin Lung Lu and Chuan Yi Tang. SPRING genome rearrangement server. http://http://algorithm.cs.nthu.edu.tw/tools/SPRING/.

[127] Ingrid Filliol, Alifiya S. Motiwala, Magali Cavatore, Weihong Qi, Manzour Hernando Hazbón, Miriam Bobadilla del Valle, Janet Fyfe, Lourdes García-García, Nalin Rastogi, Christophe Sola, Thierry Zozio, Marta Inírida Guerrero, Clara InÃ©s León, Jonathan Crabtree, Sam Angiuoli, Kathleen D. Eisenach, Riza Durmaz, Moses L. Joloba, Adrian Rendón, José Sifuentes-Osornio, Alfredo Ponce de León, M Donald Cave, Robert Fleischmann, Thomas S. Whittam, and David Alland. Global phylogeny of mycobacterium tuberculosis based on single nucleotide polymorphism (snp) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other dna fingerprinting systems, and recommendations for a minimal standard snp set. *J Bacteriol*, 188(2):759–772, Jan 2006.

[128] Inaki Comas, Jaidip Chakravartti, Peter M. Small, James Galagan, Stefan Niemann, Kristin Kremer, Joel D. Ernst, and Sebastien Gagneux. Human t cell epitopes of mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet*, 42(6):498–503, Jun 2010.

[129] Emma L. Best, Warren N. Fawley, Peter Parnell, and Mark H. Wilcox. The potential for airborne dispersal of clostridium difficile from symptomatic patients. *Clin Infect Dis*, 50(11):1450–1457, Jun 2010.